МИНИСТЕРСТВО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

МИНИСТЕРСТВО РОССИЙСКОЙ ФЕДЕРАЦИИ ПО АТОМНОЙ ЭНЕРГИИ

МИНИСТЕРСТВО ПРОМЫШЛЕННОСТИ, НАУКИ И ТЕХНОЛОГИЙ
РОССИЙСКОЙ ФЕДЕРАЦИИ

РОССИЙСКАЯ АССОЦИАЦИЯ НЕЙРОИНФОРМАТИКИ

МОСКОВСКИЙ ИНЖЕНЕРНО-ФИЗИЧЕСКИЙ ИНСТИТУТ
(ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ)

НАУЧНАЯ СЕССИЯ МИФИ–2003

НЕЙРОИНФОРМАТИКА–2003

**V ВСЕРОССИЙСКАЯ
НАУЧНО-ТЕХНИЧЕСКАЯ
КОНФЕРЕНЦИЯ**

ЛЕКЦИИ
ПО НЕЙРОИНФОРМАТИКЕ

Часть 1

По материалам Школы-семинара
«Современные проблемы нейроинформатики»

Москва 2003

УДК 004.032.26 (06)
ББК 32.818я5
М82

НАУЧНАЯ СЕССИЯ МИФИ–2003. V ВСЕРОССИЙСКАЯ НАУЧНО-ТЕХНИЧЕСКАЯ КОНФЕРЕНЦИЯ «НЕЙРОИНФОРМАТИКА–2003»: ЛЕКЦИИ ПО НЕЙРОИНФОРМАТИКЕ. Часть 1. – М.: МИФИ, 2003. – 188 с.

В книге публикуются тексты лекций, прочитанных на Школе-семинаре «Современные проблемы нейроинформатики», проходившей 29–31 января 2003 года в МИФИ в рамках V Всероссийской конференции «Нейроинформатика–2003».

Материалы лекций связаны с рядом проблем, актуальных для современного этапа развития нейроинформатики, включая ее взаимодействие с другими научно-техническими областями.

Ответственный редактор
Ю. В. Тюменцев, кандидат технических наук

ISBN 5–7262–0471–9

Содержание

УДК 004.032.26 (06) Нейронные сети

ПРЕДИСЛОВИЕ

1. В этой книге (она выходит в двух частях) содержатся тексты лекций, прочитанных на Школе-семинаре «Современные проблемы нейроинформатики», проходившей 29–31 января 2003 года в МИФИ в рамках V Всероссийской научно-технической конференции «Нейроинформатика–2003».

При отборе и подготовке материалов для лекций авторы и редактор следовали принципам и подходам, сложившимся при проведении двух предыдущих Школ (см. [1–3]).

А именно, основной целью Школы было рассказать слушателям о современном состоянии и перспективах развития важнейших направлений в теории и практике нейроинформатики, о ее применениях.

При подготовке программы Школы особенно приветствовались лекции междисциплинарные, лежащие по охватываемой тематике «на стыке наук», рассказывающие о проблемах не только собственно нейроинформатики (т. е. о проблемах, связанных с нейронными сетями, как естественными, так и искусственными), но и о взаимосвязях нейроинформатики с другими областями мягких вычислений (нечеткие системы, генетические и другие эволюционные алгоритмы и т. п.), с системами, основанными на знаниях, с традиционными разделами математики, биологии, психологии, инженерной теории и практики.

Основной задачей лекторов, приглашаемых из числа ведущих специалистов в области нейроинформатики и ее приложений, смежных областей науки, было дать живую картину современного состояния исследований и разработок, обрисовать перспективы развития нейроинформатики в ее взаимодействии с другими областями науки.

Помимо междисциплинарности, приветствовалась также и дискуссионность излагаемого материала. Как следствие, не со всеми положениями, выдвигаемыми авторами, можно безоговорочно согласиться, но это только повышает ценность лекций — они стимулируют возникновение дискуссии, выявление пределов применимости рассматриваемых подходов, поиск альтернативных ответов на поставленные вопросы, альтернативных решений сформулированных задач.

2. В программу Школы-семинара «Современные проблемы нейроинформатики» на конференции «Нейроинформатика–2003» вошли следующие семь лекций [1]:

[1] Первые четыре из перечисленных лекций публикуются в части 1, а оставшиеся три — в части 2 сборника «Лекции по нейроинформатике».

УДК 004.032.26 (06) Нейронные сети

1. *А. А. Фролов, Д. Гусек, И. П. Муравьев.* Информационная эффективность ассоциативной памяти типа Хопфилда с разреженным кодированием.
2. *Б. В. Крыжановский, Л. Б. Литинский.* Векторные модели ассоциативной памяти.
3. *Н. Г. Макаренко.* Эмбедология и нейропрогноз.
4. *С. А. Терехов.* Введение в байесовы сети.
5. *А. А. Ежов.* Некоторые проблемы квантовой нейротехнологии.
6. *А. Ю. Хренников.* Классические и квантовые модели мышления, основанные на p-адическом представлении информации.
7. *Ю. И. Нечаев.* Математическое моделирование в бортовых интеллектуальных системах реального времени.

Характерная объединяющая черта семи лекций, публикуемых в настоящем сборнике, состоит в том, что все они посвящены обсуждению различных подходов к моделированию интеллектуальных процессов и систем. Эти подходы едва ли можно назвать конкурирующими, скорее их надо расценивать как взаимодополняющие — в духе принципа дополнительности *Нильса Бора.*

3. Первая пара лекций, открывавшая Школу, была посвящена одной из классических тем — *ассоциативной памяти*, причем истоки подходов, рассмотренных в обеих лекциях, также относятся к классике, к модели Хопфилда, оказавшей очень большое влияние на развитие нейроинформатики.

Общеизвестен факт — именно с публикации в 1982 году физиком *Джоном Хопфилдом* статьи [4] началось возрождение и последующее бурное развитие нейроинформатики после примерно полутора десятилетий относительного затишья.

Сети Хопфилда в многочисленных разновидностях до сих пор остаются популярной нейросетевой моделью, привлекающей к себе внимание исследователей. Не в последнюю очередь такая популярность объясняется способностью хопфилдовых сетей выполнять функции ассоциативной памяти[2], т. е. памяти, адресуемой по содержимому, обеспечивающей хранение и извлечение паттернов (образов).

[2]Об ассоциативной памяти и двух ее основных разновидностях — *гетероассоциативной* памяти и *автоассоциативной* памяти см., например, статью *А. А. Фролова* в сборнике [8].

Один из возможных вариантов решения проблемы ассоциативной памяти рассмотрен в лекции **А. А. Фролова, Д. Гусека, И. П. Муравьева** «Информационная эффективность ассоциативной памяти типа Хопфилда с разреженным кодированием». В ней исследуется сеть хопфилдового типа, которая действует как автоассоциативная память для статистически независимых бинарных паттернов, т. е. паттернов, элементы которых могут принимать только два значения (например, 0 и 1).

Применительно к сетям такого вида существует несколько типичных проблем, в том числе: проблема информационной емкости (сколько паттернов-эталонов можно записать в такую сеть и затем воспроизвести их?); проблема качества воспроизведения (какова будет доля ошибок в выходных паттернах в сравнении с воспроизводимыми эталонами?); проблема размеров областей притяжения (насколько сильно может быть искажен эталон, чтобы сохранить свойство воспроизводимости?).

Одним из серьезных недостатков сети Хопфилда в ее первоначальной формулировке была невысокая информационная емкость таких сетей. Сеть из N нейронов может иметь 2^N состояний, но максимальная емкость памяти оказывается значительно меньшей. Предполагалось вначале, что максимальное количество запоминаемых паттернов, которые могут безошибочно извлекаться, будет доходить до величины cN^2, где $c > 1$ — положительная константа [10]. Эта оценка оказалась слишком оптимистической. Было показано, что число запоминаемых паттернов не может превышать N, причем в общем случае оно будет ближе к $0.14N$ ([5, 6]; см. также [7, 10]).

Один из возможных подходов, позволяющих увеличить информационную емкость сети Хопфилда — *разреженное кодирование*, т. е. такое кодирование, при котором количество активных нейронов n в записанных паттернах (эталонах) много меньше общего количества N нейронов в сети. В предельном случае, когда $n/N \to 0$, оценка максимального числа запоминаемых паттернов составляет $0.72N$.

В лекции, основываясь на теоретическом анализе и компьютерном эксперименте, даются ответы на вопросы, сформулированные выше, причем прежде всего анализируется влияние разреженности на размер областей притяжения.

4. Резкое увеличение числа элементов и использование разреженного кодирования в сетях хопфилдова типа с традиционными бинарными нейронами — это один из возможных путей повышения информационной эффективности сетей данного вида и ассоциативной памяти на их основе. Существует, однако, альтернативный вариант, в основе которого — исполь-

зование сравнительно небольшого числа нейронов, каждый из которых может принимать q состояний, т. е. так называемых *q-нарных нейронов*. Сети из элементов такого вида рассматриваются в лекции **Б. В. Крыжановского, Л. Б. Литинского** «Векторные модели ассоциативной памяти».

Исследования в области моделей ассоциативной памяти с q-нарными нейронами ведутся уже примерно в течение 15 лет. Был предложен целый ряд схем, позволяющих приписать нейрону q различных состояний, а также нейросетей с такими элементами. Совсем недавно был предложен еще один вариант сетей с q-нарными нейронами, получивший наименование *«параметрическая нейронная сеть»* [12, 13]. Вначале она была ориентирована на нелинейно-оптические принципы обработки информации, затем была формализована для общего случая в рамках векторного подхода к описанию нейронов.

Если, как уже отмечалось выше, традиционная модель Хопфилда (т. е. модель с бинарными элементами и плотным кодированием) может эффективно запомнить сравнительно небольшое число паттернов, а именно, порядка $0.14N$, где N — число элементов в сети (в случае разреженного кодирования этот показатель может быть существенно выше), то в моделях с q-нарным нейроном, особенно в параметрической нейронной сети, данный показатель удается существенно превзойти, в частности, число запоминаемых паттернов может превышать число нейронов в два и более раз, при этом обеспечивается высокая вероятность правильного восстановления сильно зашумленных паттернов.

5. Как уже отмечалось выше, все семь лекций Школы 2003 года были посвящены различным аспектам проблемы моделирования интеллектуальных процессов и систем.

Проблема моделирования процессов и систем «стара как мир», она существует столько же лет, сколько и сама наука. Как сказано в известной книге *Леннарта Льюнга* (см. [14], с. 15): «Формирование моделей на основе результатов наблюдений и исследование их свойств — вот, по существу, основное содержание науки [3]. Модели (“гипотезы”, “законы природы”, “парадигмы” и т. п.) могут быть более или менее формализованными, но все обладают той главной особенностью, что связывают наблюдения в некую общую картину».

[3] Иными словами — «извлечение идей» (сущностей, как называл идеи *Платон*) из объектов и систем материального мира; с другой стороны, науку можно описать также и как деятельность, направленную на объективизацию, «материализацию» сущностей.

Одному из новейших и перспективных подходов к построению математических моделей непосредственно из наблюдений, из данных эксперимента, посвящена лекция **Н. Г. Макаренко** «Эмбедология и нейропрогноз».

В ней изучается случай, когда есть результаты наблюдений за некоторым объектом, причем эти результаты представлены в виде скалярного временного ряда, как это чаще всего и бывает на практике.

В традиционной трактовке принято считать временной ряд [4] дискретным случайным процессом (точнее, наблюдаемой конечной реализацией дискретного случайного процесса), анализ которого осуществляется методами теории вероятностей и математической статистики.

В последние 15–20 лет анализ временных рядов стал одной из наиболее активно развиваемых областей теории вероятностей и математической статистики, имеющей многочисленные приложения в физике, технике, экономике, социологии, биологии, лингвистике, т. е. в тех областях, где приходится иметь дело со стохастическими стационарными рядами наблюдений, или же с рядами наблюдений, отличающихся от стационарных легко выделяемым трендом, периодическими составляющими и т. п.

Практически для всех вариантов статистического подхода характерно то, что ответом в них будет некая *функция*, более или менее «хорошо» описывающая исходные экспериментальные данные.

Зададимся, однако, вопросом — а что явилось источником [5] анализируемого временного ряда? Вполне логичным представляется предположение, вводимое в лекции *Н. Г. Макаренко*, о том, что «. . . отсчеты ряда являются нелинейной проекцией движения фазовой точки некоторой динамической системы, продуцирующей ряд. . . »

Если встать на эту точку зрения, то намного более привлекательным (с точки зрения потенциально достижимых прикладных результатов) выглядит подход, позволяющий «восстановить» не просто одну фазовую траекторию (реализацию временного ряда), полученную для конкретных начальных условий и возмущений «продуцирующей системы», как это имеет место в статистических подходах, но попытаться восстановить «природу»

[4] Понятие *временного ряда* не обязательно связано с процессами, развертывающимися во времени; в качестве независимой переменной t может быть взята, например, некоторая пространственная координата.

[5] Заметим, что в традиционных методах анализа вопрос о природе источника вообще не имеет смысла. Это происходит потому, что ответ на него обычно заложен уже в самом методе. Так, Фурье-анализ временного ряда сразу предполагает полигармоническую модель источника.

этой системы, т. е. ее динамическое описание, диффеоморфизм, отвечающий (по-возможности) *всем* временным рядам, которые могли бы быть порождены исследуемой системой-оригиналом при всех возможных значениях начальных условий и возмущающих воздействий.

Надежда на успешное решение такого рода задач появилась после публикации *Ф. Такенсом* в 1981 году статьи [6], где доказывалась *теорема о типичном вложении временного ряда* в n-мерное евклидово пространство. В лекции *Н. Г. Макаренко* отмечается: «Так возник новый способ построения модели из наблюдаемого сигнала (или реализации), который тут же инициировал совершенно новую область численных методов топологической динамики — *эмбедологию* [7]». И далее: «С ее помощью наконец-то удалось подойти к проблеме моделирования динамики с “правильного конца”: модель начиналась с “ответа” — наблюдений, а затем ставился “правильный” вопрос: что их продуцирует?»

Лекция *Н. Г. Макаренко* как раз и посвящена изложению основных идей эмбедологии — этого многообещающего подхода, который приводит к получению многомерного варианта авторегрессионного прогноза временных рядов.

Здесь следует отметить, что «стандартный» авторегрессионный прогноз (см., например, [25]) основан на линейной комбинации прошлых значений. «Предсказательные возможности» прогноза можно, очевидно, повысить, если перейти от использования линейной комбинации к некоторой нелинейной функции. Именно такой подход и осуществляется в рамках эмбедологии. Поиск предиктора при этом сведен к проблеме аппроксимации функции нескольких переменных с использованием искусственной нейронной сети.

При определенных условиях, подробно обсуждаемых в лекции, эмбедология позволяет: из временного ряда восстановить фазовый портрет неизвестной системы, порождающей ряд; оценить размерность аттрактора и стохастичность системы; восстановить (хотя и не всегда) уравнения системы; реализовать нелинейную схему прогноза; обнаружить и оценить взаимодействие двух систем.

По материалу лекции *Н. Г. Макаренко* можно рекомендовать для ознакомления помимо тех источников, что указаны в списке литературы к лекции, еще и книги: [16–23], а также его лекцию на Школе 2002 года [15].

[6] Ссылка на нее есть в списке литературы к лекции *Н. Г. Макаренко*.

[7] Название данной области происходит от английского термина *embedology*, который произошел, в свою очередь, от названия теоремы Такенса: *embedding theorem*.

Кроме того, довольно много публикаций (статей, препринтов и т. п.) можно найти в библиотеке ResearchIndex [72], если задать, например, поиск по терминам “embedding theorem”, “Takens theorem”.

6. В лекции «Введение в байесовы сети» **С. А. Терехов** возвращается к теме, лишь вскользь затронутой им в лекции на Школе 2002 года [24] при обсуждении подходов к аппроксимации плотности распределения вероятности в рамках задачи информационного моделирования.

Моделирование интеллектуальных процессов, создание интеллектуальных систем различного назначения в рамках как парадигмы искусственного интеллекта, так и парадигмы мягких вычислений [8], всегда существенно опиралось на понятие неопределенности, понимаемой традиционно как неопределенность *невероятностного* характера (нечеткость, недоопределенность, неполнота, неточность и т. п.). Отношение к неопределенности вероятностного типа в ИИ- и МВ-сообществе стало меняться в последние 10–15 лет и произошло это вследствие появления *байесовых сетей* [9] — графических моделей для представления неопределенностей и взаимосвязей между ними.

В математическом плане байесова сеть представляет собой ориентированный ациклический граф специального вида. Каждая вершина в нем отвечает некоторой переменной из решаемой задачи, распределение вероятностей для которой интерпретируется как «значение ожидания» (belief value) [10], а каждая дуга — как зависимость между переменными-вершинами, численно выражаемая таблицей условных вероятностей.

Можно соглашаться или не соглашаться с оценкой степени значимости

[8] Согласно первоначальному определению *мягких вычислений* (МВ), которое дал *Л. Заде* ([28, 29]; см. также о развитии этого направления в [28, 29]), в перечень дисциплин, объединяемых в составе МВ, традиционный искусственный интеллект (ИИ) как область, где изучаются проблемы представления, извлечения и использования знаний, не входил. Однако, к настоящему времени грань между МВ и ИИ все больше размывается и их вместе можно считать «расширенным ИИ» или «расширенными МВ».

[9] Первооткрывателем модели, получившей впоследствии наименование «байесова сеть», был американский статистик *С. Райт* (*S. Wright*), предложивший соответствующую модель в 1921 году (ссылка на его работу есть в списке литературы к лекции *С. А. Терехова*). Как это часто бывало в истории науки, данную модель несколько раз «переоткрывали», давая ей различные имена: сети причинности (causal nets), сети вывода (inference nets), сети ожиданий (belief networks), диаграммы влияния (influence diagrams).

[10] Belief — слово многозначное, в контексте байесовых сетей (и шире — систем ИИ) может переводиться как «ожидание» (ожидание того, что произойдет то или иное событие), «вера», «доверие», означая при этом степень ожидания (веры, уверенности) в том, что произойдет то или иное событие.

байесовых сетей как «самой революционной технологии десятилетия в области ИИ» [11], но то, что это область, потенциально богатая приложениями, едва ли подлежит сомнению.

Примеры таких приложений, приводимые в лекции *С. А. Терехова*, не исчерпывают, разумеется, всех уже имеющихся на сегодняшний день. Ряд примеров применений байесовых сетей рассматривается в тематическом выпуске журнала “Communications of the ACM” [30] — это отладка и сопровождение компьютерных программ, поиск информации в базах данных, задачи диагностики систем различного назначения и многое другое. Еще несколько примеров применений содержится в статьях [31] — автоматическое формирование тезауруса в информационных системах [32, 33] — распознавание и анализ изображений.

Байесовы сети тесно связаны с другими сетевыми моделями, в том числе и с нейросетевой моделью. В частности, в лекции *С. А. Терехова* указывается на взаимосвязи байесовых сетей и нейронных сетей со встречным распространением (Counter-Propagation Network). Еще один вариант такого рода взаимосвязи — с нейросетевой соревновательной архитектурой (WTA-архитектурой) [12] приводится в работе [34].

Много информации по байесовым сетям можно найти в цифровой библиотеке ResearchIndex [72], если запросить публикации по темам *Bayesian network* (net), *belief network* (net), *inference network* (net).

Популярное изложение материала о байесовых сетях, а также пакет расширения (Bayes Net Toolbox) для системы Matlab содержится по адресам, указанным в позиции [35] списка литературы к предисловию.

Изложение вероятностно-статистического аппарата, используемого при работе с байесовыми сетями, можно найти в книгах [25–27].

7. В лекции **А. А. Ежова** «Некоторые проблемы квантовой нейротехнологии» обсуждаются возможные взаимосвязи нейротехнологии с быстро развивающейся областью *квантовых вычислений*.

Квантовые вычисления, квантовые нейронные сети постоянно присутствует в перечне тем, обсуждаемых на конференциях «Нейроинформатика». До сих пор эта тематика была представлена только докладами (пленарным, секционными, на Рабочем совещании [36]), теперь же она излагается в виде лекции на Школе-семинаре.

[11] Такая оценка — со ссылкой на мнение *Билла Гейтса* — приводилась *С. А. Тереховым* в его лекции на Школе 2002 года [24].

[12] WTA — Winner-Take-All, «победитель забирает все».

Цель лекции — попытаться дать ответ на вопрос о том, какой может быть квантовая нейротехнология, каковы ее связи с квантовыми вычислениями и квантовыми компьютерами.

Обобщение нейронных технологий на квантовую область заставляет искать также ответы и на такие вопросы: какие нейросетевые модели можно назвать квантовыми, каковы перспективы квантовой нейротехнологии, чем отличается квантовая нейротехнология от квантовых вычислений?

Изложению этого круга вопросов, а также неизбежно сопутствующих им проблем квантовой механики (интерференция, запутанность квантовых состояний, многомировая интерпретация квантовой механики и др.) и посвящена лекция *А. А. Ежова.*

Вначале в ней кратко излагаются необходимые понятия и сведения, связанные с квантовыми вычислениями. Затем обсуждается роль запутанности и интерференции в квантовых вычислениях, а также многомировая интерпретация квантовой механики применительно к квантовым вычислениям. После этого приводятся соображения о возможной природе нейросетевых систем, которые можно было бы назвать квантовыми. И, наконец, обсуждается проблема создания квантовых компьютеров для применения их в физическом моделировании, в которых значительную роль могут сыграть квантовые нейронные системы.

Подробный обзор результатов исследований в области квантовых вычислений и квантовых компьютеров (касающийся как теоретических результатов, так и возможных аппаратных реализаций) по состоянию на 2000 год приводится в книге [37].

По теме лекции *А. А. Ежова*, кроме книги [37] и книги [40], а также публикаций, содержащихся в списке литературы к лекции, можно указать также следующие источники. Издательством «Регулярная и хаотическая динамика» (РХД) выпускается серия сборников «Квантовый компьютер и квантовые вычисления» (см. [38, 39]); этим же издательством выпускается журнал «Квантовый компьютер и квантовые вычисления». По тематике квантовых вычислений издательством «РХД» выпущены также книги [41, 42]. Понимание идей, которые положены в основу квантовых вычислений, существенно облегчается, если знать историю возникновения и развития основных понятий и концепций квантовой механики. Для знакомства с этим кругом вопросов можно рекомендовать книги [41, 43, 44].

Много публикаций по тематике квантовых вычислений, квантовых нейронных сетей и смежным вопросам можно найти в цифровом архиве arXiv.org e-Print archive [73].

8. В определенной степени к тематике лекции *А. А. Ежова* примыкает и лекция **А. Ю. Хренникова** «Классические и квантовые модели мышления, основанные на p-адическом представлении информации». Она дает еще один, существенно отличающийся от других, взгляд на проблему математического моделирования процессов и систем.

Эта лекция посвящена изучению возможностей решения такой важнейшей проблемы, как математическое моделирование сознания и когнитивных процессов.

Результаты нейрофизиологических и психологических исследований позволяют говорить об иерархической структуре когнитивных процессов. В качестве одного из возможных перспективных подходов к математическому моделированию таких процессов предлагается использовать p-адические иерархические деревья.

Модели, рассматриваемые в лекции *А. Ю. Хренникова*, основаны на математическом аппарате, сравнительно мало известном в нейроинформационном сообществе. В связи с этим, целесообразно дать некоторые пояснения, облегчающие понимание материала лекции.

Для физических процессов, протекающих в пространстве и во времени, математической моделью физического пространства принято обычно [13] считать вещественное евклидово трехмерное пространство (псевдоевклидово четырехмерное — для случая пространства-времени [14]). Пространственно-временные координаты при этом задаются обычно вещественными (действительными) числами.

Такие представления привычны, но всегда ли они будут соответствовать физической реальности? Другими словами, во всех ли случаях можно считать евклидово пространство «хорошей» моделью физического пространства? Чтобы ответить на этот вопрос, надо проверить, как отвечают реальности соответствующие геометрические аксиомы, достаточно известные еще из школьного курса геометрии.

[13] «Обычно» здесь надо трактовать как «из опыта человеческой деятельности», протекающей, большей частью, в макромире.

[14] *Евклидово пространство* является непосредственным обобщением обычного трехмерного пространства. Для двух векторов («точек») этого пространства $\mathbf{x} = (x_1, \ldots, x_n)$ и $\mathbf{y} = (y_1, \ldots, y_n)$, заданных своими декартовыми координатами, может быть определено скалярное произведение $(\mathbf{xy}) = (x_1 y_1, \ldots, x_n y_n)$, удовлетворяющее ряду условий, в числе которых условие $(\mathbf{xx}) \geqslant 0$, $(\mathbf{xx}) = 0$ лишь при $\mathbf{x} = 0$ (положительная определенность). Число $|\mathbf{x}| = \sqrt{(\mathbf{xx})}$ называется *нормой* (или *длиной*) вектора $\mathbf{x} = 0$. Пространство, в котором нарушено условие положительной определенности, называется *псевдоевклидовым пространством*.

Одна из них — *аксиома Архимеда*, в которой речь идет о двух отрезках, A и B, $A < B$, отложенных на прямой линии и имеющих начало в одной точке. Согласно данной аксиоме, если последовательно откладывать меньший отрезок A вдоль прямой, то в конце концов мы выйдем за пределы отрезка B, или, более формально, для данной величины имеет место аксиома Архимеда, если для любых двух значений A и B, $A < B$, этой величины *всегда* можно найти целое число m такое, что $Am > B$.

По существу, аксиома Архимеда описывает процедуру измерения, постулируя при этом возможность измерять сколь угодно малые расстояния.

Однако, из квантовой теории известно, что принципиально невозможно измерять длины, меньшие так называемой *планковской длины* (величина ее порядка 10^{-33} см). Значит, в реальном физическом пространстве условия аксиомы Архимеда будут выполняться не всегда, пределы ее применимости устанавливаются существованием планковской длины. Но из этого сразу же следует, что и геометрия обычного евклидова пространства [15] не может считаться адекватной свойствам реального физического пространства в случаях, относящихся к миру, где характерные размеры — величины «квантового» порядка малости.

Как уже отмечалось выше, координаты в евклидовом пространстве описываются вещественными числами. Между геометрическим и аналитическим (с помощью чисел) описаниями существует тесная взаимосвязь. Поэтому, если приходится признать, что евклидова геометрия перестает «работать» в микромире, то, следовательно, придется признать неправомочным и использование вещественных чисел в качестве средства аналитического описания закономерностей в микромире. В этом случае требуется привлекать какую-то другую числовую систему.

Можно ли построить такую числовую систему и на каких принципах она могла бы основываться? Отправной точкой здесь могли бы послужить следующие соображения.

Общепринята [16] убежденность в возможности получить для измеряемой величины сколь угодно много знаков после запятой, проводя измерения со все большей точностью. Эта убежденность, однако, основана на

[15] Варианты геометрий, опирающиеся на аксиому Архимеда и отрицающие ее, правильнее было бы именовать собирательно «архимедовы геометрии» и «неархимедовы геометрии», соответственно. В рассматриваемом контексте это обеспечивало бы более точную расстановку смысловых акцентов. Однако, если второй из этих двух терминов уже вполне устоялся, то первый едва ли можно назвать общеупотребительным.

[16] В смысле примечания 13 на странице 14.

идеализации, которую можно считать верной лишь до того предела, что устанавливается существованием планковской длины. Другими словами, есть предельное *конечное* число «знаков после запятой», которое можно получить в любом физическом эксперименте, с *бесконечным* числом «знаков после запятой» в измеряемых величинах не приходится иметь дело никогда.

Но, как известно, вещественные числа разделяются на рациональные и иррациональные. Рациональные числа могут быть представлены в виде дроби p/q, где p и q — целые, $q \neq 0$, либо в виде конечной или бесконечной десятичной периодической дроби. Иррациональные же числа представимы только в виде бесконечной десятичной непериодической дроби.

Именно иррациональные числа оказываются «под запретом» в случаях, когда существенными оказываются ограничения, основанные на планковской длине; правомерность использования рациональных чисел при этом особых сомнений не вызывает [17].

Однако просто «запретить» иррациональные числа нельзя, одних рациональных чисел для целей моделирования физической реальности недостаточно. В самом деле, всякое иррациональное число можно заключить между двумя рациональными числами, одно из которых меньше, а другое — больше рассматриваемого иррационального числа. При этом разность между данной парой рациональных чисел может быть сделана сколь угодно малой [18], т. е. множество рациональных чисел является плотным в множестве вещественных чисел, но оно, однако, не обладает свойством *непрерывности* (полноты). Потеря непрерывности [19] — это, по-видимому, слишком высокая плата за строгое соответствие аппарата рациональных чисел нашим «измерительным возможностям». Следовательно, надо найти альтернативный вариант пополнения множества рациональных чисел, который не опирался бы на аксиому Архимеда.

И такой вариант существует. Он основывается на p-адических числах, введенных *К. Гензелем* в конце XIX века.

При построении аппарата p-адических чисел уточняется понятие *нормы*

[17] Точно так же, как, по-видимому, не вызывает сомнений обоснованность использования вещественных чисел и евклидовой геометрии в макромире.

[18] Если отвлечься опять от существования планковской длины.

[19] Хотя это тоже идеализация, не имеющая «природных» аналогов, если верна гипотеза о дискретной (квантованной) структуре пространственно-временного мира в области малых масштабов. По поводу обоснованности различного рода идеализаций, привлекаемых в ходе построения математических моделей, см. также [45].

на множестве рациональных чисел, являющееся аналитическим аналогом геометрического понятия *расстояния*. Вводится p-адическая норма, которая в качестве частного случая включает в себя «обычную» норму. При этом пополнение множества рациональных чисел по обычной норме приводит к множеству вещественных чисел, а пополнение его по p-адической норме — к множествам p-адических чисел (каждому простому числу p будет отвечать свое p-адическое множество).

Тогда, если вещественным числам отвечает евклидово («архимедово») пространство, то p-адическим числам — так называемые *неархимедовы пространства*.

Специфика p-адической нормы обусловливает необычность свойств неархимедовых пространств в сравнении с евклидовом пространством. Например, в неархимедовой геометрии все треугольники — равнобедренные, два разных шара не могут частично пересекаться — они либо не имеют общих точек, либо один из них целиком содержится в другом и т. д.

Наряду с чисто математической ролью p-адических чисел, основанного на них p-адического анализа, неархимедовой геометрии, начиная примерно с 70-х гг. XX века все более широко осуществляется внедрение данного аппарата в физику и другие естественные науки.

К сожалению, литература на русском языке по p-адическим числам, p-адическому анализу, их возможным применениям, небогата — это книги [46, 47], именно им, в основном, следует изложение материала, представленного выше.

О перспективах применения p-адических чисел и p-адического анализа в физике и других науках в [47] (см. с. 11) сказано следующее: «Итак, мы приходим к следующей картине. В основе фундаментальной физической теории должны лежать рациональные числа, на очень малых расстояниях важную роль должны играть p-адические числа, а на больших — вещественные. Это приводит к необходимости переработки основ математической и теоретической физики, начиная с классической механики и кончая теорией струн, с использованием теории чисел, p-адического анализа, алгебраической геометрии $\langle \ldots \rangle$ Неархимедова геометрия и p-адический анализ применяются в физике не только для описания геометрии на малых расстояниях, но и в рамках традиционной теоретической физики для описания сложных систем типа спиновых стекол или фракталов. Мы находимся только в начале p-адической математической физики. Мы надеемся, что p-адические числа найдут применение в таких областях, как теория турбулентности, биология, динамические системы, компьютеры. пробле-

мы передачи информации, криптография и других естественных науках, в которых изучаются системы с хаотическим фрактальным поведением и иерархической структурой».

К сказанному можно добавить, что спиновые стекла, как известно, являются моделью, которая привела к появлению сетей Хопфилда. Один из видов взаимосвязей фрактальной геометрии с нейросетевой тематикой был хорошо показан в лекции *Н. Г. Макаренко* на Школе 2002 года [15]. Есть уже и примеры построения p-адических нейросетей (см., например, [48, 49]).

Таким образом, даже беглый взгляд обнаруживает сразу же области пересечения p-адического анализа с теорией нейронных сетей и смежными с ней областями.

Важность математики, основанной на p-адическом анализе, для перспектив информатики вообще и нейроинформатики, в частности, обусловлена по крайней мере такими тремя причинами:

- переходом от микротехнологий к нанотехнолоигям и дальше в направлении уменьшения характерных размеров элементов электронных, оптических, опто-электронных и других устройств; еще немного [20], и эти устройства на низшем уровне «алгоритмики» будут работать на уровне атома и ниже;
- развитием исследований и разработок в области квантовых вычислений;
- возможностью использовать p-адическую топологию для описания иерархической структуры когнитивной информации.

Как уже отмечалось, p-адический подход может быть применен не только в физике для описания геометрии на малых расстояниях, но существенно шире — и в физике, и в других естественных науках как аппарат математического моделирования сложных систем. Именно к этому направлению относится материал лекции *А. Ю. Хренникова*, где основной упор делается на p-адическую топологизацию мыслительных процессов. В предлагаемой им математической модели ментальное пространство реализуется как p-адическое пространство. В каком-то смысле это — аналог классического пространства-времени. На этом ментальном конфигурационном пространстве развивается вероятностная модель мышления. Здесь предполагается, что мозг способен оперировать с вероятностными распределениями на p-

[20] Может быть эта оценка («немного») и чрезмерно оптимистична, но не считаться с такой тенденцией нельзя.

адических деревьях. В мозге такие деревья могут состоять из иерархических цепей нейронов.

В основе модели лежит фундаментальное предположение о квантовых вероятностных законах преобразования когнитивной информации. Важнейшим квантовым отличием является возможность интерференции вероятностей. Однако *А. Ю. Хренников* отнюдь не пытается свести мыслительные процессы к квантовым процессам в микромире. Более того он считает, что такая редукция в принципе невозможна. В частности, использование p-адических чисел в его когнитивной модели не имеет ничего общего с идеями об использовании этих чисел для описания пространства времени на планковских расстояниях. В течение последних лет *А. Ю. Хренников* развивал так называемый контекстуальный подход к квантовым вероятностям. В контекстуальной квантовой модели (она может быть связана с физическими, биологическими, социальными системами) интерференция вероятностей не связана напрямую с планковской шкалой.

Материалы. связанные с тематикой лекции *А. Ю. Хренникова*, в том числе по различным аспектам p-адического подхода, можно найти в библиотеках [72] и [73].

9. В лекции **Ю. И. Нечаева** «Математическое моделирование в бортовых интеллектуальных системах реального времени» обсуждается еще один подход к моделированию процессов и систем — применительно к разработке, испытаниям и функционированию в различных условиях эксплуатации бортовых интеллектуальных систем реального времени для динамических объектов (кораблей, самолетов, вертолетов).

На фоне лекций *Н. Г. Макаренко, А. А. Ежова* и *А. Ю. Хренникова*, где речь идет об использовании весьма нестандартного для современной нейроинформатики, пока еще даже «экзотического», математического аппарата, лекция *Ю. И. Нечаева* может даже показаться традиционной по направленности, по используемому аппарату.

И она действительно более традиционна в той степени, в которой можно говорить о каких-либо устоявшихся традициях в такой молодой области, как *интеллектуальное управление* динамическими системами различного назначения.

Следует уточнить смысл, в котором здесь понимается термин «интеллектуально управление». Будем трактовать его в соответствии с определением, предложенным техническим комитетом по интеллектуальному управлению Общества специалистов по системам управления (IEEE Control

Systems Society).

Согласно данному определению, наиболее общим требованием к интеллектуальным системам управления является то, что они должны быть в состоянии воспроизводить такие человеческие умения и способности, как планирование поведения, обучение и адаптация. Особенно важны в этом перечне способности к обучению и адаптации.

Интеллектуальное управление — это область комплексных, междисциплинарных исследований. В последние годы выработалась точка зрения, согласно которой область интеллектуального управления базируется, с одной стороны, на идеях, методах и средствах традиционной теории управления, а с другой — на заделе, накопленном в ряде нетрадиционных (для задач управления) областей, истоки которых лежат в различного рода психологических и биологических метафорах (искусственные нейронные сети, нечеткая логика, искусственный интеллект, генетические алгоритмы, различного рода процедуры поиска, оптимизации и принятия решений в условиях неопределенности).

В русле данного направления, активно развивающегося в последнее десятилетие, следует и лекция *Ю. И. Нечаева*, в этом и состоит ее «традиционность», упомянутая выше. В лекции продолжено рассмотрение темы, начатой на прошлой Школе (см. [50]), связанной с бортовыми интеллектуальными системами реального времени, включая такие важные вопросы, как управление динамическим объектом, идентификация экстремальных ситуаций, оценка параметров динамического объекта и внешней среды. При этом в лекции, представленной в данном сборнике, акцент смещен на задачи математического моделирования, решение которых необходимо для обеспечения управления сложными динамическими объектами в быстропротекающих процессах их взаимодействия с внешней средой. Задача математического моделирования трактуется при этом как задача работы со знаниями, включая такие ее аспекты, как формализация знаний, их извлечение, организация работы с ними. Основным при этом является комбинированный подход, объединяющий нечеткие и нейросетевые технологии, используются также подходы, основанные на принципе нелинейной самоорганизации, а также традиционные методы математического моделирования.

Целесообразность применения данной совокупности средств демонстрируется на примерах решения конкретных задач.

Дополнительные сведения по затронутым в лекции *Ю. И. Нечаева* вопросам можно получить в следующих источниках: по системам, осно-

ванным на знаниях — в [59–61]; по нечеткой логике, нечетким системам — в [55–58]; по нейросетевым технологиям — в [8–11, 60]; по смешанным технологиям мягких вычислений — в [53, 54]; по информационной обработке и управлению на основе технологий мягких вычислений — в [62–70]. Значительное число программ и публикаций по таким темам, как искусственные нейронные сети, нечеткие системы, генетические алгоритмы, а также их применениям можно найти через портал научных вычислений, адрес которого содержится в позиции [71] списка литературы к предисловию.

* * *

Как это уже было в [1–3]), помимо традиционного списка литературы каждая из лекций сопровождается списком интернетовских адресов, где можно найти информацию по затронутому в лекции кругу вопросов, включая и дополнительные ссылки, позволяющие расширить, при необходимости, зону поиска.

Кроме этих адресов, можно порекомендовать еще такой уникальный источник научных и научно-технических публикаций, как цифровая библиотека ResearchIndex (ее называют также CiteSeer, см. позицию [72] в списке литературы в конце предисловия). Эта библиотека, созданная и развиваемая отделением фирмы NEC в США, на конец 2002 года содержала около миллиона публикаций, причем это число постоянно и быстро увеличивается за счет круглосуточной работы поисковой машины.

Каждый из хранимых источников (статьи, препринты, отчеты, диссертации и т. п.) доступен в полном объеме в нескольких форматах (PDF, PostScript и др.) и сопровождается очень подробным библиографическим описанием, включающим, помимо данных традиционного характера (авторы, заглавие, место публикации и/или хранения и др.), также и большое число ссылок-ассоциаций, позволяющих перейти из текущего библиографического описания к другим публикациям, «похожим» по теме на текущую просматриваемую работу. Это обстоятельство, в сочетании с весьма эффективным полнотекстовым поиском в базе документов по сформулированному пользователем поисковому запросу, делает библиотеку ResearchIndex незаменимым средством подбора материалов по требуемой теме.

Перечень проблем нейроинформатики и смежных с ней областей, требующих привлечения внимания специалистов из нейросетевого и родственных с ним сообществ, далеко не исчерпывается, конечно, вопросами, рассмотренными в предлагаемом сборнике, а также в сборниках [1–3].

В дальнейшем предполагается расширение данного списка за счет рассмотрения насущных проблем собственно нейроинформатики, проблем «пограничного» характера, особенно относящихся к взаимодействию нейросетевой парадигмы с другими парадигмами, развиваемыми в рамках концепции мягких вычислений, проблем использования методов и средств нейроинформатики для решения различных классов прикладных задач. Не будут забыты и взаимодействия нейроинформатики с такими важнейшими ее «соседями», как нейробиология, нелинейная динамика (синергетика — в первую очередь), численный анализ (вейвлет-анализ и др.) и т.п.

Замечания, пожелания и предложения по содержанию и форме лекций, перечню рассматриваемых тем и т.п. просьба направлять электронной почтой по адресу tium@mai.ru *Тюменцеву Юрию Владимировичу*.

Литература

1. *Лекции по нейроинформатике*: По материалам Школы-семинара «Современные проблемы нейроинформатики» // III Всероссийская научно-техническая конференция «Нейроинформатика-2001», 23–26 января 2001 г. / Отв. ред. *Ю. В. Тюменцев*. – М.: Изд-во МИФИ, 2001. – 212 с.
2. Лекции по нейроинформатике: По материалам Школы-семинара «Современные проблемы нейроинформатики» // IV Всероссийская научно-техническая конференция «Нейроинформатика-2002», 23–25 января 2002 г. / Отв. ред. *Ю. В. Тюменцев*. Часть 1. – М.: Изд-во МИФИ, 2002. – 164 с.
3. Лекции по нейроинформатике: По материалам Школы-семинара «Современные проблемы нейроинформатики» // IV Всероссийская научно-техническая конференция «Нейроинформатика-2002», 23–25 января 2002 г. / Отв. ред. *Ю. В. Тюменцев*. Часть 2. – М.: Изд-во МИФИ, 2002. – 172 с.
4. *Hopfield J. J.* Neural network and physical systems with emergent collective computational abilities // *Proceedings of the National Academy of Science, USA*. – 1982. – **79**. – pp. 2554–2558.
5. *Abu-Mostafa Y. S., St. Jacques J.* Information capacity of the Hopfield model // *IEEE Trans. on Information Theory*. – 1985. – v. 31, No. 4. – pp. 461–464.
6. *Доценко Вик. С., Иоффе Л. Б., Фейгельман М. В., Цодыкс М. В.* Статистические модели нейронных сетей // В кн.: Итоги науки и техники. Серия *«Физические и математические модели нейронных сетей»*. Том 1. Часть I. *«Спиновые стекла и нейронные сети»* / Ред.: *А. А. Веденов*. – М.: ВИНИТИ, 1990. – с. 4–43.
7. *Веденов А. А., Ежов А. А., Левченко Е. Б.* Архитектурные модели и функции нейронных ансамблей // В кн.: Итоги науки и техники. Серия *«Физические и*

математические модели нейронных сетей». Том 1. Часть I. *«Спиновые стекла и нейронные сети»* / Ред.: *А. А. Веденов*. – М.: ВИНИТИ, 1990. – с. 4–43.

8. *Нейрокомпьютер как основа мыслящих ЭВМ*: Сб. науч. статей / Отв. ред. *А. А. Фролов* и *Г. И. Шульгина*. – М.: Наука, 1993. – 239 с.
9. *Горбань А. Н., Россиев Д. А.* Нейронные сети на персональном компьютере. – Новосибирск: Наука, 1996. – 276 с.
10. *Уоссерман Ф.* Нейрокомпьютерная техника: Теория и практика: Пер. с англ. – М.: Мир, 1992. – 240 с.
11. *Ежов А. А., Шумский С. А.* Нейрокомпьютинг и его приложения в экономике и бизнесе. – М.: МИФИ, 1998. – 222 с.
12. *Fonarev A., Kryzhanovsky B. V. et al.* Parametric dynamic neural network recognition power // *Optical Memory and Neural Networks*. – 2001. – v. 10, No. 4, pp. 31–48.
13. *Крыжановский Б. В., Микаэлян А. Л.* О распознающей способности нейросети на нейронах с параметрическим преобразованием частот //*ДАН (мат.-физ.)*, т. 65(2), с. 286–288 (2002).
14. *Льюнг Л.* Идентификация систем: Теория для пользователя: Пер. с англ. под ред. *Я. З. Цыпкина*. – М.: Наука, 1991. – 432 с.
15. *Макаренко Н. Г.* Фракталы, аттракторы, нейронные сети и все такое // В сб.: *«Лекции по нейроинформатике»*. Часть 2. – М.: Изд-во МИФИ, 2002. – с. 121–169.
16. *Борисович Ю. Г., Близняков Н. М., Израилевич Я. А., Фоменко Т. Н.* Введение в топологию. 2-е изд., доп. Под ред. *С. П. Новикова*. – М.: Наука, 1995. – 416 с.
17. *Милнор Дж., Уоллес А.* Дифференциальная топология: Начальный курс. Пер. с англ. *А. А. Блохина* и *С. Ю. Аракелова* под ред. *Д. В. Аносова*. – М.: Мир, 1972. – 304 с. (Серия «Современная математика: Популярная серия»)
18. *Стинрод Н., Чинн У.* Первые понятия топологии: Геометрия отображений отрезков, кривых, окружностей и кругов. Пер. с англ. *И. А. Вайнштейна* с предисл. *И. М. Яглома*. – М.: Мир, 1967. – 224 с. (Серия «Современная математика: Популярная серия»)
19. *Николис Дж.* Динамика иерархических систем: Эволюционное представление: Пер. с англ. *Ю. А. Данилова*. Предисл. *Б. Б. Кадомцева*. – М.: Мир, 1989. – 488 с.
20. *Неймарк Ю. И., Ланда П. С.* Стохастические и хаотические колебания. – М.: Наука, 1987. – 424 с.
21. *Лихтенберг А., Либерман М.* Регулярная и стохастическая динамика. – М.: Мир, 1984. – 528 с.
22. *Лоскутов А. Ю., Михайлов А. С.* Введение в синергетику. – М.: Наука, 1990. – 272 с.

23. *Анищенко В. С.* Сложные колебания в простых системах: Механизмы возникновения, структура и свойства динамического хаоса в радиофизических системах. – М.: Наука, 1990. – 312 с.

24. *Терехов С. А.* Нейросетевые аппроксимации плотности распределения вероятности в задачах информационного моделирования // В сб.: «Лекции по нейроинформатике». Часть 2. – М.: Изд-во МИФИ, 2002. – с. 94–120.

25. *Бендат Дж., Пирсол А.* Прикладной анализ случайных данных: Пер. с англ. – М.: Мир, 1989. – 540 с.

26. *Боровков А. А.* Математическая статистика: Оценка параметров, проверка гипотез. – М.: Наука, 1984. – 472 с.

27. *Королюк В. С., Портенко Н. И., Скороход А. В., Турбин В. Ф.* Справочник по теории вероятностей и математической статистике. – М.: Наука, 1985. – 640 с.

28. *Zadeh L. A.* Fuzzy logic, neural networks and soft computing // *Commun. ACM*, 1994, **37**, No. 3, pp. 77–84.

29. *Zadeh L. A.* From computing with numbers to computing with words — From manipulation of measurements to manipulation of perceptions *// IEEE Trans. on Circuits and Systems-I: Fundamental Theory and Applications.* – 1999. – v. 45, № 1. – pp.105–119.

30. Special Issue *"Uncertainty in AI"* // Communications of the ACM. – March 1995. – v. 38, No. 5. – pp. 26–57.
[*Heckerman D., Wellman M. P.* Bayesian networks. – pp. 27–30.
Burnell L., Horvitz E. Structure and chance: Melding logic and probability for software debugging. – pp. 31–41, 57.
Fung R., Del Favelo B. Applying Bayesian networks to information retrieval. – pp. 42–48, 57.
Heckerman D., Breese J. C., Rommelse K. Decision-theoretic troubleshooting. – pp. 49–57]

31. *Park Y. C., Choi K.-S.* Authomatic thesaurus construction using Bayesian networks *// Information Processing & Management*. – 1996. – v. 32, No. 5. – pp. 543–553.

32. *Luttrell S. P.* Partitioned mixture distribution: An adaptive Bayesian network for low-level image processing *// IEE Proc. Vision, Image and Signal Processing.* – 1994. – v. 141, No. 4. – pp. 251–260.

33. *McMichael D. W.* Decision-theoretic approach to visual inspection using neural networks *// IEE Proc. Vision, Image and Signal Processing.* – 1994. – v. 141, No. 4. – pp. 223–229.

34. *Liu J., Desmarais M. C.* A method of learning implication networks from empirical data: Algorithm and Monte-Carlo simulation-based validation *// IEEE Trans. on Knowledge and Data Engineering.* – Nov./Dec. 1997. – v. 9, No. 6. – pp. 990–1004.

35. Bayes net toolbox for Matlab:
URL: http://www.cs.berkeley.edu/∼murphyk/Bayes/bnt.html A Brief Introduction to Graphical Models and Bayesian Networks:
URL: http://www.cs.berkeley.edu/∼murphyk/Bayes/bayes.html

36. Квантовые нейронные сети: Материалы рабочего совещания «Современные проблемы нейроинформатики» // 2-я Всероссийская научно-техническая конференция «Нейроинформатика-2000», 19–21 января 2000 г.; III Всероссийская научно-техническая конференция «Нейроинформатика-2001», 23–26 января 2001 г. / Отв. ред. *А. А. Ежов*. – М.: Изд-во МИФИ, 2001. – 104 с.

37. *Валиев К. А., Кокин А. А.* Квантовые компьютеры: Подходы и реальность. – Ижевск: НИЦ «Регулярная и хаотическая динамика», 2001. – 352 с.

38. Квантовые вычисления: За и против: Сб. статей: Пер. с англ. – Ижевск: Издательский дом «Удмуртский университет», 1999. – 212 с. – (Б-ка «Квантовый компьютер и квантовые вычисления», том I / Гл. ред. *В. А. Садовничий*)

39. Квантовый компьютер и квантовые вычисления: Сб. статей. – Ижевск: НИЦ «Регулярная и хаотическая динамика», 1999. – 288 с. – (Б-ка «Квантовый компьютер и квантовые вычисления», том II / Гл. ред. *В. А. Садовничий*)

40. *Китаев А., Шень А., Вялый М.* Классические и квантовые вычисления. – М.: МЦНМО; ЧеРо, 1999. – 192 с.

41. *Белокуров В. В., Тимофеевская О. Д., Хрусталев О. А.* Квантовая телепортация — обыкновенное чудо. – Ижевск: НИЦ «Регулярная и хаотическая динамика», 2000. – 256 с.

42. *Стин Э.* Квантовые вычисления: Пер. с англ. *И. Д. Пасынкова*. – Ижевск: НИЦ «Регулярная и хаотическая динамика», 2000. – 112 с.

43. *Данин Д. С.* Неизбежность странного мира. – М.: Мол. гвардия, 1966. – 375 с. – (Серия «Эврика»)

44. *Пономарев Л. И.* Под знаком кванта. – М.: Наука, 1989. – 368 с.

45. *Блехман И. И., Мышкис А. Д., Пановко Я. Г.* Механика и прикладная математика: Логика и особенности приложений математики. 2-е изд., испр. и доп. – М.: Наука, 1990. – 360 с.

46. *Коблиц Н.* p-адические числа, p-адический анализ и дзета-функции. Пер. с англ. *В. В. Шокурова* под ред. *Ю. И. Манина*. – М.: Мир, 1982. – 192 с. (Серия «Современная математика: Вводные курсы»)

47. *Владимиров В. С., Волович И. В., Зеленов Е. И.* p-адический анализ и математическая физика. – М.: Наука, 1994. – 352 с.

48. *Khrennikov A., Tirozzi B.* Learning of p-adic neural networks // *Canadian Math. Soc. Proc. Series,* **29**, pp. 395–401 (2000).

49. *Albeverio S., Khrennikov A. Yu., Tirozzi B.* $p-$adic Neural Networks // *Mathematical Models and Methods in Applied Sciences*, **9**, No. 9, pp. 1417–1437 (1999).

50. *Нечаев Ю. И.* Нейросетевые технологии в бортовых интеллектуальных системах реального времени // В сб.: «Лекции по нейроинформатике». Часть 1. – М.: Изд-во МИФИ, 2002. – с. 114–163.

51. Special Issue *"Evolutionary Computations"* / Ed.: *David B. Fogel* and *Lawrence J. Fogel* // IEEE Transactions on Neural Networks. – January 1994. – v. 5, No. 1. – pp. 1–147.

52. Special Issue *"Genetic Algorithms"* / Eds.: *Anup Kumar* and *Yash P. Gupta* // Computers and Operations Research. – January 1995. – v. 22, No. 1. – pp. 3–157.

53. Special Issue *"Artificial Intelligence, Evolutionary Programming and Operations Research"* / Eds.: *James P. Ignizio* and *Laura I. Burke* // Computers and Operations Research. – June 1996. – v. 23, No. 6. – pp. 515–622.

54. Special Issue *"Neuro-Fuzzy Techniques and Applications"* Eds.: *George Page* and *Barry Gomm* // Fuzzy Sets and Systems: Intern. J. of Soft Computing and Intelligence. – Apr. 8, 1996. – v. 79, No. 1. – pp. 1–140.

55. *Заде Л.* Понятие лингвистической переменной и его применение к принятию приближенных решений: Пер. с англ. – М.: Мир, 1976. – 165 с. (Серия «Новое в зарубежной науке: Математика», вып. 3 / Ред. серии *А. Н. Колмогоров* и *С. П. Новиков*)

56. *Кофман А.* Введение в теорию нечетких множеств: Пер. с франц. – М.: Радио и связь, 1982. – 432 с.

57. *Орловский С. А.* Проблемы принятия решений при нечеткой исходной информации. – М.: Наука, 1981. – 208 с. (Серия «Оптимизация и исследование операций»)

58. Прикладные нечеткие системы / Под. ред. *Т. Тэрано, К. Асаи* и *М. Сугэно*: Пер. с япон. – М.: Мир, 1993. – 368 с.

59. *Нильсон Н.* Принципы искусственного интеллекта: Пер. с англ. – М.: Радио и связь, 1985. – 376 с.

60. Компьютер обретает разум: Пер. с англ. Под ред. *В. Л. Стефанюка*. – М.: Мир, 1990. – 240 с.

61. Будущее искусственного интеллекта / Ред.-сост. *К. Е. Левитин* и *Д. А. Поспелов*. – М.: Наука, 1991. – 302 с.

62. Special Issue *"Fuzzy Information Processing"* / Ed.: *Dan Ralescu* // Fuzzy Sets and Systems: Intern. J. of Soft Computing and Intelligence. – Feb. 10, 1995. – v. 69, No. 3. – pp. 239–354.

63. Special Issue *"Fuzzy Signal Processing"* / Eds.: *Anca L. Ralescu* and *James G. Shanahan* // Fuzzy Sets and Systems: Intern. J. of Soft Computing and Intelligence. – Jan. 15, 1996. – v. 77, No. 1. – pp. 1–116.

64. Special Issue *"Fuzzy Multiple Criteria Decision Making"* / Eds.: *C. Carlsson* and *R. Fullér* // Fuzzy Sets and Systems: Intern. J. of Soft Computing and Intelligence. – March 11, 1996. – v. 78, No. 2. – pp. 139–241.

65. Special Issue *"Fuzzy Modelling"* / Ed.: *J. M. Barone* // Fuzzy Sets and Systems: Intern. J. of Soft Computing and Intelligence. – May 27, 1996. – v. 80, No. 1. – pp. 1–120.

66. Special Issue *"Fuzzy Optimization"* / Ed.: *J.-L. Verdegay* // Fuzzy Sets and Systems: Intern. J. of Soft Computing and Intelligence. – July 8, 1996. – v. 81, No. 1. – pp. 1–183.

67. Special Issue *"Fuzzy Methodology in System Failure Engineering"* / Ed.: *Kai-Yuan Cai* // Fuzzy Sets and Systems: Intern. J. of Soft Computing and Intelligence. – Oct. 8, 1996. – v. 83, No. 2. – pp. 111–290.

68. Special Issue *"Analytical and Structural Considerations in Fuzzy Modelling"* / Ed.: *A. Grauel* // Fuzzy Sets and Systems: Intern. J. of Soft Computing and Intelligence. – Jan. 16, 1999. – v. 101, No. 2. – pp. 205–313.

69. Special Issue *"Soft Computing for Pattern Recognition"* / Ed.: *Nikhil R.Pal* // Fuzzy Sets and Systems: Intern. J. of Soft Computing and Intelligence. – Apr. 16, 1999. – v. 103, No. 2. – pp. 197–367.

70. Special Issue *"Fuzzy Modeling and Dynamics"* / Eds.: *Horia-Nicolai Teodorescu, Abraham Kandel, Moti Schneider* // Fuzzy Sets and Systems: Intern. J. of Soft Computing and Intelligence. – Aug. 16, 1999. – v. 106, No. 1. – pp. 1–97.

71. Портал научных вычислений (Matlab, Fortran, C++ и т.п.)
URL: http://www.mathtools.net/

72. NEC Research Institute CiteSeer (also known as ResearchIndex) — Scientific Literature Digital Library.
URL: http://citeseer.nj.nec.com/

73. The Archive arXiv.org e-Print archive — Physics, Mathematics, Nonlinear Sciences, Computer Science.
URL: http://arxiv.org/

Редактор материалов выпуска,
кандидат технических наук *Ю. В. Тюменцев*

E-mail: tium@mai.ru

А. А. ФРОЛОВ [1)], **Д. ГУСЕК** [2)], **И. П. МУРАВЬЕВ** [1)]
[1)] Институт высшей нервной деятельности и нейрофизиологии РАН, Москва,
[2)] Институт информатики Академии наук Чешской республики, Прага
E-mail: aafrolov@mail.ru

ИНФОРМАЦИОННАЯ ЭФФЕКТИВНОСТЬ АССОЦИАТИВНОЙ ПАМЯТИ ТИПА ХОПФИЛДА С РАЗРЕЖЕННЫМ КОДИРОВАНИЕМ

Аннотация

Аттракторная нейронная сеть хопфилдового типа исследуется аналитически и путем компьютерного моделирования. Вычисляются информационная емкость, качество воспроизведения и размер областей притяжения. Выявляются асимптотические свойства нейронных сетей путем компьютерного моделирования для больших размеров сетей (до 15000 нейронов). Показано, что существующие аналитические подходы недостаточны для случая большой разреженности. Основная причина их недостатков заключается в игнорировании различия в поведении нейронов, активных и неактивных на предшествующих тактах процесса воспроизведения. Показано, что размер областей притяжения меняется немонотонно при возрастании разреженности: вначале он возрастает, а потом — убывает. Поэтому в пределе при $p \to 0$ разреженность ухудшает способность сети к коррекции эталонов, в то же время повышая информационную емкость и качество воспроизведения. Информационная эффективность сети, достигаемая за счет коррекции искаженных эталонов, используется как обобщенная характеристика способности сети выполнять функцию автоассоциативной памяти. Существует оптимальная разреженность, для которой информационная эффективность максимальна. Оптимальная разреженность оказывается близкой к уровню нервной активности мозга.

УДК 004.032.26 (06) Нейронные сети

A. FROLOV [1)], **D. HUSEK** [2)], **I. MURAVIEV** [1)]
[1)] Institute of Higher Nervous Activity and Neurophysiology
of the Russian Academy of Sciences, Moscow
[2)] Institute of Computer Science,
Academy of Sciencies of the Czech Republik, Prague
E-mail: aafrolov@mail.ru

INFORMATIONAL EFFICIENCY OF SPARSELY ENCODED HOPFIELD-LIKE ASSOCIATIVE MEMORY

Abstract

A sparsely encoded Hopfield-like attractor neural network is investigated analytically and by computer simulation. Informational capacity, recall quality and attractor basin sizes are evaluated. Asymptotic properties of the neural network are revealed by computer simulation for a large network size (up to 15000 of neurons). This shows that all existing analytical approaches fail in the case of a large sparseness. The main reason of their invalidity is that they ignore the difference in behavior of previously active and nonactive neurons during the recall process. It is shown that the size of attraction basins changes nonmonotonically while sparseness increases: it increases initially and then decreases. Thus at the limit case for $p \to 0$ sparseness worsens the network ability to correct distorted prototypes while it improves both informational capacity and recall quality. The gain of information provided by the network due to correction of the distorted prototypes is used as cumulative index of the network ability to perform the functions of autoassociative memory. There exists an optimal sparseness for which the gain is maximal. Optimal sparseness obtained is corresponded to brain neural activity.

Введение

Исследуемая нейронная сеть хопфилдового типа является полносвязной сетью, которая действует как автоассоциативная память для статистически независимых бинарных паттернов на основе корреляционного правила обучения Хебба. Кодирование называется *разреженным*, если количество активных нейронов n в записанных паттернах (эталонах) мало по сравнению с общим количеством нейронов в сети N. Типичные вопросы, возникающие при исследовании свойств сетей такого типа:

1. Как много эталонов может быть записано и воспроизведено в таких сетях? Это проблема *информационной емкости*.
2. Какова доля ошибок в финальных (выходных) паттернах по сравнению с воспроизводимыми эталонами? Это проблема *качества воспроизведения*.
3. Как сильно эталоны могут быть искажены при сохранении свойства воспроизводимости? Это проблема *размера областей притяжения* устойчивых состояний (аттракторов) в окрестности эталонов.

Важный теоретический вклад в оценку емкости сети и качества воспроизведения сетей хопфилдового типа с разреженным кодированием сделан методами статистической физики (метод реплик) (см. [2, 6, 10, 15, 24, 26] и др.). Было показано, что разреженность приводит к возрастанию емкости сети, причем не только если она определяется количеством запомненных эталонов (что является слабым результатом, потому что энтропия каждого эталона уменьшается при разрежении), но это верно даже если рассматривается *информационная* емкость, которая определяется общей *энтропией* записанных эталонов. Например, в предельном случае, когда $p = n/N \to 0$, эта энтропия была оценена в работе [2] как $(N^2/2)\log_2 e \simeq 0.72N^2$, в то время, как для оригинальной сети Хопфилда с плотным кодированием ($p = 0.5$) эта энтропия составляет только около $0.14N^2$ [5, 14].

Качество воспроизведения сетей хопфилдового типа обычно измеряется коэффициентом корреляции между эталонами и стабильными паттернами в их окрестностях. Эти паттерны являются аттракторами сетевой нейродинамики и рассматриваются как выходные (финальные) паттерны в процедуре декодирования (воспроизведения). Хорошо известно, что при относительно малом числе записанных эталонов (малой информационной нагрузке) качество воспроизведения сетей хопфилдового типа довольно высоко (корреляция между выходными и записанными паттернами близка к 1), но качество воспроизведения резко падает, когда информационная нагрузка близка к насыщению. Для оригинальной сети Хопфилда с плотным кодированием оценка корреляции между выходными паттернами и эталонами при максимальной нагрузке, сделанная методом реплик, дает 0.968 [5]. Ее оценка, сделанная тем же методом в работе [10] для сети с разреженным кодированием, показывает, что качество воспроизведения для сети с разреженным кодированием меняется немонотонно при возрастании разреженности. Вначале оно убывает, и коэффициент корреляции между финальными паттернами и эталонами при максимальной нагрузке достигает 0.82 для

$p \simeq 0.02$, а потом возрастает и стремится к 1, когда $p \to 0$. Таким образом, в пределе высокой разреженности качество воспроизведения и информационная емкость выше, чем в случае плотного кодирования. Однако, при увеличении разреженности они возрастают очень медленно. Например, качество воспроизведения при максимальной нагрузке достигает величины 0.9 только при p около 10^{-5} и величины 0.95 только при p около 10^{-44}.

Почти ничего не известно о влиянии разреженности на размер областей притяжения вокруг эталонов, что является основной характеристикой способности сети к коррекции эталонов после их частичного искажения. Анализ этого размера требует исследования динамики сети, что довольно трудно проделать точно. Трудности возникают из-за наличия корреляции между историей активности сети и матрицей связей. Исследование первых нескольких шагов по времени было проведено *Gardner et al.* [12] для сети с плотным кодированием и параллельной (синхронной) динамикой. Однако, сложность их расчетов экспоненциально возрастает в зависимости от числа рассматриваемых временных шагов, поэтому практически приходится ограничиваться малым числом шагов. Исследование этой ранней стадии нейродинамики часто оказывается недостаточным для заключения о размере областей притяжения, так как начальное возрастание корреляции между текущим паттерном активности сети и воспроизводимым эталоном может наблюдаться даже вне области притяжения (см. [4] и многие другие). С другой стороны, начальное убывание корреляции может наблюдаться внутри области притяжения (см. раздел «Компьютерное моделирование» настоящей статьи, с. 48–64). *Amari* & *Maginu* [4] предложили теорию статистической нейродинамики (Statistical Neurodynamics, SN) с целью исследования долговременного поведения активности сети с параллельной динамикой. Они вывели систему рекурсивных уравнений для двух макропараметров: 1) корреляции между текущим паттерном нейронной активности и воспроизводимым эталоном и 2) дисперсии шума. Их оценки информационной емкости и качества воспроизведения для сети Хопфилда с плотным кодированием оказались очень близки к оценкам, полученным методом реплик, а их оценки размера областей притяжения качественно совпадают с результатами компьютерного моделирования. *Okada* [22] обобщил SN-метод *Amari–Maginu* и вывел рекурсивные уравнения для более чем двух макропараметров. Он показал, что возрастание порядка системы рекурсивных уравнений делает оценки информационной емкости все более близкими к получаемым методом реплик. *Coolen* & *Sherington* [7] предложили более изощренный и точный теоретический подход к анализу параметров сети

в случае асинхронной (последовательной) динамики, свободный от предположения, сделанного в теориях *Amari-Maginu* и *Okada*, что распределение синаптических возбуждений нейронов является гауссовским. Недавно *Koyama et al.* [21] предложили разумное приближение к теории, построенной в работе [12], и вывели нейродинамические уравнения для полного процесса воспроизведения с параллельной динамикой. Как и указанная теория, этот метод дает достаточно точную оценку информационных и динамических свойств сети Хопфилда с неразреженным кодированием.

Имеются лишь несколько статей, в которых исследуется размер областей притяжения в сетях хопфилдового типа с разреженным кодированием. *Horner et al.* [16] исследовали их для асинхронной, а *Okada* [22] — для синхронной динамики. Метод, предложенный в работе [16], основан на интерполяции между первыми шагами воспроизведения (когда может быть сделано разложение по степеням шага воспроизведения) и последними шагами воспроизведения (когда может использоваться стационарное решение). *Okada* использовал SN-приближение, близкое к предложенному *Amari-Maginu* для плотного кодирования. К сожалению, влияние разреженности на размер области притяжения систематически в этих статьях не анализировалось.

Обобщенной характеристикой способности сети выполнять функцию автоассоциативной памяти является информация I_g, извлекаемая из сети за счет коррекции искаженных записанных эталонов. Следуя *Дунину-Барковскому* [1], относительную информацию, извлекаемую из сети $E = I_g/N^2$, где N^2 — общее количество элементов, способных к модификации (для хопфилдовой сети — число синаптических связей), мы называем *информационной эффективностью*. Информация, извлеченная из сети, не может, очевидно, превысить информацию, загруженную в сеть $I_l = LNh(p)$, где L — количество записанных эталонов, и $h(p) = -p\log_2 p - (1-p)\log_2(1-p)$ — функция Шеннона. Как относительную информационную нагрузку мы используем $\alpha = I_l/N^2 = Lh(p)/N$. Эта мера информационной нагрузки учитывает, что возрастание разреженности приводит к убыванию энтропии эталонов, в отличие от параметра α, определенного как L/N, который часто используется в качестве меры нагрузки сети, начиная с появления статьи Хопфилда [14]. В случае плотного кодирования обе эти меры совпадают, так как $h(0.5) = 1$. Для разреженного кодирования мы предпочитаем учитывать энтропию эталонов при расчете информационной нагрузки по двум причинам. Во-первых, в этом случае относительная информационная нагрузка измеряется в тех же единицах (бит/синапс), что и

информационная эффективность, и поэтому эти две величины легко сравнивать. Во-вторых, информационная емкость, измеряемая в битах на синапс, стремится к конечной величине $(\log_2 e)/2$ [10], когда разреженность возрастает, в то время как емкость, определенная как L/N, расходится как $1/|p\ \log_2 p|$.

Информационную эффективность можно сравнить с ее верхним пределом, задаваемым энтропией матрицы связей. Для полносвязной нейронной сети, обучаемой по правилу Хебба, верхний предел информационной эффективности был оценен как $E_{lim} = (\log_2 e)/4 \simeq 0.36$ [11, 23]. Для плотного кодирования максимальная информационная эффективность значительно меньше и составляет только 0.092 [16]. Однако, она возрастает при возрастании разреженности и, например, для $p = 0.1$ оценивается величиной 0.143 [16].

Основная цель настоящей лекции заключается в оценке информационной эффективности сети хопфилдового типа при разреженном кодировании. Она требует оценки всех трех упомянутых выше характеристик эффективности сети: информационной емкости, качества воспроизведения и размера областей притяжения. В статье [10] мы исследовали только информационную емкость и качество воспроизведения. В настоящей лекции мы исследуем также размер областей притяжения. Исследование производилось аналитически и путем компьютерного моделирования. В качестве аналитической техники мы использовали статистическую нейродинамику [4] и одношаговое (Single Step, SS) приближение [18]. Система рекурсивных уравнений, используемая здесь для оценок размеров областей притяжения, была выведена в работах [9] и [10]. К сожалению, мы не могли сравнить ее с системой уравнений, независимо выведенных *Shigemitsu & Okada* [25], так как их статья опубликована на японском языке. На основе результатов, полученных для плотного кодирования, может возникнуть сомнение в точности SS-приближения [22], которое игнорирует корреляции между историей активности сети и матрицей связей. Однако, как показано в работе [10], различие между результатами, полученными *методом реплик* (RM), SN и SS-приближениями убывает при возрастании разреженности. Более того, результаты компьютерного моделирования (см. раздел «Компьютерное моделирование», с. 48–64) показывают, что при разреженном кодировании результаты SS-приближения дают даже лучшее согласие с результатами компьютерного моделирования, чем методы RM и SN. Так как результаты, полученные с помощью SS-приближения, легко интерпретируются и справедливы по крайней мере для первых шагов нейродинамики,

мы сохранили в наших исследованиях эту «наивную» [22] технику, как один из используемых аналитических подходов.

Большинство результатов, представленных в этой лекции, получено путем компьютерного моделирования. Результаты показывают, что все существующие пока аналитические подходы дают только качественную оценку параметров сети хопфилдового типа с разреженным кодированием и параллельной динамикой. Основной недостаток существующих аналитических методов заключается в игнорировании различия в поведении нейронов, активных и неактивных в предшествующих состояниях динамики. Различие между результатами компьютерного моделирования и аналитическими становится очевидным только при достаточно больших размерах сети. Например, для сети Хопфилда с плотным кодированием количество нейронов в моделируемой сети должно превышать, по крайней мере, несколько тысяч (в моделировании, проведенном в работах [19] и [20] оно достигало 10^5). Для моделирования таких больших сетей *Коринг* [19] разработал специальный алгоритм, позволяющий обойтись без записи матрицы связей в компьютерную память. Мы применили алгоритм Коринга для сетей с разреженным кодированием и дополнительно модифицировали его, избежав необходимости записи в компьютерную память полного набора запомненных эталонов. Поэтому, для моделировании сети большого размера мы были ограничены только временем счета. В наших компьютерных экспериментах размер сети достигал 15000 нейронов.

В последующих разделах лекции мы опишем вначале модель и технику, используемую в компьютерном моделировании (с. 34–39), далее выведем основные уравнения для вычисления информационной эффективности (с. 39–43), затем рассмотрим аналитические подходы (с. 43–48). Результаты проведенного компьютерного моделирования сравниваются с аналитическими результатами (с. 48–64), после чего используются для оценки информационной эффективности (с. 64–68). Завершается лекция кратким обсуждением результатов (с. 68–69).

Описание модели

Как и в работе [10], мы использовали для обучения сети *корреляционное правило Хебба*, более эффективное, чем обычное правило Хебба [6, 26]. Правило Хебба называется корреляционным, если компоненты матрицы

связей определяются уравнением

$$J_{ij} = \frac{1}{Np\,(1-p)} \sum_{l=1,L} (X_i^l - p)(X_j^l - p), \quad i \neq j, \quad J_{ii} = 0\,, \tag{1}$$

где $X_i^l \in \{0,1\}$ — компоненты эталонов (1 — для активного, 0 — для неактивного нейронов), $p = n/N$ — вероятность нейрона быть активным и L — количество записанных эталонов. В нашей модели эталоны предполагаются равномерно и независимо распределенными на множестве всех паттернов с $n = pN$ компонентами 1 и $(N-n)$ компонентами 0 (т. е. количество активных нейронов в каждом эталоне фиксировано. Мы показали в работе [10], что такой тип кодирования обеспечивает возрастание информационной емкости по сравнению с кодированием, когда количество активных нейронов в эталонах случайно и только их среднее количество равно n.

Мы ограничились анализом только случая параллельной динамики, когда паттерн нейронной активности $\mathbf{X}$ в каждый момент времени t вычисляется как

$$X_i(t+1) = \Theta(\eta_i(t) - T(t)), \quad i = 1, \ldots, N\,, \tag{2}$$

где

$$\eta_i(t) = \sum_j J_{ij} X_j(t) \tag{3}$$

представляет собой синаптическое возбуждение, Θ — пороговая функция Хевисайда и T — порог активации. Порог $T(t)$ выбирается в каждый момент времени так, чтобы количество активных нейронов было бы равно $n = pN$. Таким образом, в каждый такт времени активны только n нейронов, имеющих наибольшее синаптическое возбуждение. Поскольку количество активных нейронов в каждом эталоне фиксировано и тоже равно n, этот выбор порога активации способствует стабилизации активности сети в окрестности одного из эталонов. Следовательно, тип кодирования эталонов согласуется с используемой простой процедурой воспроизведения, которая позволяет обойтись без точного контроля порога активации. Это второе преимущество кодирования с фиксированным числом активных нейронов в эталонах. Для того, чтобы исключить конкуренцию между нейронами, имеющими одинаковые синаптические возбуждения, равные порогу активации, к возбуждениям добавлен шум χ_i, различный для всех

нейронов сети, величина которого столь мала, что не меняет распределения нейронов по величине синаптического возбуждения от активных нейронов сети, но достаточна, чтобы упорядочить их по величине суммарного возбуждения, если вклады от активных нейронов сети одинаковы. Для того, чтобы обеспечить стабилизацию активности сети, распределение шума χ_i по нейронам полагается фиксированным во все моменты времени. Тогда, как и в случае, когда порог активации фиксирован [13], в нейронной динамике существуют только два типа аттракторов — точки или циклы длины 2.

Для доказательства введем $\eta'_i(t) = \eta_i(t) + \chi_i$, где $\eta_i(t)$ — синаптические возбуждения от активных нейронов сети, а χ_i — синаптический шум. Тогда все значения $\eta'_i(t)$ различны, даже если некоторые значения $\eta_i(t)$ равны. Введем порог $T'(t)$ такой что n нейронов с $\eta'_i(t) > T'(t)$ активны и $N - n$ нейронов с $\eta'_i < T'$ — неактивны. В этом случае уравнение (2) заменяется на

$$X_i(t+1) = \Theta(u_i(t))\,, \tag{4}$$

где $u_i(t) = \eta'_i(t) - T'(t)$.

Введем функцию $F(t) = \mathbf{X}^T(t+1)\mathbf{J}\mathbf{X}(t) + \left[\mathbf{X}^T(t+1) + \mathbf{X}^T(t)\right]\chi$, где χ — вектор с компонентами χ_i. Поскольку число активных нейронов фиксировано на каждом временном шаге, $\mathbf{X}^T(t+2)\,\mathbf{e} = \mathbf{X}^T(t)\,\mathbf{e} = n$, где $\mathbf{e}$ — вектор из N единиц. Следовательно,

$$\begin{aligned}\Delta = F(t+1) - F(t) &= \left[\mathbf{X}^T(t+2) - \mathbf{X}^T(t)\right]\left[\mathbf{J}\mathbf{X}(t+1) + \chi\right] = \\ &= \left[\mathbf{X}^T(t+2) - \mathbf{X}^T(t)\right]\left[\mathbf{J}\mathbf{X}(t+1) + \chi - T'(t+1)\,\mathbf{e}\,\right] = \sum_i \delta_i\,,\end{aligned} \tag{5}$$

где $\delta_i = [X_i(t+2) - X_i(t)]u_i(t+1)$. Величины δ_i неотрицательны, поскольку, если $u_i(t+1) > 0$, то, в соответствии с (4), $X_i(t+2) = 1$ и $X_i(t+2) - X_i(t) \geqslant 0$ независимо от $X_i(t)$. Если $u_i(t+1) < 0$, то согласно (4), $X_i(t+2) = 0$ и $X_i(t+2) - X_i(t) \leqslant 0$ независимо от $X_i(t)$. Следовательно, в обоих случаях $\delta_i \geqslant 0$ и, следовательно, $\Delta \geqslant 0$. Поскольку количество состояний сети конечно, через конечное число шагов Δ достигает нулевого значения, когда все $\delta_i = 0$. Так как в используемой процедуре восстановления $u_i \neq 0$ для всех i, это возможно только если $X_i(t+2) = X_i(t)$ для всех i. Таким образом, в процессе восстановления достигается неподвижная точка или цикл длины два. Выходным (финальным) паттерном $\mathbf{X}^f$ процедуры воспроизведения полагается стабильное состояние сети или первый паттерн цикла длины 2.

Сходство между текущим паттерном нейронной активности $\mathbf{X}(t)$ и восстанавливаемым эталоном $\mathbf{X}^l$ задается так называемым *пересечением* (overlap) этих паттернов:

$$m(\mathbf{X}^l, \mathbf{X}(t)) = \sum_{i=1,N} (X_i^l - p) X_i(t) / \left[Np\,(1-p) \right]. \tag{6}$$

Качество воспроизведения измеряется средним пересечением между восстанавливаемым эталоном и финальным паттерном $m_f = m(\mathbf{X}^l, \mathbf{X}^f)$. Так как средняя активность нейрона равна p, а ее дисперсия равна $p\,(1-p)$, то среднее пересечение фактически равно коэффициенту корреляции между паттерном нейронной активности и восстанавливаемым эталоном.

Для компьютерного моделирования сетей большого размера при ограничениях на компьютерную память мы используем процедуру, близкую к процедуре, предложенной в [19] для плотного кодирования. Эта процедура позволяет моделировать работу сети без записи матрицы связей. С использованием (1)–(3) можно легко выразить синаптическое возбуждение прямо через пересечения между текущими паттернами активности сети и записанными эталонами:

$$\begin{aligned} \eta_i(t) &= \frac{1}{Np\,(1-p)} \sum_{l=1,L} (X_i^l - p) \sum_{j \neq i} (X_j^l - p) X_j(t) = \\ &= \sum_{l=1,L} \left[(X_i^l - p)\, m(\mathbf{X}^l, \mathbf{X}(t)) - \frac{1}{Np\,(1-p)} (X_i^l - p)^2 X_i(t) \right] = \\ &= A_i - \frac{1-2p}{Np(1-p)} B_i - p\,C - \frac{LpX_i}{N(1-p)}, \end{aligned} \tag{7}$$

где

$$\begin{aligned} A_i &= \sum_{l=1,L} X_i^l m(\mathbf{X}^l, \mathbf{X}(t)), \\ B_i &= \sum_{l=1,L} X_i^l X_i(t), \\ C &= \sum_{l=1,L} m(\mathbf{X}^l, \mathbf{X}(t)). \end{aligned}$$

Для разреженного кодирования более эффективно представлять записанные эталоны как векторы длины n, компонентами которых являются

номера активных нейронов вместо векторов длины N с компонентами состояния (0 или 1) всех нейронов. Тогда более удобно представлять пересечения в виде

$$m(\mathbf{X}^l, \mathbf{X}(t)) = \frac{1}{Np\,(1-p)}\Big[\sum_{i=1,N} X_i^l X_i(t) - Np^2\Big] = \\ = \frac{1}{Np\,(1-p)}\Big[\sum_{j=1,n} X_{i(j)}(t) - Np^2\Big],$$

где $i(j)$ — номер активного нейрона, задаваемый вектором, представляющим l-тый эталон. Поэтому требуется только n сложений для вычисления каждого пересечения и Ln сложений для вычисления всех пересечений. Аналогично, для вычисления всех N величин A_i и B_i требуется только $2Ln$ сложений и в целом требуется примерно $3Ln$ сложений для вычисления всех N величин синаптических возбуждений η_i. Заметим, что для обычной процедуры, основанной на прямом вычислении η_i согласно (3), необходимо nN сложений. Таким образом, для малых информационных нагрузок ($L < N/3$) модифицированная процедура имеет преимущество перед обычной даже по числу вычислений.

Описанная процедура позволяет обойтись без записи матрицы связей, но не набора эталонов. Однако компьютерная память, требуемая для записи матрицы связей и требуемая для записи набора эталонов, сравнимы. Фактически, для записи около $N^2/2$ независимых коэффициентов матрицы связей как действительных чисел требуется $2N^2$ байтов памяти (если положить, что для записи каждого числа требуется 4 байта). В то же время для записи Ln индексов активных нейронов в эталонах как переменных типа «слово» требуется $2Ln$ байтов (если положить, что для записи слова требуется 2 байта). Для разреженного кодирования мы имеем $L \simeq \alpha N/|p \log_2 p|$. Поэтому, память, требуемая для записи набора эталонов, по порядку только в $|\log_2 p|$ раз меньше, чем требуемая для записи матрицы связей. Чтобы обойтись без записи эталонов, мы использовали процедуру, в которой эталоны заново генерируются на каждом шаге воспроизведения, с помощью одной и той же последовательности псевдослучайных величин. Для генерации L эталонов требуется Ln псевдослучайных величин. Для этого требуется примерно то же компьютерное время, что и для вычисления синаптических возбуждений по уравнению (7). Таким образом, при моделировании работы сети можно обойтись не только без записи в компьютерную память

матрицы связей, но и без записи набора эталонов, оплачивая эту экономию памяти уменьшением скорости счета примерно в два раза.

Информация, извлекаемая из сети за счет коррекции искаженных эталонов

Для вычисления количества информации, извлекаемой из памяти за счет коррекции искаженных эталонов, мы применяем основные понятия теории информации. *Взаимная информация* случайных величин A и B вычисляется по формуле

$$I(A,B) = H(A) - H(A|B) = H(B) - H(B|A)\,.$$

Взаимная условная информация случайных величин A и B при условии, что задана случайная величина C, вычисляется по формуле

$$I(A,B|C) = H(A|C) - H(A|B,C)\,,$$

где

$$H(A) = -\sum_{A} \mathcal{P}\{A\} \log_2 \mathcal{P}\{A\}$$

есть энтропия случайной величины A и

$$H(A|B) = -\sum_{A,B} \mathcal{P}\{A,B\} \log_2 \mathcal{P}\{A|B\}$$

есть условная энтропия случайной величины A.

Информация, полученная из сети за счет коррекции входного паттерна, задается выражением

$$i_g = I(\mathbf{X}^l, \mathbf{X}^f|\mathbf{X}^{in}) = H(\mathbf{X}^l|\mathbf{X}^{in}) - H(\mathbf{X}^l|\mathbf{X}^{in}, \mathbf{X}^f)\,, \tag{8}$$

где $\mathbf{X}^{in}$ и $\mathbf{X}^f$ — начальный и финальный паттерны в процессе воспроизведения эталона $\mathbf{X}^l$, соответственно. В (8) первый член дает величину информации, требуемой для нахождения эталона, когда известен лишь начальный паттерн (начальная неопределенность восстанавливаемого эталона), а второй — остаточную неопределенность, когда становится известным

также финальный паттерн. Их разность есть информация, извлеченная из сети в процессе воспроизведения.

Для нахождения $H(\mathbf{X}^{in}|\mathbf{X}^{l})$ введем вероятности $q_\mu = \mathcal{P}\{X_i^{in} = 1|X_i^l = \mu\}$ того, что нейроны, принадлежащие ($\mu = 1$) и не принадлежащие ($\mu = 0$) восстанавливаемому эталону $\mathbf{X}^l$ активны в начальном паттерне $\mathbf{X}^{in}$. Поскольку количество активных нейронов в начальном паттерне совпадает с количеством активных нейронов в эталоне, то $H(\mathbf{X}^l|\mathbf{X}^{in}) = H(\mathbf{X}^{in}|\mathbf{X}^l) = Nh_{in}$, где

$$h_{in} = ph(q_1) + (1-p)h(q_0)\,. \tag{9}$$

Два члена в (9) дают количество информации (в пересчете на один нейрон), требуемой для нахождения нейронов, принадлежащих эталону, среди активных и неактивных нейронов начального паттерна, соответственно.

Поскольку количество активных нейронов в начальном паттерне полагается равным $n = pN$, то

$$p = q_1 p + q_0(1-p) \tag{10}$$

и, следовательно, начальное пересечение определяется по формуле

$$m(\mathbf{X}^l, \mathbf{X}^{in}) = m_{in} = q_1 - q_0 = (q_1 - p)/(1-p)\,. \tag{11}$$

Для нахождения $H(\mathbf{X}^l|\mathbf{X}^{in}, \mathbf{X}^f)$ введем вероятности

$$p_{\mu\nu} = \mathcal{P}\{X_i^f = 1|X_i^l = \mu, X_i^{in} = \nu\}$$

того, что данный нейрон, который *принадлежит* ($\mu = 1$) или *не принадлежит* ($\mu = 0$) воспроизводимому эталону и *активен* ($\nu = 1$) или *не активен* ($\nu = 0$) в начальном паттерне, активен в финальном паттерне. Тогда

$$P_{11} = \mathcal{P}\{X_i^{in} = 1, X_i^f = 1\} = p_{11}q_1 p + p_{01}q_0(1-p)$$

есть вероятность того, что данный нейрон активен в начальном и финальном паттернах,

$$P_{10} = \mathcal{P}\{X_i^{in} = 1, X_i^f = 0\} = (1-p_{11})\,q_1 p + (1-p_{01})q_0(1-p)$$

вероятность того, что он активен в начальном и не активен в финальном паттерне,

$$P_{01} = \mathcal{P}\{X_i^{in} = 0, X_i^f = 1\} = p_{10}(1-q_1)p + p_{00}(1-q_0)(1-p)$$

вероятность того, что он не активен в начальном и активен в финальном паттерне, и

$$P_{00} = \mathcal{P}\{X_i^{in} = 0, X_i^f = 0\} = \\ = (1 - p_{10})(1 - q_1)p + (1 - p_{00})(1 - q_0)(1 - p)$$

вероятность того, что он не активен в начальном и не активен в финальном паттерне.

Тогда $H(\mathbf{X}^l|\mathbf{X}^{in}, \mathbf{X}^f) = Nh_f$, где

$$h_f = P_{11}h(p_{11}q_1p/P_{11}) + P_{10}h((1 - p_{11})q_1p/P_{10}) + \\ + P_{01}h\big[p_{10}(1 - q_1)p/P_{01}\big] + P_{00}h\big[(1 - p_{10})(1 - q_1)p/P_{00}\big]\,. \tag{12}$$

Четыре члена в (12) дают количества информации, которые требуются, соответственно, для нахождения нейронов, принадлежащих записанному паттерну среди нейронов, *активных* и в начальном и в финальном паттернах, *активных* в начальном и *не активных* в финальном паттерне, *не активных* в начальном и *активных* в финальном паттерне и *не активных* и в начальном и в финальном паттернах.

Так как количество активных нейронов в финальном паттерне полагается равным $n = pN$, то

$$p = p_{11}q_1p + p_{01}q_0(1 - p) + p_{10}(1 - q_1)p + p_{00}(1 - q_0)(1 - p)\,. \tag{13}$$

Полезно также ввести вероятности

$$p_1 = p_{11}q_1 + p_{10}(1 - q_1), \quad p_0 = p_{01}q_0 + p_{00}(1 - q_0)$$

того, что нейрон, принадлежащий или не принадлежащий восстанавливаемому эталону, активен в финальном паттерне. Так как количество активных нейронов в финальном паттерне равно Np, то

$$p = p_1p + p_0(1 - p)\,. \tag{14}$$

Финальное пересечение $m(\mathbf{X}^l, \mathbf{X}^f)$, которое равно

$$m_f = p_1 - p_0 = (p_1 - p)/(1 - p)\,, \tag{15}$$

может быть легко выражено через вероятности $p_{\mu\nu}$.

Если эталон восстановлен без ошибок, т. е. $p_1 = p_{11} = p_{10} = 1$ и $p_0 = p_{01} = p_{00} = 0$, то $H(\mathbf{X}^l|\mathbf{X}^{in}, \mathbf{X}^f) = 0$ (дополнительной информации об эталоне не требуется, потому что он полностью определен финальным паттерном). Однако, даже в этом случае выигрыш информации не равен энтропии эталона, потому что начальный паттерн уже содержит некоторую информацию о нем. В соответствии с (10) и (9), эта информация

$$\begin{aligned} I(\mathbf{X}^l, \mathbf{X}^{in}) &= H(\mathbf{X}^l) - H(\mathbf{X}^l|\mathbf{X}^{in}) = H(\mathbf{X}^{in}) - H(\mathbf{X}^{in}|\mathbf{X}^l) = \\ &= N[h(p) - p\,h(q_1) - (1-p)h(p\,(1-q_1)/(1-p))]\,. \end{aligned}$$

Она мала, только когда высок уровень искажения эталонов. Она равна нулю, когда входной паттерн статистически независим от эталона (т. е. $q_1 = p$). Однако, в большинстве статей (например, [3,6,24]) эта информация игнорируется. Это приводит к переоценке количества информации, извлекаемой из сети. Это количество, очевидно, равно нулю, если начальный паттерн совпадает с восстанавливаемым эталоном (потому что эталон уже полностью определен), в то время как игнорирование информации, данной начальным паттерном, приводит к ненулевой оценке для количества информации, извлекаемой из сети. Информация, содержащаяся относительно эталона в начальном паттерне, учитывается в наших работах (например, [11]) и в работе [16]. Аналогично, в большинстве статей (например [16]) информация о восстанавливаемом эталоне, содержащаяся в начальном паттерне, игнорируется при оценке финальной неопределенности в восстанавливаемом эталоне. То есть, вместо условной энтропии $H(\mathbf{X}^l|\mathbf{X}^{in}, \mathbf{X}^f)$ в (8) используется $H(\mathbf{X}^l|\mathbf{X}^f)$. Для совпадения этих условных энтропий достаточно выполнения двух условий. Во-первых, начальный паттерн является статистически независимым от эталона (т. е. $q_1 = q_0 = p$) и, во-вторых, активность нейронов в финальном паттерне не зависит от их активности в начальном паттерне (т. е. $p_{11} = p_{10} = p_1$ и $p_{01} = p_{00} = p_0$). Тогда $H(\mathbf{X}^l|\mathbf{X}^{in}, \mathbf{X}^f) = H(\mathbf{X}^l|\mathbf{X}^f) = Nh'_f$, где

$$h'_f = ph(p_1) + (1-p)h(p_0)\,. \tag{16}$$

Два члена в (16) определяют среднюю информацию, необходимую для нахождения нейронов, активных в эталоне, среди, соответственно, активных и неактивных нейронов в финальном паттерне. Уравнение (16) совпадает с уравнением (9), если заменить p_1 на q_1 и p_0 на q_0.

Однако, в общем случае игнорирование информации, имеющейся в начальном паттерне, ведет к недооценке количества информации, извлекаемой из сети. Например, если начальный паттерн совпадает с эталоном

(т. е. $q_1 = 1$, $q_0 = 0$) вычисление по (8) дает ожидаемое нулевое количество информации, в то время как при использовании $H(\mathbf{X}^l|\mathbf{X}^f)$ вместо $H(\mathbf{X}^l|\mathbf{X}^{in}, \mathbf{X}^f)$ нулевое количество получается, только если эталоны воспроизводятся точно (т. е. $p_1 = 1$, $p_0 = 0$). В остальных случаях оно оказывается парадоксально отрицательным.

Извлеченная из памяти информация о всех записанных эталонах $I_g = Li_g$ и, соответственно, информационная эффективность

$$E = I_g/N^2 = i_{in} - i_f\,, \tag{17}$$

где

$$i_{in} = \alpha h_{in}/h(p) \tag{18}$$

и

$$i_f = \alpha h_f/h(p)\,. \tag{19}$$

Когда информационная нагрузка мала, эталоны воспроизводятся точно и области притяжения имеют большой размер (т. е. $h_f = 0$ и $h_{in} \simeq h(p)$). Тогда информационная эффективность близка к α. Однако при возрастании информационной загрузки h_{in} убывает (из-за убывания размера областей притяжений) и h_f возрастает (из-за уменьшения качества воспроизведения). Поэтому возрастание E за счет возрастания α преодолевается убыванием $h_{in} - h_f$ и информационная эффективность начинает падать. В связи с этим, для каждого значения разреженности существует оптимальная информационная нагрузка, обеспечивающая максимальную информационную эффективность. Ее оценка в зависимости от p является основной целью настоящей лекции.

Аналитические подходы

Размер областей притяжения рассчитывается здесь с использованием двух аналитических подходов: одношагового приближения (SS) и статистической нейродинамики (SN). Соответствующие нейродинамические уравнения для нейронных сетей с разреженным кодированием были получены в работах [9] и [10].

Одношаговое приближение

Принципиальной особенностью SS-приближения [18] является то, что на каждом временном шаге нейродинамики игнорируется статистическая зависимость между активностью сети и матрицей связей и учитывается только один макропараметр нейродинамики: пересечение текущего паттерна с воспроизводимым эталоном $m(t)$. Пренебрежение статистической зависимостью справедливо только для первого шага нейродинамики, когда начальная активность сети действительно устанавливается независимо от матрицы связей. Именно поэтому данный подход называется *одношаговым приближением* или приближением первого шага. Другими, более точными аналитическими методами, а также с помощью компьютерного моделирования (см. работы [4, 5] и многие другие) было показано, что SS-приближение очень неточно в случае плотного кодирования ($p = 0.5$). Однако, оно оказалось достаточно точным для случая разреженного кодирования [10]. Более того, в разделе «Компьютерное моделирование» будет показано, что в этом случае оно дает даже лучшее соответствие с экспериментом, чем SN-приближение.

При воспроизведении эталона $\mathbf{X}^l$ синаптические возбуждения каждого нейрона на первом временном шаге выражаются, в соответствии с уравнениями (1) и (3), в виде

$$\eta_i = (X_i^l - p)m_{in} + N_i\,,$$

где m_{in} — пересечение между воспроизводимым эталоном и начальным паттерном $\mathbf{X}^{in}$ и N_i — шум, обусловленный записью в матрицу связей других эталонов. Для используемой модели [10]

$$\mathcal{M}\{N_i\} = -\alpha p/h(p), \qquad \sigma^2 = D\{N_i\} = \alpha r/h(p)\,, \tag{20}$$

где

$$r = p\,(1-p)\,. \tag{21}$$

Для больших N и L распределение N_i близко к нормальному. Тогда

$$p_{01} = p_{00} = p_0 = \Phi(\theta_0), \qquad p_{11} = p_{10} = p_1 = \Phi(\theta_1)\,, \tag{22}$$

где

$$\theta_0 = \theta + m_{in}p/\sigma, \qquad \theta_1 = \theta - m_{in}(1-p)/\sigma\,, \tag{23}$$

$$\Phi(x) = 1/(2\pi)^{1/2} \int_x^\infty \exp(-u^2/2)du$$

и θ — масштабированный порог возбуждения. В нашей модели он выбирается так, что общее количество активных нейронов остается фиксированным и равным n, т. е. θ выбирается так, чтобы выполнялось уравнение (14). В результате первого шага воспроизведения пересечение меняется на

$$m(1) = p_1 - p_0\,. \tag{24}$$

В SS-приближении эти уравнения полагаются выполняющимися для всех временных шагов (естественно, m_{in} нужно заменить на $m(t)$ и $m(1)$ на $m(t+1)$).

Финальное стабильное состояние удовлетворяет условию

$$m(t+1) = m(t) = m_f\,. \tag{25}$$

Кривые, характеризующие поведение активности сети в зависимости от α, m_{in} и p представлены на рис. 1. Пусть начальное состояние активности сети для данного α характеризуется точкой (m_{in}, α). Если эта точка лежит под кривой, пересечение между текущим и восстанавливаемым паттернами смещается в процессе воспроизведения вправо к финальному пересечению, задаваемому правой частью кривой. Поэтому данная часть кривой задает величину пересечения между эталоном и финальным стабильным состоянием в его окрестности, т. е. качество воспроизведения. Пересечение смещается влево для любой точки выше кривой. Поэтому левая часть кривой соответствует границе области притяжения. Эта часть отсутствует при $p = 0.5$. Критическое значение α_{cr}, которое дает информационную емкость сети для данного значения p, соответствует максимуму каждой кривой.

Рис. 1 показывает, что SS-приближение предсказывает монотонное убывание размера областей притяжения при возрастании разреженности т. е. при убывании p.

В предельном случае

$$p \to 0\,,$$
$$p_1 = \Phi(\theta_1) \simeq m(1), \quad p_0 = \Phi(\theta_0) \simeq p\,(1 - m(1))\,,$$
$$\theta_0 - \theta_1 \simeq \frac{m_{in}}{\sigma} = m_{in}\frac{\sqrt{\ln(1/p)}}{\sqrt{\alpha \ln 2}}\,.$$

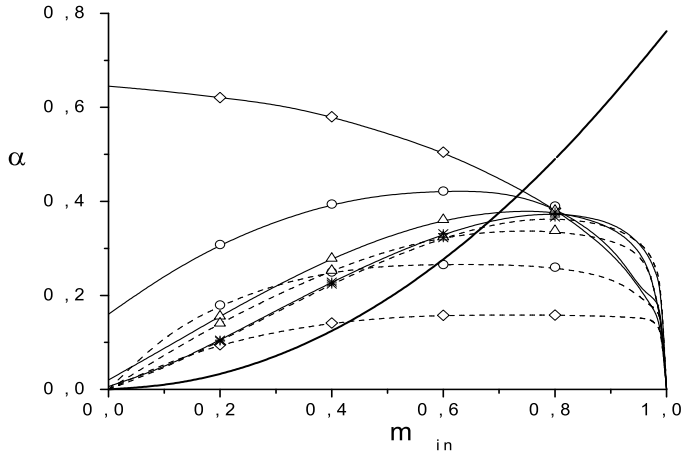

Рис. 1. Области притяжения, предсказываемые SS (тонкие сплошные линии) и SN (пунктирные линии)
Толстая сплошная линия соответствует предельному случаю большой разреженности ($p \to 0$); $\diamond$ — для $p = 0.5$, $\circ$ — для $p = 0.1$, Δ — для $p = 0.02$, $*$ — для $p = 0.004$.

Тогда

$$\theta_0 \simeq \sqrt{2\ln(1/p)}\,, \quad \theta_1 \simeq \sqrt{2\ln(1/p)}\,(1 - m_{in}/\sqrt{2\alpha\ln 2})\,.$$

Так как $\ln(1/p) \gg 1$, то

$$\theta_1 \ll -1\,, \quad \text{т. е. } m(1) \simeq 1, \text{ если } \alpha < \frac{m_{in}^2}{2\ln 2}\,,$$
$$\theta_1 \gg 1\,, \quad \text{т. е. } m(1) \ll 1, \text{ если } \alpha > \frac{m_{in}^2}{2\ln 2}\,.$$

Поэтому

$$\alpha = \frac{m_{in}^2}{2\ln 2} \tag{26}$$

дает границу области притяжения в этом предельном случае. Формула (26) дает ту же зависимость α от m_{in}, что полученная в работе [3], однако с другим коэффициентом. В этой работе границы областей притяжения соответствуют требованию, чтобы вероятность ошибки (на первом шаге) хотя бы в одном нейроне и хотя бы в одном из записанных эталонов была мала, в то время как в наших расчетах мала ошибка в каждом отдельно взятом нейроне при воспроизведении каждого отдельно взятого эталона.

Статистическая нейродинамика

Метод статистической нейродинамики был разработан *Amari & Maginu* [4] для плотного кодирования и модифицирован в работах [9,10,25] для разреженного кодирования. Принципиальной особенностью этого подхода является учет статистической зависимости между активностью сети и матрицей связей. Это достигается включением дополнительной переменной $c(t)$ в систему рекурсивных нейродинамических уравнений. В SN-приближении остаются справедливыми все уравнения SS-приближения, кроме уравнения (21), которое задает дисперсию синаптических возбуждений $\mathcal{D}\{N_i(t)\}$. В SN-приближении оно принимает следующий вид

$$\begin{aligned} r(t) &= p\,(1-p) + 2p\,(1-p)m(t)m(t-1)c(t) \\ &+c^2(t)\Big[r(t-1) + \frac{\alpha p\,(1-p)}{h(p)}(1-c^2(t-1))\Big]\,, \end{aligned} \tag{27}$$

где

$$c(t) = \frac{p\,\exp\Big[-\theta_1^2(t-1)/2\Big] + (1-p)\exp\Big[-\theta_0^2(t-1)/2\Big]}{\sqrt{2\pi\alpha r(t-1)/h(p)}}\,, \tag{28}$$

$$c(0) = 0\,.$$

Параметры конечного стабильного состояния находятся из условий

$$m(t+1) = m(t) = m(t-1) = m_f, \quad c(t+1) = c(t) = c^f\,.$$

Кривые, характеризующие размер областей притяжения в SN-приближении также показаны на рис. 1. Видно, что для $p = 0.5$ результаты SS- и SN-приближений сильно различаются, но разница убывает при возрастании разреженности. Размер областей притяжения в SN меньше, чем в

SS. Так как на первом шаге оба подхода дают одинаковые результаты, из этого следует немонотонное изменение пересечения в процессе воспроизведения, если начальная активность находится внутри области притяжения, предсказываемой SS, но вне области притяжения, предсказываемой SN. В этом случае пересечение вначале возрастает, но потом уменьшается до нуля. Такое немонотонное поведение $m(t)$ действительно наблюдалось при компьютерном моделировании (см. [4, 16, 22] и др. работы). Заметим, что отличие размеров областей притяжений, предсказываемых SS и SN, убывает, когда убывает p. Они практически неразличимы уже при $p = 0.004$.

Как показано на рис. 1, SN предсказывает немонотонную зависимость размера областей притяжения от p. Когда p убывает от 0.5 до 0.1, кривые на рис. 1 смещаются вверх и влево (что означает возрастание размера области притяжения). Когда p продолжает убывать, они смещаются вверх и вправо (что означает убывание размера области притяжения).

Компьютерное моделирование

Мы тестировали сети размером от 200 до 15000 нейронов с p от 0.02 to 0.5. Начальные пересечения были в интервале от 0.1 до 1. Общее число испытаний для каждой комбинации p, N, α и m_{in} было 2000 для $N < 3000$, 1000 для $3000 \leqslant N \leqslant 5000$ и 250 для $N \geqslant 5000$.

Плотное кодирование

Полученные распределения для пересечений m_f между финальными состояниями активности сети и восстанавливаемыми эталонами показаны на рис. 2. Общие свойства этих распределений хорошо известны (см. статьи [5, 16, 19] и многие другие). Каждое распределение состоит из двух хорошо разделенных мод. Первая мода близка к $m_f = 1$ («правильные» аттракторы) и соответствует восстановленным, а вторая мода — невосстановленным эталонам («ложные» аттракторы). Когда начальный паттерн находится достаточно далеко от границы внутри области притяжения, преобладает первая мода. Когда он далеко от границы вне области притяжения — вторая. Для каждого значения начального пересечения m_{in} существует узкая переходная область вблизи критического значения α, где обе моды имеют сходный размер. Для $m_{in} = 1$ переход происходит между $\alpha = 0.14$ и $\alpha = 0.15$, для $m_{in} = 0.5$ — между 0.12 и 0.13, для $m_{in} = 0.3$ — между

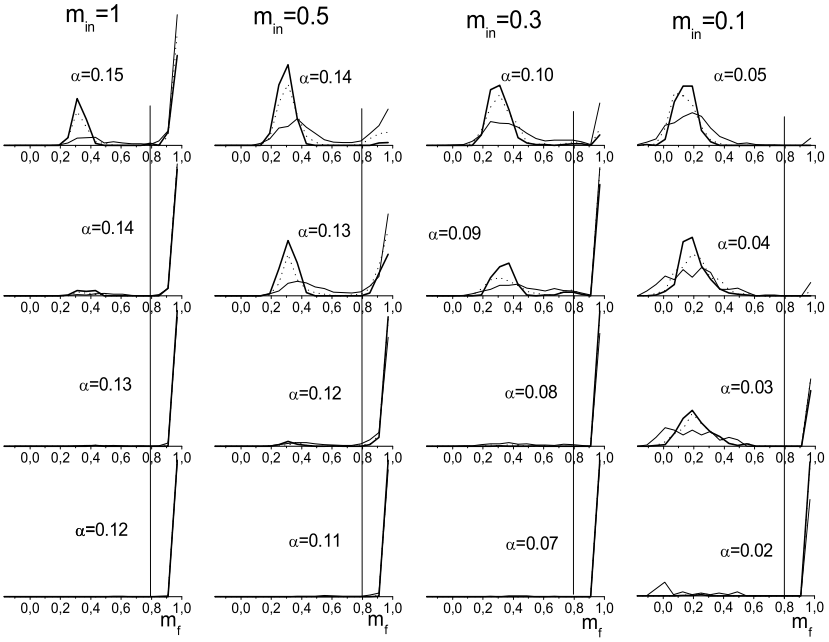

Рис. 2. Распределение финальных пересечений для $p = 0.5$ и различных значений N, α и m_{in}

Распределение финальных пересечений для $p = 0.5$ и различных значений N, α и m_{in}. Вертикальные линии соответствуют выбранным порогам для разделения траекторий, сходящихся к «правильным» ($m_f \simeq 1$) и «ложным» ($m_f \ll 1$) аттракторам. Здесь: $N = 1000$ — тонкие сплошные линии, $N = 3000$ — пунктирные линии, $N = 5000$ — толстые сплошные линии.

0.08 and 0.09, и для $m_{in} = 0.1$ — между 0.02 и 0.03. Переходная область сужается при возрастании размера сети.

Для более точного вычисления $\alpha_{cr}(m_{in})$ мы вычисляем «площадь» первой моды, т. е. вероятность $P(\alpha, m_{in}, N)$ того, что траектория, стартующая из начального состояния, имеющего пересечение с эталоном m_{in}, заканчивается около соответствующего эталона. Эта вероятность зависит от выбора границы разделения мод m_b. Хотя этот выбор в какой-то сте-

пени произволен, это не критично из-за хорошей разделенности мод. Как в [5, 10, 16], мы выбираем $m_b = 0.8$. Эта граница показана на рис. 2 вертикальными линиями. Так как P изменяется от 0 до 1, удобно [5, 8] сделать преобразование

$$P = \frac{\exp F}{1 + \exp F}\,, \tag{29}$$

где F меняется от $-\infty$ до $+\infty$. Полученные оценки F для N в области от 200 до 5000 показаны на рис. 3 светлыми знаками. Тонкие сплошные линии — результат их аппроксимации по формуле

$$F = a_0 + a_1\alpha + a_2 N(\alpha - \alpha_{cr}) + a_3 \ln N\,. \tag{30}$$

Для проверки справедливости этой аппроксимации мы дополнили наши данные на рис. 3а ($m_{in} = 1$) данными, представленными в работе [19] для больших размеров сети (N меняется до 32000) и аппроксимировали полный набор данных по (30). Данные, взятые из работы [19], показаны на рис. 3а темными знаками. Видно, что все данные достаточно единообразны и могут быть хорошо аппроксимированы выбранной формулой.

В работах [5, 8, 10, 16, 17] оценки F аппроксимировались аналогичной формулой, но без последнего (логарифмического) члена. Его влияние важно лишь тогда, когда N достаточно велико, а α очень близко к α_{cr}. Например, немонотонная зависимость P от N проявляется для $\alpha = 0.14$ только когда N достигает 10000.

Экспериментальная оценка α_{cr} соответствует толстой прямой линии, которая разделяет на рис. 3а два множества кривых. Верхние кривые вогнуты и стремятся к $+\infty$ ($P \to 1$), когда $N \to \infty$. Нижние — выпуклы и стремятся к $-\infty$ ($P \to 0$), когда $N \to \infty$. Когда α становится близким к α_{cr}, кривизна линий уменьшается. Поэтому, верхние кривые при $\alpha \simeq \alpha_{cr}$ имеют очень широкую нижнюю часть: F близка к константе для большой области значений N. Мы предсказываем (на основе используемой аппроксимации), что при возрастании N кривые в конечном счете поворачивают вверх и $P \to 1$. Но при приближении α к α_{cr} все большие значения N требуются для того, чтобы F начало возрастать.

Как *Amit et al.* [5], *Horner et al.* [16] и другие, мы рассматриваем значение α_{cr}, полученное в случае, когда начальное состояние совпадает с одним из эталонов ($m_{in} = 1$), как экспериментальную оценку информационной емкости. При совмещении наших данных и данных, полученных

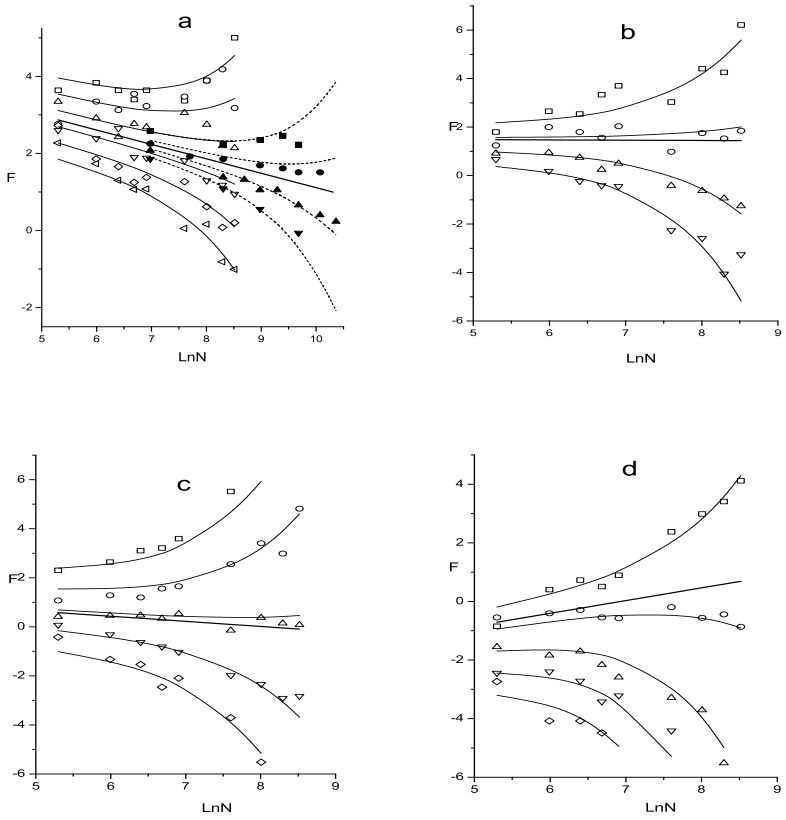

Рис. 3. Зависимость F от α и $\ln N$ для $p = 0.5$ и $m_{in} = 1$ (a), $m_{in} = 0.5$ (b), $m_{in} = 0.3$ (c) и $m_{in} = 0.1$ (d).

Светлые точки — наши результаты, темные — данные [19]. На **(a)** для наших результатов α меняется от 0.13 (верхняя кривая) до 0.155 (нижняя кривая) с шагом 0.005, для данных [19] — от 0.14 до 0.146 с шагом 0.002, на **(b)** α меняется от 0.11 до 0.14 с шагом 0.01, на **(c)** α меняется от 0.07 до 0.11 с шагом 0.01, на **(d)** α меняется от 0.02 до 0.06 с шагом 0.01. Толстые прямые линии соответствуют $\alpha = \alpha_{cr} = 0.1425$ на (a), $\alpha = \alpha_{cr} = 0.121$ на (b), $\alpha = \alpha_{cr} = 0.091$ на (c) и $\alpha = \alpha_{cr} = 0.027$ на (d).

в работе [19], $\alpha_{cr} = 0.1429 \pm 0.0003$. Если используются только наши данные, то $\alpha_{cr} = 0.1425 \pm 0.002$. Если только данные работы [19], то $\alpha_{cr} = 0.1430 \pm 0.0002$. Таким образом наши данные хорошо согласуются с данными работы [19] и дают довольно точную оценку α_{cr}. Все приведенные оценки хорошо согласуются также со значением $\alpha_{cr} = 0.1455 \pm 0.001$, полученным в работе [16] для асинхронной (последовательной) динамики, и близки (хотя и достоверно отличны) от значений $\alpha_{cr} = 0.138$ и $\alpha_{cr} = 0.158$, предсказанных RM- и SN-приближениями в [5] и [4], соответственно.

В работе [25] SN-приближение обобщено на случай, когда корреляции между шумовыми членами учитываются на нескольких шагах нейродинамики. В этом обобщении приближение первого порядка соответствует уравнениям Амари-Магину [4]. Приближение второго порядка дает более точную оценку информационной емкости: $\alpha_{cr} = 0.142$. Однако далее, когда порядок приближения возрастает, точность приближения уменьшается: (0.140, 0.139), достигая значения 0.138, предсказанного RM.

Для случая, когда начальные паттерны отличались от эталонов (рис. 3b, 3c, 3d), аппроксимация данных по формуле (30) дает: $\alpha_{cr} = 0.121 \pm 0.003$ для $m_{in} = 0.5$, $\alpha_{cr} = 0.091 \pm 0.0015$ для $m_{in} = 0.3$ и $\alpha_{cr} = 0.027 \pm 0.001$ для $m_{in} = 0.1$. Эти значения приведены на рис. 4 вместе с кривыми, полученными, полученными SN-методом; эти данные взяты из работы *Окада* [22]. Точка, построенная при $m_{in} = 0.8$, показывает полученную оценку информационной емкости. Видно, что значения, полученные нами с помощью компьютерного моделирования (кроме значений для информационной емкости), лежат между кривыми, полученными по методу Окада по приближениям второго и четвертого порядка. Это свидетельствует в пользу точности как этого метода, так и экспериментальных оценок.

Наши экспериментальные данные хорошо согласуются также с результатами компьютерного моделирования сети с асинхронной динамикой, приведенными в [16]. Однако результаты компьютерного моделирования, приведенные в [22], дают заниженную величину области притяжения. Например, при $m_{in} = 0.1$ в этой работе приведена оценка $\alpha_{cr} \simeq 0.02$ вместо 0.027, полученной в наших экспериментах. Эта разность может быть объяснена двумя причинами. Первое, в работе [22] моделировалась сеть значительно меньшего размера ($N = 1000$). Второе, в этой работе граница области притяжения оценивалась как точка, в которой среднее финальное пересечение становилось равным 0.5. Однако эта оценка является довольно грубой, так как среднее конечное пересечение определяется двумя фактора-

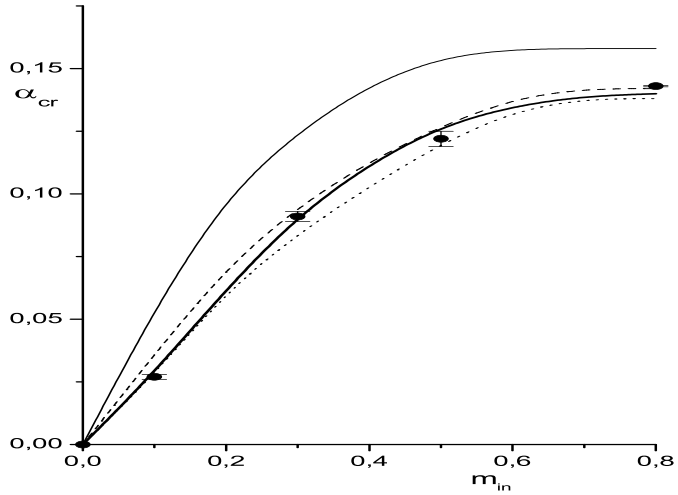

Рис. 4. Базы притяжения для $p = 0.5$
Получены SN-методом первого порядка (сплошная линия), второго порядка (штриховая линия) и четвертого порядка (пунктирная линия). Точками представлены результаты компьютерного моделирования. Результаты, полученные SN-методом второго и четвертого порядка, взяты из работы [22].

ми. Во-первых, оно зависит от перераспределения финальных пересечений между модой, близкой к единице (конечные состояния в окрестности эталонов), и модой, близкой к нулю (ложные аттракторы). Во-вторых, оно зависит от положения этих мод на оси финальных пересечений. Так как влияние этих факторов по-разному проявляется при увеличении N, нельзя экстраполировать данные, полученные для финальных пересечений при относительно малых значениях N, на случай $N \to \infty$.

На рис. 5 показаны средние значения параметров активности сети в зависимости от шага нейродинамического процесса при $\alpha = 0.09$, $m_{in} = 0.3$ и $N = 5000$. Значение m_{in} выбрано очень близко к границе области притяжения для данного α. Как показано на рис. 2 и рис. 3, при этих значениях α и m_{in} две моды конечных пересечений m_f почти равны по площади (т. е. $P \simeq 0.5$). Кривые получены усреднением отдельно траекторий, сходящихся

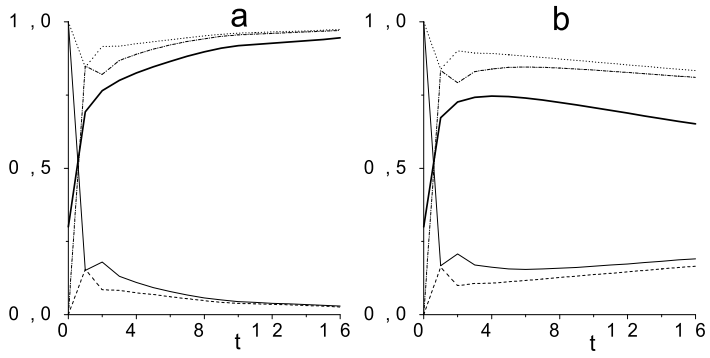

Рис. 5. Осредненные характеристики воспроизведения
Получены по траекториям, сходящимся к правильным **(a)** и ложным **(b)** аттракторам для $p = 0.5$, $m_{in} = 0.3$, $\alpha = 0.09$ and $N = 5000$. Здесь p_{11} — пунктирные линии; p_{10} — штрих-пунктирные линии; p_{01} — тонкие сплошные линии; p_{00} — штриховые линии; m — толстые сплошные линии.

к правильным (рис. 5a) и ложным (рис. 5b) аттракторам (около 250 траекторий для каждого случая). На первом шаге нейродинамики усредненные параметры для траекторий, сходящихся к правильным и ложным аттракторам, совпадают между собой и с предсказанием SS-приближения. На этом шаге пересечение с воспроизводимым эталоном возрастает. Для траекторий, сходящихся к правильным аттракторам, оно продолжает возрастать и на следующих шагах. Однако для траекторий, сходящихся к ложным аттракторам, оно возрастает только на нескольких первых шагах, а затем медленно падает. Это хорошо знакомое поведение траекторий, сходящихся к правильным и ложным аттракторам, которые начинаются внутри области притяжения, предсказанной SS-приближением. Более интересно взглянуть на поведение вероятностей $p_{\mu\nu}$, определенных в разделе «Информация, извлекаемая из сети за счет коррекции искаженных эталонов» (см. с. 41). Для плотного кодирования поведение нейронов, активных и неактивных в воспроизводимых эталонах, симметрично, т. е. $p_{01} = 1 - p_{10}$, а $p_{00} = 1 - p_{11}$. На

первом шаге, в соответствии с предсказаниями SS- и SN-приближений, вероятности p_{11} и p_{10} равны. На следующем шаге они становятся различными, затем медленно сближаются для траекторий, сходящихся к правильным аттракторам, и остаются различными для траекторий, сходящихся к ложным аттракторам. Нейроны, которые были активны в начальном паттерне, имеют большую вероятность активироваться: $p_{11} > p_{10}$, а $p_{01} > p_{00}$. Таким образом, нейроны как бы «помнят» свою предысторию в течение многих шагов нейродинамики. Этот факт игнорируется SN-приближениями всех порядков. Однако для плотного кодирования различие между p_{11} и p_{10} не так велико. Поэтому это приближение оказывается достаточно точным для плотного кодирования эталонов. Для разреженного кодирования, как показано далее, различие в поведении нейронов, активных и неактивных в начальном паттерне, более значительно.

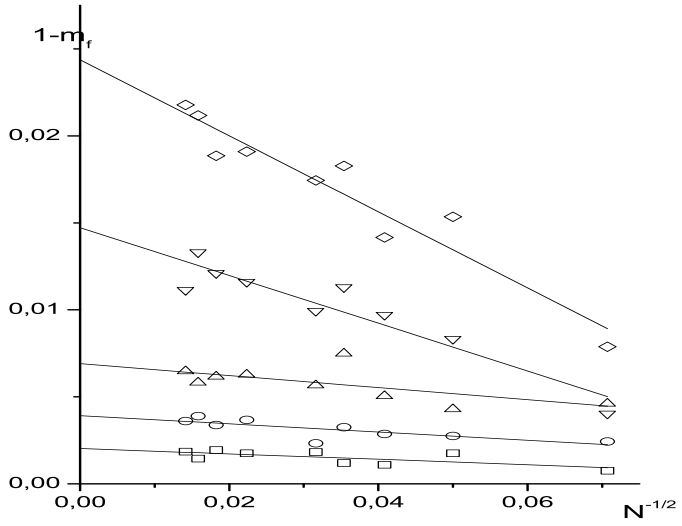

Рис. 6. Зависимость качества воспроизведения от α и N для $p = 0.5$, полученные компьютерным моделированием. Значения α меняются от 0.1 (нижняя линия) до 0.14 (верхняя линия) с шагом 0.1

Для оценки качества воспроизведения мы использовали средние значения финальных пересечений m_f для траекторий, сходящихся к правильным аттракторам из начальных состояний, совпадающих с эталонами (т. е. когда $m_{in} = 1$). Результаты показаны на рис. 6. Для каждого значения α значения средних финальных пересечений были аппроксимированы прямыми линиями в зависимости от $1/\sqrt{N}$, и их пересечение с осью ординат рассматривалось как асимптотическая оценка качества воспроизведения для данного значения α. Полученные значения m_f приведены на рис. 7. Они сравниваются с оценками качества воспроизведения, полученными с помощью RM- и SN-приближений, а также с данными, полученными *Корингом* [19] для гораздо больших размеров сети (N до 50000). Точки, полученные с помощью компьютерного моделирования, опять лежат между кривыми, полученными с помощью RM- и SN-приближений, а точки, полученные нами, практически совпадают с данными *Коринга.* Следует заметить, что на рис. 7 значения $(1 - m_f)$ умножены на 5, чтобы привести их в тот же диапазон значений, что и для разреженного кодирования.

Разреженное кодирование

Распределения финальных пересечений m_f, полученных при $p = 0.02$, показаны на рис. 8. Главной особенностью этих распределений является то, что они имеют две явно различающиеся моды только в случае, когда начальные состояния достаточно далеки от эталонов ($m_{in} = 0.1$ и $m_{in} = 0.3$). Когда $m_{in} = 0.5$ или $m_{in} = 1$ присутствует только одна мода, которая медленно смещается к меньшим значениям m_f при возрастании α. Та же особенность имеется при $p = 0.1$. Можно все-таки предположить, что распределение имеет две моды и при больших значениях m_{in}, которые слишком близки и имеют слишком большую дисперсию, чтобы их было легко разделить. При возрастании N моды становятся более узкими и их можно разделить, как это показано на рис. 8 для $p = 0.02$, $m_{in} = 0.5$, $\alpha = 0.34$ и $N = 15000$. Наличие дополнительной моды вблизи моды, близкой к $m_f = 1$, можно заметить даже для плотного кодирования (см. рис. 2, $m_{in} = 0.3$, $\alpha = 0.09$). Эту моду можно объяснить наличием метастабильных состояний в окрестности эталонов, которые исчезают при возрастании N [20].

Несмотря на плохое разделение мод, для оценки α_{cr} мы использовали ту же процедуру, что и при плотном кодировании. В качестве границы для разделения правильных и ложных аттракторов мы приняли $m_b = 0.75$ (вер-

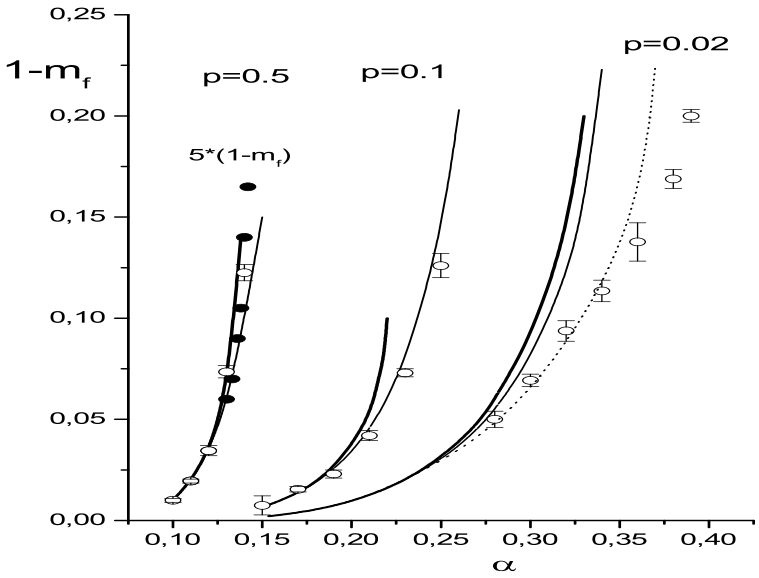

Рис. 7. Зависимость качества воспроизведения, полученная методом RM (толстые сплошные линии), SN (тонкие сплошные линии) и SS (пунктирные линии) и экстраполяцией результатов компьютерного моделирования к пределу $N \to \infty$ (светлые точки), от α и p. Темные точки — данные, взятые из работы [19].

тикальные линии на рис. 8). При $m_{in} = 0.1$ и $m_{in} = 0.3$ эта граница имеет тот же смысл, что и при плотном кодировании. Она явно разделяет хорошо выраженные моды для правильных и ложных аттракторов. При $m_{in} = 0.5$ и $m_{in} = 1$ вероятность сходимости нейродинамического процесса к правильным аттракторам сильно зависит от выбора m_b. Выбор $m_b = 0.75$ представляется разумным, потому что он обеспечивает изменение вероятности P, что некоторая траектория сойдется к аттрактору в окрестности эталона, от значения 1 при малых α до 0 при больших α. Оказалось, что выбор m_b в некотором разумном диапазоне от 0.7 до 0.8 не так сильно влияет на оценку α_{cr}, как на оценку величины P для каждой комбинации α и N.

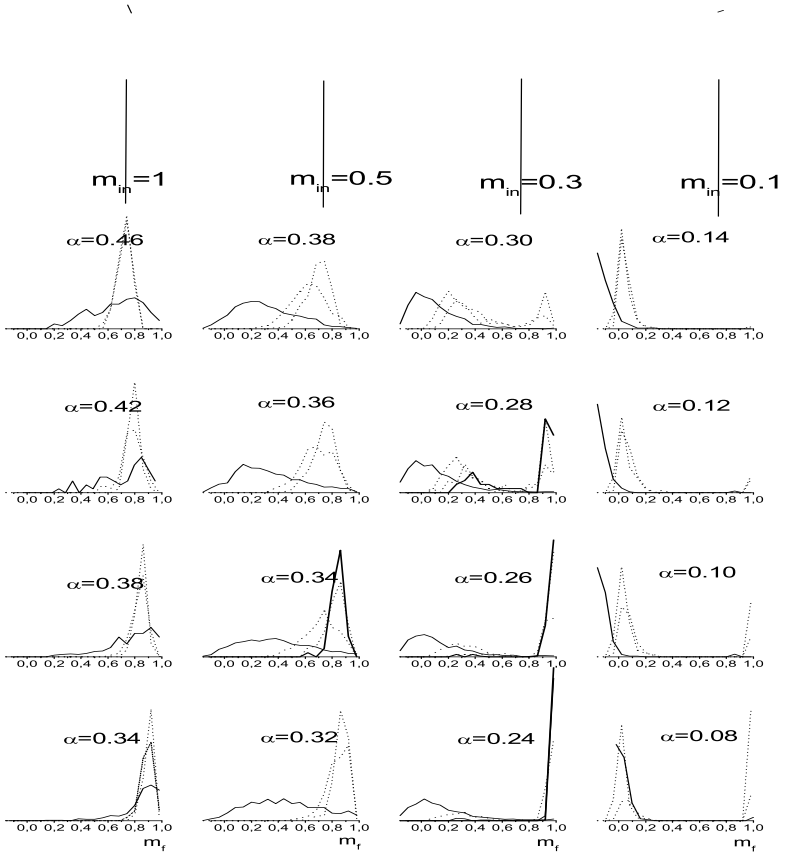

Рис. 8. Распределение финальных пересечений для $p = 0.02$ и различных значений N, α и m_{in}

Вертикальные линии соответствуют выбранным порогам ($m_b = 0.75$) для разделения траекторий, сходящихся к правильным и ложным аттракторам. Тонкие сплошные линии — $N = 850$ для $m_{in} = 0.1$, $N = 1100$ для $m_{in} = 0.3$ и $N = 1000$ для $m_{in} \geqslant 0.5$. Штриховые линии — $N = 5000$. Пунктирные линии — $N = 10000$. Толстые сплошные линии — $N = 15000$.

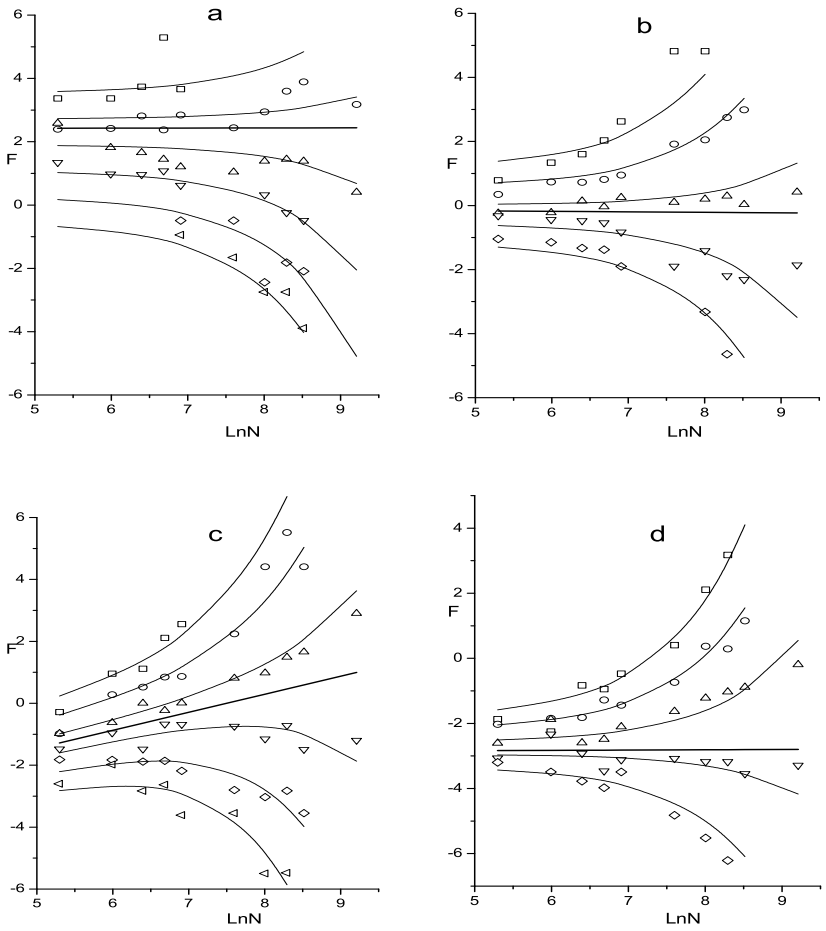

Рис. 9. Зависимость F от α и $\ln N$ для $p = 0.1$ и $m_{in} = 1$ (a), $m_{in} = 0.5$ (b), $m_{in} = 0.3$ (c) и $m_{in} = 0.1$ (d)

Значения α изменяются с шагом 0.02 от 0.23 (верхняя кривая) до 0.33 (нижняя кривая) на **(a)**, от 0.21 до 0.29 на **(b)**, от 0.15 до 0.25 на **(c)** и от 0.05 до 0.13 на **(d)**. Толстые прямые линии соответствуют $\alpha = \alpha_{cr} = 0.257$ на (a), $\alpha = \alpha_{cr} = 0.252$ на (b), $\alpha = \alpha_{cr} = 0.2$ на (c) и $\alpha = \alpha_{cr} = 0.104$ на (d).

Для получения оценки α_{cr} вероятности P трансформировались, как и для плотного кодирования, в значения F по формуле (29), а значения F аппроксимировались формулой (30). Результаты аппроксимации показаны на рис. 9 ($p = 0.1$) и рис. 10 ($p = 0.02$).

Как и для плотного кодирования, формула (30) обеспечивает хорошее приближение для зависимости F от α и N. Среднеквадратичная ошибка рассогласования экспериментальных значений F от рассчитанных по формуле (30) была примерно той же, что и для плотного кодирования (в диапазоне от 0.35 до 0.45). Информационная емкость (оцененная как α_{cr} при $m_{in} = 1$) составляет 0.257 ± 0.006 для $p = 0.1$ и 0.411 ± 0.007 для $p = 0.2$. Ее можно сравнить с предсказаниями RM-, SN- и SS-приближений, которые, соответственно, дают 0.22, 0.27 и 0.42 для $p = 0.1$; 0.31, 0.325 и 0.365 для $p = 0.02$. Таким образом, для $p = 0.1$, как и для плотного кодирования, компьютерное моделирование дает оценку информационной емкости между значениями, предсказанными RM- и SN-приближениями. В противоположность этому, для $p = 0.02$ компьютерное моделирование дает удивительно высокую оценку информационной емкости. Эта оценка даже больше предсказания SS-приближения. Для того, чтобы проверить, не является ли полученная завышенная оценка информационной емкости следствием выбора m_b, она была пересчитана при $m_b = 0.8$. Изменение m_b привело лишь к статистически незначимым изменениям информационной емкости: $\alpha_{cr} = 0.396 \pm 0.008$.

Гистограммы, представленные на рис. 8, также показывают, что полученная завышенная оценка информационной емкости при $p = 0.02$ не является столь нереалистичной. Распределение финальных пересечений m_f остается близким к единице вплоть до $N = 10000$ даже при $\alpha = 0.42$. То же самое демонстрирует рис. 10a: вероятность P продолжает возрастать при увеличении N даже при $\alpha = 0.42$. Мы можем предсказать ее падение при $\alpha = 0.42$ и $N > 10000$ только на основе используемой аппроксимации. Таким образом простейшее SS-приближение дает наиболее точную оценку информационной емкости в пределе большой разреженности кодирования. Однако это не столь уж и большое его преимущество, поскольку все три приближения SS, SN и RM дают все более близкие результаты при увеличении разреженности [10].

Результаты для областей притяжения представлены на рис. 11. При $p = 0.1$ (рис. 11a) данные компьютерного моделирования очень близки к полученным SN-методом. Тогда, как при $p = 0.02$ (рис. 11b) они очень близки к полученным SS-методом.

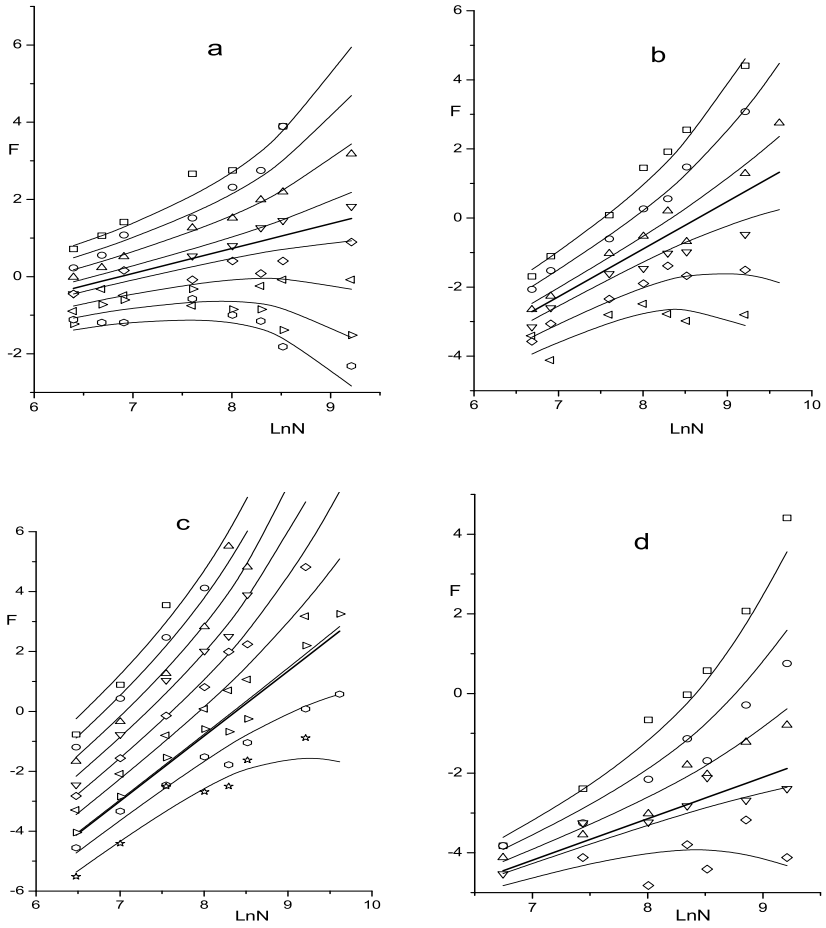

Рис. 10. Зависимость F от α и $\ln N$ для $p = 0.02$ и $m_{in} = 1$ (a), $m_{in} = 0.5$ (b), $m_{in} = 0.3$ (c) и $m_{in} = 0.1$ (d)

Значения α изменяются с шагом 0.02 от от 0.34 (верхняя кривая) до 0.48 (нижняя кривая) на **(a)**, от 0.28 до 0.38 на **(b)**, 0.14 до 0.3 на **(c)** и от 0.06 до 0.14 на **(d)**. Толстые прямые линии соответствуют $\alpha = \alpha_{cr} = 0.41$ на (a), $\alpha = \alpha_{cr} = 0.33$ на (b), $\alpha = \alpha_{cr} = 0.261$ на (c) и $\alpha = \alpha_{cr} = 0.115$ на (d).

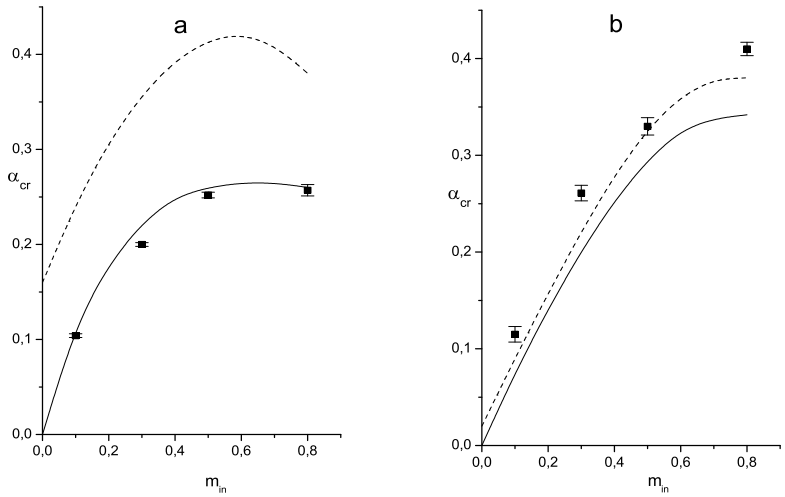

Рис. 11. Размеры баз притяжения для $p = 0.1$ **(a)** и $p = 0.02$ **(b)** Получены SN-методом (сплошные линии) и SS-методом (штриховые линии). Темные точки представляют результаты компьютерного моделирования.

При $p = 0.02$ области притяжения оказались даже больше, чем предсказывает SS-метод. Это означает, что вблизи границы области притяжения траектории, сходящиеся к правильным аттракторам, вначале будут удаляться от эталонов и только затем приближаться к ним. Такое поведение траекторий противоположно тому, что наблюдается в случае плотного кодирования. Среднее пересечение $m(t)$ между текущим паттерном нейронной активности и воспроизводимым эталоном для $p = 0.02$, $\alpha = 0.24$, $m_{in} = 0.3$ и $N = 15000$ показано на рис. 12a. В полном согласии с предсказанием SS-приближения, на первом шаге среднее пересечение уменьшается, однако затем оно увеличивается, приближаясь к единице. Заметим, что для выбранных параметров сети все 300 тестируемых траекторий сошлись к аттракторам в окрестности соответствующих эталонов. Как и для информационной емкости, полученное для $m_{in} = 0.3$ значение $\alpha_{cr} = 0.26 \pm 0.008$

не выглядит переоценкой критического значения информационной нагрузки при данном начальном пересечении, так как даже при $\alpha = 0.28$ (см. рис. 8 и рис. 10c) мода для правильных аттракторов продолжает возрастать вплоть до $N = 15000$ и мы можем предсказать ее падение при дальнейшем увеличении N только на основе используемой аппроксимации зависимости F от α и N.

Для $p = 0.1$ поведение пересечения $m(t)$ то же самое, что и для $p = 0.5$: оно монотонно возрастает для траекторий, сходящихся к правильным аттракторам, и может вначале увеличиваться, но затем уменьшается для ложных аттракторов (т. е. размер областей притяжения меньше предсказанного SS-методом, см. рис. 11a). Колебания $p_{\mu\nu}$ также, как и разница между p_{11} и p_{10}, возрастают для разреженного кодирования. Таким образом влияние начального состояния нейрона на его активность в процессе восстановления эталонов увеличивается при увеличении разреженности кодирования. Заметим, что для $p = 0.02$ даже $m(t)$ имеет слабые колебания, несмотря на гладкое и монотонное изменение порога активации $T(t)$ (рис. 12b).

К сожалению, невозможно напрямую сравнить наши расчеты с результатами, полученными в работах [16] и [22] для $p = 0.1$. В первой работе границы области притяжения представлены только для таких параметров сети, которые обеспечивают извлечение из нее максимальной информации. Для разреженного кодирования этот максимум достигается с случае, когда начальный паттерн не содержит активных нейронов, не принадлежащих восстанавливаемому эталону (т. е. когда начальный паттерн является фрагментом эталона с числом активных нейронов, меньшим чем в эталоне). В отличие от этого мы моделировали случай, когда уровень активности в начальных паттернах был тот же, что в эталонах. В работе [22] рассматривался только случай постоянного порога активации T. В этом случае области притяжения вокруг эталонов значительно меньше, чем для случая переменного порога, рассмотренного нами.

Как и для плотного кодирования, качество воспроизведения оценивалось по среднему финальному пересечению для правильных траекторий (т. е. траекторий с $m_f > 0.75$), стартующих из неискаженных эталонов. Результаты приведены на рис. 13. Для каждого значения α данные компьютерного моделирования аппроксимировались прямой линией в зависимости от $1/\sqrt{N}$. Пересечение прямой с осью ординат рассматривалось как асимптотическая оценка качества воспроизведения для данного α. Полученные асимптотические значения m_f показаны на рис. 7. Они сравниваются на этом рисунке со значениями, полученными методами RM, SN и

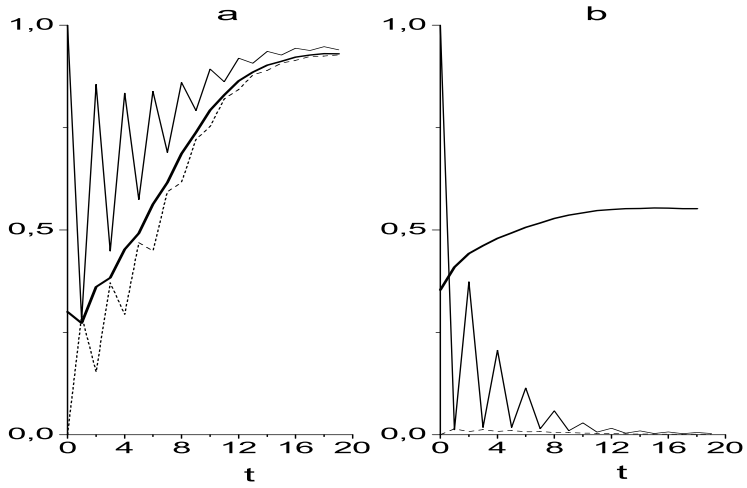

Рис. 12. Усредненные характеристики траекторий нейронной активности для $p = 0.02$, $m_{in} = 0.3$, $\alpha = 0.24$ и $N = 15000$
(a) Сплошная тонкая линия — p_{11}; штриховая линия — p_{10}; сплошная толстая линия — пересечение m между текущим паттерном нейронной активности и воспроизводимым эталоном. **(b)** Сплошная тонкая линия — p_{01}; штриховая линия — p_{00}; сплошная толстая линия — средний порог активации T.

SS. Для $p = 0.1$ экспериментальные данные очень близки к полученным SN-методом, тогда как для $p = 0.02$ они более близки к полученным SS-методом.

Информационная эффективность

Количество информации, извлекаемой из сети благодаря коррекции начальных паттернов, задается формулами (8), (9) и (12). Для каждого значения p оно оценивалось аналитически тем методом, который оказался наиболее точным в сравнении с данными компьютерного моделирования. Для $p = 0.5$ это был SN-метод второго порядка, предложенный в работе [22], для $p = 0.1$ — SN-метод первого порядка, предложенный в работе [10] и для

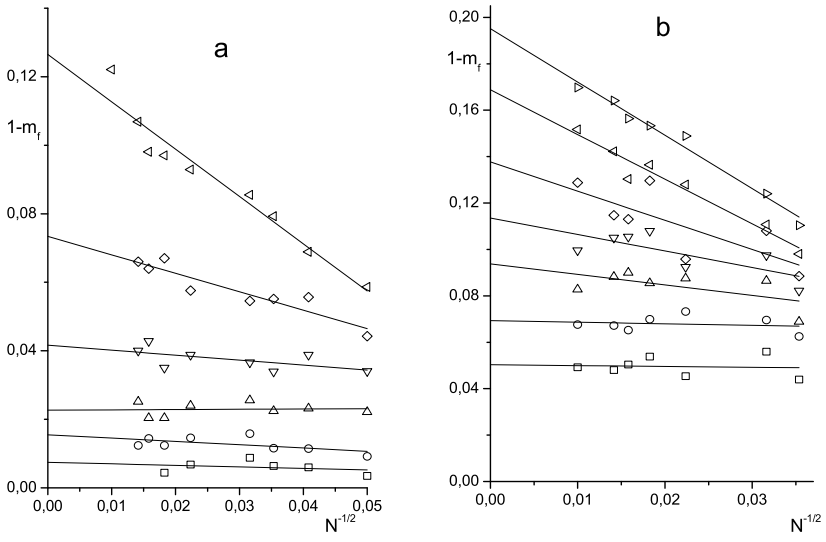

Рис. 13. Зависимость качества воспроизведения от α и N для $p = 0.1$ (a) и $p = 0.02$ (b), полученные компьютерным моделированием **(a)** Значения α изменяются от 0.15 (нижняя линия) до 0.25 (верхняя линия) с шагом 0.2. **(b)** Значения α изменяются от 0.28 (нижняя линия) до 0.40 (верхняя линия) с шагом 0.2.

$p = 0.02$ — SS-метод (который фактически является SN-методом нулевого порядка). Результаты этих оценок приведены на рис. 14 в зависимости от информационной нагрузки α. Для каждого значения α начальное пересечение с воспроизводимым эталоном бралось на границе базы притяжения. Задание начального пересечения полностью определяет значения вероятностей q_1 и q_0 (см. формулу (11)), а тем самым и начальную энтропию воспроизводимого эталона Nh_{in}, заданную формулой (9). Конечная энтропия воспроизводимого эталона определяется конечными значениями вероятностей $p_{\mu\nu}$. Во всех используемых методах влияние начального состояния нейрона на его конечное состояние игнорируется, т. е. p_{10} полагается равным p_{11}, а p_{00} полагается равным p_{01}. Тогда вероятности $p_{11} = p_{10} = p_1$ и

$p_{01} = p_{00} = p_0$ полностью определяются значением конечного пересечения m_f (см. уравнение (15)), и поэтому конечная энтропия воспроизводимого эталона Nh_f, заданная формулой (12), полностью определяется значением m_{in} (взятым для данного α на границе базы притяжения) и качеством воспроизведения, т. е. m_f.

Коэффициенты информационной эффективности E, рассчитанные по формуле (17) для различных значений p, показаны на рис. 14 толстыми линиями. Значения i_{in}, рассчитанные по формуле (18), которые задают начальную энтропию воспроизводимых эталонов, показаны тонкими линиями. Значения i_f, рассчитанные по формуле (19), задающие конечную энтропию эталонов, показаны штриховыми линиями. Значения $i'_f = \alpha h'_f / h(p)$, где h'_f рассчитано по формуле (16), показаны пунктирными линиями.

Когда информационная нагрузка α относительно мала, размеры баз притяжения велики ($h_{in} \simeq h(p)$), а качество воспроизведения близко к единице ($h_f \ll h(p)$). Тогда $E \simeq \alpha$, т. е. из сети извлекается почти столько же информации, сколько в нее заложено. Когда α возрастает, различие между α и E возрастает как за счет уменьшения баз притяжения, так и за счет уменьшения качества воспроизведения. Для $p = 0.5$ максимальная информационная эффективность составляет 0.092, что совпадает с результатом работы [16]. Для $p = 0.1$ она составляет 0.167, что превышает значение 0.143, полученное в работе [16]. Это различие не может объясняться тем фактом, что в этой работе при расчете конечной энтропии воспроизводимого эталона пренебрегается информацией, данной начальным паттерном (т. е. используется формула (16) вместо (12)). Такое пренебрежение наоборот приводит к завышению информационной эффективности. Основная причина расхождения нашей оценки информационной эффективности и оценки работы [16] объясняется тем, что в этой работе использовался метод, основанный на RM-методе, а этот метод дает заниженную оценку информационной емкости при $p = 0.1$ (см. раздел «Компьютерное моделирование», с. 48–64). При такой разреженности кодирования мы оценивали информационную эффективность SN-методом первого порядка, который, как показано на рис. 11а, дает намного более точную оценку всех характеристик сети. Различие между нашей оценкой и результатом работы [16] могло бы быть еще больше, если бы мы считали информационную эффективность для случая, когда начальные паттерны являются фрагментами эталонов, как это сделано в работе [16], а не для случая, когда начальные паттерны содержат активные нейроны, совпадающие и не совпадающие с активными нейронами эталонов, как это сделано в наших исследованиях.

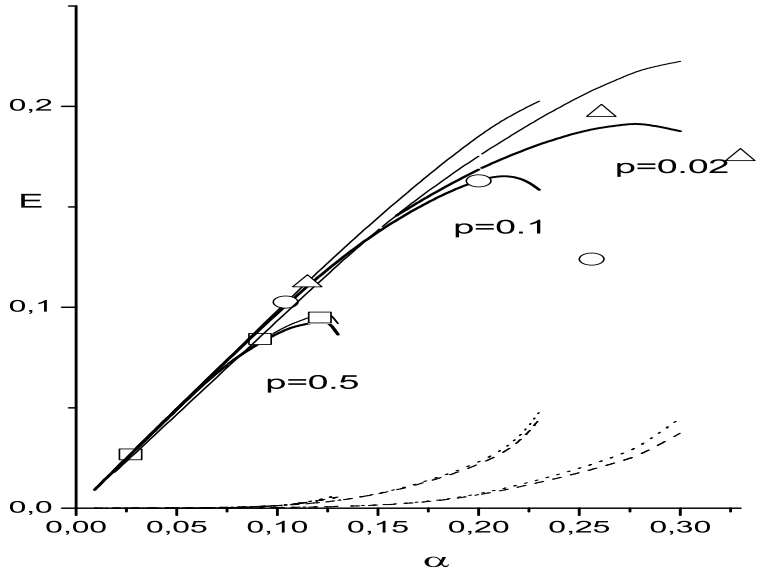

Рис. 14. Информационная эффективность E (сплошные толстые линии) в зависимости от α и p

Рассчитаны по SN-методу второго порядка ($p = 0.5$), первого порядка ($p = 0.1$) и нулевого порядка ($p = 0.02$), и с помощью компьютерного моделирования ($\diamond$ — для $p = 0.5$, $\circ$ — для $p = 0.1$ и Δ — для $p = 0.02$). Тонкие сплошные линии — i_{in}; штриховые линии — i_f; пунктирные линии — i'_f.

Как хорошо известно [11, 16], в случае разреженного кодирования восстановление эталонов по их фрагментам позволяет извлечь из сети больше информации, чем их восстановление по начальным паттернам, содержащим активные и неактивные нейроны эталонов.

Для $p = 0.02$ информационная эффективность оценивалась SS-методом. Ее максимальное значение составило 0.192. Компьютерное моделирование дает даже слегка большее значение.

Для $p \to 0$ уравнения (9–11) и (18) дают $i_{in} = \alpha(1 - m_{in})$. В этом пределе все три метода SS, SN и RM предсказывают высокое качество воспроизведения (т. е. $i_f = 0$) [10]. Оба метода SS и SN (см. уравнение (26)) предсказывают, что размер области притяжения в этом случае задается

формулой $m_{in} = \sqrt{2\alpha \ln 2}$, а потому $E = \alpha(1 - \sqrt{2\alpha \ln 2})$. Максимальное значение информационной емкости составляет $E = 2/(27 \ln 2) \simeq 0.11$. Таким образом, существует оптимальная разреженность кодирования, при которой информационная эффективность максимальна. Эта эффективность по крайней мере не менее значения 0.2, достигаемого при $p = 0.02$.

Заключение

В настоящей работе мы исследовали аналитически и с помощью компьютерного моделирования влияние разреженности на размер областей притяжения в полносвязной автоассоциативной сети хопфилдового типа. Компьютерное моделирование показало, что существующие аналитические подходы недостаточно точны при больших разреженностях. Для $p = 0.02$ траектории, описывающие активность сети, демонстрируют парадоксальное поведение: пересечение с восстанавливаемым эталоном вначале убывает, но потом траектория возвращается и активность сети приближается к восстанавливаемому эталону. Поэтому, размер области притяжения оказывается даже больше, чем получаемый в SS-приближении. Такое поведение не предсказывается существующими аналитическими методами. Основной причиной их неспособности отразить это свойство является игнорирование различия в поведении нейронов, активных и неактивных на предыдущих шагах.

Компьютерное моделирование показало, что размер области притяжения меняется немонотонно по мере возрастания разреженности: вначале он возрастает, а затем убывает. Поэтому в пределе $p \to 0$ разреженность ухудшает способность к коррекции искаженных эталонов, хотя улучшает как информационную емкость, так и качество воспроизведения. Количество информации I_g, извлекаемой из сети за счет коррекции предъявленных искаженных вариантов записанных эталонов, используется как обобщенная характеристика способности сети к выполнению функции ассоциативной памяти. Существует оптимальная разреженность (p порядка 10^{-2}), для которой информационная эффективность $E = I_g/N^2$ максимальна. Эта эффективность составляет около 0.2. Интересно, что оптимальная разреженность соответствует нейронной активности мозга. Мы полагаем, что это совпадение не случайно и отражает важную особенность работы мозга.

Максимальная информационная эффективность оказывается примерно вдвое выше, чем для плотного кодирования ($p = 0.5$) и чем в предельном

случае ($p \to 0$). Однако, она остается значительно ниже максимального предела информационной эффективности $E = 1/(4 \ln 2) \simeq 0.36$, задаваемого энтропией матрицы связей [11]. Это означает, что в принципе процедура воспроизведения может быть улучшена путем повышения количества информации, извлекаемой при коррекции искаженных паттернов. С другой стороны, оценка информационной эффективности, сделанная в данной работе, относится только к части записанной информации, извлекаемой из сети в процессе воспроизведения. Другая часть информации может быть извлечена за счет распознавания новых (далеких от записанных эталонов) и знакомых (близких к одному из записанных эталонов) паттернов среди тестирующего множества входных паттернов информации [11]. Таким образом, перед тем, как пытаться улучшить процедуру коррекции, уместно оценить также количество информации, которая может быть извлечена из сети за счет распознавания входных паттернов как новых и знакомых.

Литература

1. *Дунин-Барковский В. Л.* Информационные процессы в нейронных структурах. – М.: Наука, 1978. – 166 с.
2. *Фролов А. А., Мушинский А. М., Цодыкс М. В.* Имитационная модель автоассоциативной памяти в виде нейронной сети с малым уровнем активности // *Биофизика*. – **36**. – с. 339–343.
3. *Amari S.* Characteristics of sparsely encoded associative memory // *Neural Networks*. – 1989. – **2**. – pp. 451–457.
4. *Amari S., Maginu K.* Statistical neurodynamics of associative memory // *Neural Networks*. – 1988. – **1**. – pp. 63–73.
5. *Amit D. J., Gutfreund H., Sompolinsky H.* Statistical mechanics of neural networks near saturation // *Annals of Physics*. – 1987. – **173**. – pp. 30–67.
6. *Buhmann J., Divko R., Schulten K.* Associative memory with high information content // *Physical Review A*. – 1989. – **39**. – pp. 2689–2692.
7. *Coolen A. C. C., Sherington D.* Order-parameter flow in the fully connected Hopfield model near saturation // *Physical Rewiew E*. – 1993. – **49**. – pp. 1921–1934.
8. *Forrest B. M.* Content-addressability and learning in neural networks // *Journal of Physics A*. – 1988. – **21**. – pp. 245–255.
9. *Frolov A. A., Husek D., Muraviev I. P.* Statistical neurodynamics of sparsely encoded Hopfield-like associative memory // *Neural Netw. World*. – 1996. – **6**. – pp. 609–617.
10. *Frolov A. A., Husek D., Muraviev I. P.* Informational capacity and recall quality in sparsely encoded Hopfield-like neural network: Analytical approaches and computer simulation // *Neural Networks*. – 1997. – **10**. – pp. 845–855.

11. *Frolov A. A, Muraviev I. P.* Informational characteristics of neural networks capable of associative learning based on Hebbian plasticity // *Network*. – 1993. – **4**. – pp. 495–536.

12. *Gardner E., Derrida B., Mottishaw P.* Zero temperature parallel dynamics for infinite range spin glasses and neural networks // *J. Physique*. – 1987. – **48**. – pp. 741–755.

13. *Goles-Chacc E., Fogelman-Soulie F., Pellegrin D.* Decreasing energy functions as a tool for studying threshold networks // *Discrete Mathematics*. – 1985. – **12**. – pp. 261–277.

14. *Hopfield J. J.* Neural network and physical systems with emergent collective computational abilities // *Proceedings of the National Academy of Science, USA*. – 1982. – **79**. – pp. 2544–2548.

15. *Horner H.* Neural networks with low levels of activity: Ising vs. McCulloch-Pitts neurons // *Z. Physik B*. – 1989. – **75**. – pp. 133–136.

16. *Horner H., Bormann D., Frick M., Kinzelbach H., Schmidt A.* Transients and basins of attraction in neural network models // *Z. Physik B*. – 1989. – **76**. – pp. 381–398.

17. *Husek D., Frolov A. A.* Evaluation of the informational capacity of Hopfield network by computer simulation // *Neural Network World*. – 1994. – **4**. – pp. 53–65.

18. *Kinzel W.* Learning and pattern recognition in spin glass models // *Z. Physik B*. – 1985. – **60**. – pp. 205–213.

19. *Kohring G. A.* A high-precision study of the Hopfield model in the phase of broken replica symmetry // *Journal of Statistical Physics*. – 1990. – **59**. – pp. 1077–1086.

20. *Kohring G. A.* Convergence time and finite size effects in neural networks // *Journal of Physics A: Math. Gen.* – 1990. – **23**. – pp. 2237–2241.

21. *Koyama H., Fujie N., Seyama H.* Results from the Gardner–Derrida–Mottishaw theory of associative memory *Neural Networks*. – 1999. – **12**. – pp. 247–257.

22. *Okada M.* Notions of associative memory and sparse coding // *Neural Networks*. – 1996. – **9**. – pp. 1429–1458.

23. *Palm G., Sommer F. T.* Information capacity in recurrent McCulloch–Pitts network with sparsely coded memory states // *Network*. – 1992. – **3**. – pp. 177–186.

24. *Perez-Vicente C. J., Amit D. J.* Optimized network for sparsely coded patterns // *Journal of Physics A: Math. Gen.*. – 1989. – **22**. – pp. 559–569.

25. *Shigemitsu S., Okada M.* Statistical neurodynamics of the sparsely encoded associative memory (in Japanese) // *The Brain & Neural Nerworks*. – 1996. – **3**. – pp. 58–64.

26. *Tsodyks M. V. Feigelman M. V.* The enhanced storage capacity in neural network with low activity level // *Europhysical Letters*. – 1988. – **6**. – pp. 101–105.

Александр Алексеевич ФРОЛОВ, кандидат физико-математических, доктор биологических наук, профессор, заведующий лаборатории математической нейробиологии обучения ИВНД и НФ РАН, специалист в области биомеханики и нейросетевого моделировани функций нервной системы (двигательное управление, ассоциативная память, биофизические механизмы обучения и памяти), автор более 200 статей, 6 изобретений и 2 монографий (по нейросетевым моделям ассоциативной памяти).

Душан ГУСЕК, научный сотрудник Института информатики Академии наук Чешской республики. Область научных интересов: нейронные сети, представление знаний, представление информации и базы данных. Автор более 50 научных публикаций.
URL: http://www.cs.cas.cz/∼dusan

Игорь Петрович МУРАВЬЕВ, научный сотрудник лаборатории математичесмкой нейробиологии обучения ИВНД и НФ РАН, специалист в области нейросетевого моделирования ассоциативной памяти, автор более 20 статей и 2 монографий (по нейросетевым моделям ассоциативной памяти).

Б. В. КРЫЖАНОВСКИЙ, Л. Б. ЛИТИНСКИЙ
Институт оптико-нейронных технологий РАН,
Москва
E-mail: kryzhanov@mail.ru,
litin@mail.ru

ВЕКТОРНЫЕ МОДЕЛИ АССОЦИАТИВНОЙ ПАМЯТИ

Аннотация

Дается обзор векторных моделей ассоциативной памяти. Модель Хопфилда позволяет эффективно запомнить сравнительно небольшое число паттернов – порядка 15% от размера нейронной сети. Существенно превзойти этот показатель удается только в Поттс-стекольной модели ассоциативной памяти, где нейроны могут находиться в большем чем два числе состояний. Показано, что еще большей емкостью памяти обладает параметрическая нейронная сеть (ПНС). Для оценки емкости памяти используется статистическая техника Чебышева–Чернова.

B. V. KRYZHANOVSKY, L. B. LITINSKII
Institute of Optical Neural Technologies, RAS,
Moscow
E-mail: kryzhanov@mail.ru,
litin@mail.ru

THE VECTOR MODELS OF ASSOCIATIVE MEMORY

Abstract

We present a short review of vector models for associative memory. The storage capacity of the Hopfield model is about 15% of this network size. It can be increased significantly in the Potts-glass model of the associative memory only. In this model neurons can be in more than two different states. We show that even greater storage capacity can be achieved in the parametrical neural network (PNN). We use the Chebyshev–Chernov statistical technique to estimate the PNN storage capacity.

УДК 004.032.26 (06) Нейронные сети

Введение

Последние 15 лет отмечены интересом к моделям ассоциативной памяти с q-нарными нейронами. Всё в этих моделях подобно модели Хопфилда, но число состояний q, в которых могут находиться нейроны, больше 2.

В работах [1–3] состояния нейронов изображаются *спиновой переменной* S_i, принимающей q различных значений:

$$S_i = -1 + \frac{2(k-1)}{q-1}, \quad i = 1, \ldots, N; \; k = 1, \ldots, q\,. \tag{1}$$

В другом подходе, различные состояния нейронов изображаются точками на единичной окружности. Один из вариантов такой сети был рассмотрен в [4, 5] — *фазорная модель*. Другой вариант — в [6] (*циклическая модель*).

В фазорной сети для изображения различных состояний нейронов используются двумерные векторы единичной длины (*фазоры*). Сеть состоит из N фазоров $S_i,\ i = 1, \ldots, N$ — комплексных чисел, по модулю равных единице. В первоначальной версии S_i могли принимать непрерывный ряд значений (непрерывная модель), а в более поздней — только q корней уравнения $S_i^q = 1$ (дискретная модель). Пусть p паттернов — это p наперед заданных N-мерных наборов $(\xi_1^\mu, \xi_2^\mu, \ldots, \xi_N^\mu),\ \ \mu = 1, \ldots, p$, где $\xi_j^\mu = \exp(\imath\theta_j^\mu)$ — состояние i-го нейрона в μ-м паттерне. Тогда синаптическая связь задается выражением

$$C_{ij} = (1 - \delta_{ij}) \sum_{\mu=1}^{p} \xi_i^\mu \bar{\xi}_j^\mu\,; \tag{2}$$

черта сверху означает здесь комплексное сопряжение. Динамическое решающее правило определяется с помощью (комплексного) локального поля:

$$h_i = \sum_j C_{ij} S_j, \quad S_i(t+1) = \frac{h_i}{\mid h_i \mid}\,. \tag{3}$$

В случае последовательной динамики энергия состояния

$$E = -\frac{1}{2} \sum_{i \neq j} C_{ij} S_j \bar{S}_j$$

монотонно убывает. Эта сеть изучалась также в [7].

Близкая по духу циклическая модель нейронной сети рассматривалась в [6]. Гамильтониан этой системы имеет вид

$$H = -\sum_{i \neq j}^{N} \sum_{\mu} \cos \frac{2\pi}{q} \left\{ (n_i - \xi_i^\mu) - (n_j - \xi_j^\mu) \right\},$$

где N — число нейронов, $n_i = 0, 1, \ldots, (q-1)$ — состояние i-го нейрона, а ξ_i^μ — состояние i-го нейрона в μ-м паттерне, записанном в память сети $(\mu = 1, \ldots, p)$. При $q = 2$ эта схема сводится к модели Хопфилда.

Очевидной особенностью гамильтониана этой модели является его инвариантность относительно преобразования $\{n_i\} \to \{n_i + k \pmod q\}$, где k — целое число. Связи между фазорной сетью и циклической сетью исследовались в [8], где установлено, что эти модели имеют много общего. Емкость памяти всех описанных выше моделей достигает своего максимума при $q = 2$, когда они превращаются в модель Хопфилда, а затем, с ростом q, емкость памяти падает.

Альтернативный подход к проблеме состоит в том, что q различных состояний нейронов изображаются q-мерными векторами. Исторически первой работой такого рода была статья [9], где рассматривалась модель, получившая название *Поттс-стекольной нейросети* (термин заимствован из физики). Здесь использовался так называемый *анизотропный* гамильтониан (см. ниже; некоторые подробности см. также в [10]). Согласно оценкам автора статьи [9], емкость памяти такой сети должна превосходить емкость памяти модели Хопфилда приблизительно в $q(q-1)/2$ раз. При больших значениях q это является весомым фактором. Данная модель исследовалась также в работах [11–13].

Поттс-стекольная модель с *изотропным* гамильтонианом изучалась в [10], а q-мерное обобщение непрерывной фазорной модели (когда состояния нейронов изображаются q-мерным векторами с вещественными координатами) — в работе [8]. Емкость памяти этих моделей оказалась меньше емкости памяти модели Хопфилда.

По-видимому, последней по времени создания моделью ассоциативной памяти на q-мерных нейронах является *параметрическая нейронная сеть* (ПНС), предложенная в [14, 15]. Сформулированная как вариант сети, ориентированной на нелинейно-оптические принципы обработки информации, эта модель была затем формализована в рамках векторного подхода к описанию нейронов [16].

Использование для оценки емкости памяти *статистической техники Чебышева–Чернова* [17,18] позволило установить, что среди всех q-нарных сетей именно ПНС обладает наилучшими показателями по объему памяти и помехоустойчивости. К изложению этого круга вопросов мы и переходим. Вначале будут описаны Поттс-стекольная и параметрическая нейросети, а затем, на примере анализа последней, будет изложена статистическая техника Чебышева–Чернова.

Поттс-стекольная нейросеть и параметрическая нейронная сеть

1°. Все варианты Поттс-стекольной ассоциативной памяти являются производными от известной модели магнетика, построенной Поттсом, которая обобщает *модель Изинга* на случай спиновой переменной, принимающей не два значения $\{-1,+1\}$, а произвольное число q различных значений [19–21]. Во всех случаях переход от модели магнетика к модели нейронной сети происходил по одной и той же схеме, с помощью которой модель Изинга была в свое время связана с моделью Хопфилда [22].

А именно, характерное для физических рассмотрений короткодействующее взаимодействие между двумя соседними спинами заменялось межсвязями хеббовского типа между векторами-нейронами. В результате возникающего в системе дальнодействия оказывалось возможным применить для вычислении статистической суммы приближение среднего поля и построить фазовую диаграмму системы. Различные области фазовой диаграммы интерпретировались затем в терминах способности или неспособности сети к восстановлению зашумленных паттернов.

Поттс-стекольную сеть с анизотропным гамильтонианом [9] можно описать следующим образом.

Сеть составлена из N нейронов, каждый из которых может находиться в одном из q различных состояний. При этом l-ое состояние нейрона изображается q-мерным вектором-столбцом $\vec{\xi}_l$, у которого l-я координата пропорциональна $(q-1)$, а все остальные координаты пропорциональны -1:

$$l \longrightarrow \vec{\xi}_l = \frac{1}{q} \begin{pmatrix} -1 \\ \vdots \\ q-1 \\ \vdots \\ -1 \end{pmatrix}, \quad l = 1, \ldots, q\,.$$

Состояние всей сети задается набором из N q-мерных векторов: $X = (\vec{x}_1, \ldots, \vec{x}_N)$, где $\vec{x}_i = \vec{\xi}_{l_i}$, $1 \leqslant l_i \leqslant q$. Паттерны, информация о которых хранится в межсвязях сети — это p наперед заданных наборов по N q-мерных векторов каждый:

$$X^\mu = (\vec{x}_1^\mu, \vec{x}_2^\mu, \ldots, \vec{x}_N^\mu)\,, \quad \mu = 1, \ldots, p\,,$$

где

$$\vec{x}_i^\mu = \vec{\xi}_{l_i^\mu}\,, \quad 1 \leqslant l_i^\mu \leqslant q\,.$$

Межсвязь между i-м и j-м нейронами задается $(q \times q)$-матрицей $\mathbf{J}_{ij}$, которая в духе хеббовского правила аккумулирует состояния i-го и j-го нейронов во всех p паттернах

$$\mathbf{J}_{ij} = (1 - \delta_{ij}) \sum_{\mu=1}^{p} (., \vec{\mathbf{x}}_j^\mu)\, \vec{\mathbf{x}}_i^\mu\,. \tag{4}$$

Здесь через $(., \vec{\mathbf{x}})\vec{\mathbf{y}}$ обозначена $(q \times q)$-матрица[1], которая произвольный вектор $\vec{z} \in \mathrm{R}^q$ переводит в вектор $\vec{y}$, домножая его на величину скалярного произведения $(\vec{z}, \vec{x})$: $(., \vec{\mathbf{x}})\vec{\mathbf{y}}\vec{z} = (\vec{z}, \vec{x})\vec{y}$. (Заметим, что физическая традиция использует другое обозначение для этой матрицы: $\vec{y}\vec{x}^{\mathrm{T}}$, где $\vec{x}^{\mathrm{T}} = (x_1, \ldots, x_q)$ — q-мерная вектор-строка. Формальное матричное умножение вектор-столбца $\vec{y}$, расположенного слева, на вектор-строку $\vec{x}^{\mathrm{T}}$, расположенную справа, дает $(q \times q)$-матрицу $(y_k x_l)_{k,l=1}^q$, которая у нас обозначена через $(., \vec{\mathbf{x}})\vec{\mathbf{y}}$. В то же время, скалярное произведение векторов $\vec{x}$ и $\vec{y}$, в физических обозначениях, имеет вид: $\vec{x}^{\mathrm{T}}\vec{y} = \vec{y}^{\mathrm{T}}\vec{x}$.)

Локальное поле $\vec{h}_i$, действующее на i-й нейрон, определяется аналогично (3)

$$\vec{h}_i = \sum_{j=1}^{N} \mathbf{J}_{ij}\vec{x}_j = \sum_{\mu=1}^{p} \vec{x}_i^\mu \sum_{\substack{j=1 \\ j \neq i}}^{N} (\vec{x}_j, \vec{x}_j^\mu)\,,$$

а динамика сети задается решающим правилом: в следующий момент времени i-й нейрон переходит в состояние, максимизирующее $(\vec{\xi}_l, \vec{h}_i)$,

$$\vec{x}_i(t+1) = \vec{\xi}_l, \text{ где } (\vec{\xi}_l, \vec{h}_i) \geqslant (\vec{\xi}_{l'}, \vec{h}_i) \quad \forall\, l' = 1, \ldots, q.$$

[1] О чем свидетельствует полужирный шрифт.

При $q = 2$ такая сеть эквивалентна стандартной модели Хопфилда [9]. Оценка емкости памяти этой сети при $N \gg 1$ будет приведена в следующем разделе.

2°. Параметрическая нейронная сеть опирается на *параметрический нейрон* — обладающий кубической нелинейностью элемент, способный к преобразованию и генерации частот в процессах параметрического четырехволнового смешения [23].

Нейроны в ПНС обмениваются по межсвязям квазимонохроматическими импульсами на q различных частотах $\{\omega_l\}_{l=1}^{q}$. Кроме того, импульсы имеют амплитуды, равные ± 1. Таким образом, нейроны в ПНС могут находиться в $2q$ различных состояниях. Будем описывать состояния нейронов с помощью снабженных амплитудами ± 1 единичных ортов $\vec{e}_l$ пространства $\mathbf{R}^q$:

$$\vec{x}_i = x_i \vec{e}_{l_i},\ \text{где } x_i = \pm 1,\ \vec{e}_l = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix} \in \mathbf{R}^{\mathrm{q}},\ \begin{cases} i = 1, \ldots, N; \\ l = 1, \ldots, q; \\ 1 \leqslant l_i \leqslant q. \end{cases} \tag{5}$$

Состояние сети как целого задается набором N таких q-мерных векторов $X = (\vec{x}_1, \ldots, \vec{x}_N)$, а паттерны, записанные в межсвязях сети — это p наперед заданных наборов, состоящих из N q-мерных векторов каждый:

$$X^\mu = (\vec{x}_1^{\,\mu}, \vec{x}_2^{\,\mu}, \ldots, \vec{x}_N^{\,\mu}), \quad \mu = 1, \ldots, p,$$

где

$$\vec{x}_i^{\,\mu} = x_i^\mu \vec{e}_{l_i^\mu}, \quad x_i^\mu = \pm 1, \quad 1 \leqslant l_i^\mu \leqslant q.$$

Межсвязь между i-м и j-м нейронами задается $(q \times q)$-матрицей $\mathbf{J}_{ij}$ аналогично выражению (4):

$$\mathbf{J}_{ij} = (1 - \delta_{ij}) \sum_{\mu=1}^{p} (., \vec{\mathbf{x}}_j^{\,\mu}) \vec{\mathbf{x}}_i^{\,\mu},$$

а локальное поле $\vec{h}_i$, действующее на i-й нейрон, имеет вид:

$$\vec{h}_i = \sum_{j=1}^{N} \mathbf{J}_{ij} \vec{x}_j = \sum_{\mu=1}^{p} \vec{x}_i^{\,\mu} \sum_{\substack{j=1 \\ j \neq i}}^{N} (\vec{x}_j, \vec{x}_j^{\,\mu}) = \sum_{l=1}^{q} A_{il} \vec{e}_l,$$

где амплитуды A_{il} равны:

$$A_{il} = \sum_{\substack{j=1 \\ j \neq i}}^{N} \sum_{\mu=1}^{p} (\vec{x}_i^{\,\mu}, \vec{e}_l)(\vec{x}_j, \vec{x}_j^{\,\mu}) \,. \tag{6}$$

Динамика сети задается более сложным решающим правилом, чем в случае Поттс-стекольной модели. Пусть A_{ik} — амплитуда, в данный момент времени наибольшая по модулю среди всех амплитуд (6):

$$\mid A_{ik} \mid = \max_{1 \leqslant l \leqslant q} \mid A_{il} \mid . \tag{7}$$

Тогда i-й нейрон в следующий момент времени $(t+1)$ принимает значение:

$$\vec{x}_i(t+1) = \mathrm{sign}(A_{ik}) \vec{e}_k \,. \tag{8}$$

Иначе говоря, i-й вектор-нейрон ориентируется в направлении, ближайшем к направлению локального поля $\vec{h}_i$.

Эволюция сети состоит в последовательной переориентации векторов $\vec{x}_i$ по правилам (7), (8). Условимся, что когда максимальное по модулю значение принимают несколько амплитуд и нейрон находится в одном из этих неулучшаемых состояний, его состояние не меняется. Тогда нетрудно показать, что каждый шаг эволюции сети будет сопровождаться понижением энергии

$$E = -\frac{1}{2} \sum_{i=1}^{N} (\vec{h}_i, \vec{x}_i) = -\sum_{i,j=1}^{N} (\mathbf{J}_{ij} \vec{x}_j, \vec{x}_i) \,.$$

Рано или поздно система «свалится» в локальный минимум по энергии. Все нейроны $\vec{x}_i$ будут при этом ориентированы неулучшаемым образом и эволюция сети прекратится. Такие состояния суть *неподвижные точки* системы. Необходимым и достаточным условием того, чтобы состояние X было неподвижной точкой, является выполнение системы неравенств:

$$(\vec{x}_i, \vec{h}_i) \geqslant \mid (\vec{e}_l, \vec{h}_i) \mid, \ \forall \ l = 1, \ldots, q; \ \forall \ i = 1, \ldots, N \,. \tag{9}$$

Заметим в заключение, что при $q = 1$ ПНС переходит в стандартную модель Хопфилда: векторы $\vec{x}_i$ (5) превращаются в обычные бинарные переменные $x_i = \pm 1$, а все остальное здесь просто копирует модель Хопфилда.

Статистическая техника Чебышева–Чернова

Дадим асимптотическую оценку емкости памяти ПНС при $N \gg 1$. Пусть начальным состоянием сети является искаженный m-й паттерн

$$\widetilde{X}^m = (a_1\widehat{b}_1\vec{x}_1^m,\ a_2\widehat{b}_2\vec{x}_2^m, \ldots,\ a_N\widehat{b}_N\vec{x}_N^m)\,.$$

Независимые случайные величины $\{a_i\}_1^N$ и $\{\widehat{b}_i\}_1^N$ задают мультипликативный шум. Случайная величина a_i с вероятностью a принимает значение -1, а с вероятностью $(1-a)$ — значение 1 (в оптической терминологии — *амплитудный шум*). Случайный оператор $\widehat{b}_i$ с вероятностью b изменяет состояние i-го нейрона на другое, а с вероятностью $(1-b)$ оставляет вектор $\vec{x}_i^m$ неизменным (в оптической терминологии — *частотный шум*). Оценим, при каких условиях m-й паттерн будет распознаваться сетью правильно.

Амплитуды A_{il} (6) имеют вид:

$$A_{il} = \begin{cases} x_i^m \sum_{j=1}^{N-1} \xi_j + \sum_{r=1}^{L} \eta_r, & \text{для } l = l_i^{(m)}; \\ \sum_{r=1}^{L} \eta_r, & \text{для } l \neq l_i^{(m)}, \end{cases}$$

где

$$\begin{gathered} \xi_j = a_j(\vec{x}_j^m,\ \widehat{b}_j\vec{x}_j^m)\,, \\ \eta_r \equiv \eta_j^\mu(l) = a_j(\vec{e}_l, \vec{x}_i^\mu)(\vec{x}_j^\mu, \widehat{b}_j\vec{x}_j^m)\,, \\ j = 1, \ldots, N, \quad j \neq i\,, \\ \mu = 1, \ldots, p\,, \quad \mu \neq m \end{gathered}$$

и, для простоты, величины η проиндексированы обобщенным индексом $r = (j, \mu)$, который принимает $L = (N-1)(p-1)$ различных значений; зависимость величин η от l в дальнейшем никакой роли не играет, поэтому l можно опустить.

Для рандомизированного набора паттернов $\{X^\mu\}_{\mu=1}^p$ случайные величины ξ_j и η_r независимы и имеют распределения

$$\xi_j = \begin{cases} +1\,, & \text{с вероятностью } (1-b)(1-a), \\ 0\,, & \text{с вероятностью } b, \\ -1\,, & \text{с вероятностью } (1-b)a; \end{cases}$$
$$\eta_r = \begin{cases} +1\,, & \text{с вероятностью } 1/2q^2, \\ 0\,, & \text{с вероятностью } 1 - 1/q^2, \\ -1\,, & \text{с вероятностью } 1/2q^2. \end{cases} \qquad (10)$$

Для того, чтобы i-й нейрон перешел в состояние $\vec{x}_i^m$, согласно (9) одновременно должно выполняться

$$\mathrm{sign}(A_{il_i^{(m)}}) = x_i^m, \quad \sum_{j=1}^{N-1} \xi_j + x_i^m \sum_{r=1}^{L} \eta_r \geqslant \mid \sum_{r=1}^{L} \eta_r \mid .$$

В противном случае, произойдет ошибка распознавания вектор-координаты $\vec{x}_i^m$. Поскольку случайная величина $x_i^m \eta_r$ имеет то же самое распределение, что и η_r, вероятность ошибки распознавания можно представить в виде:

$$\mathrm{Pr}_i = \mathrm{Pr}\Big\{ \sum_{j=1}^{N-1} \xi_j + \sum_{r=1}^{L} \eta_r < 0 \Big\}. \tag{11}$$

Для оценки вероятности (11) при $N \gg 1$ воспользуемся известной техникой Чебышева-Чернова [17, 18]. Для начала, запишем:

$$\mathrm{Pr}_i \leqslant \mathrm{Pr}\Big\{ \sum_{j=1}^{N-1} \xi_j + \sum_{r=1}^{L} \eta_r \leqslant 0 \Big\} = \mathrm{Pr}\Big\{ - \sum_{j=1}^{N-1} \xi_j - \sum_{r=1}^{L} \eta_r \geqslant 0 \Big\}.$$

Далее, используя экспоненциальные оценки чебышевского типа [24], для любого положительного $z > 0$, получаем:

$$\mathrm{Pr}_i \leqslant \overline{\exp\Big[z \Big(- \sum_{j=1}^{N-1} \xi_j - \sum_{r=1}^{L} \eta_r \Big) \Big]} = \Big[\overline{\exp(-z\xi_j)} \Big(\overline{\exp(-z\eta_r)} \Big)^{p-1} \Big]^{N-1}.$$

Черта сверху означает усреднение по всем возможным реализациям, а последнее равенство следует из независимости случайных величин ξ_j и η_r.

С учетом (10), легко получить выражения для средних

$$\overline{\exp(-z\xi_j)} = (1-a)(1-b)e^{-z} + b + a(1-b)e^{z},$$

$$\overline{\exp(-z\eta_r)} = e^{-z}/2q^2 + 1 - 1/q^2 + e^{z}/2q^2.$$

Делая здесь замену переменных $e^z = y$ и вводя функции $f_1(y)$ и $f_2(y)$,

$$f_1(y) = a(1-b)y + b + (1-a)(1-b)/y,$$

$$f_2(y) = 1/2q^2\,(y + 1/y) + 1 - 1/q^2,$$

получаем, что при любом положительном y для Pr_i справедливо:

$$\mathrm{Pr}_i \leqslant \left[f_1(y) f_2^{p-1}(y) \right]^{N-1} . \tag{12}$$

Для получения наинизшей оценки вероятности Pr_i необходимо отыскать значение переменной y, минимизирующее правую часть (12). Это приводит к уравнению:

$$(p-1)(y^2-1) + \frac{a(1-b)y^2 - (1-a)(1-b)}{a(1-b)y^2 + by + (1-a)(1-b)} [y^2 + 2(q^2-1)y + 1] = 0 .$$

В случае $p \gg 1$ интересующий нас корень этого уравнения, с точностью до членов более высокого порядка малости по $1/p$, равен

$$y_1 = 1 + \frac{q^2(1-2a)(1-b)}{(p-1)} .$$

Подставляя y_1 в правую часть (11), окончательно получаем оценку для вероятности неправильного распознавания вектор-координаты $\vec{x}_i^m$:

$$\mathrm{Pr}_i \leqslant \left[1 - \frac{q^2(1-2a)^2(1-b)^2}{2(p-1)} \right]^{N-1} \cong \exp\left[-\frac{N(1-2a)^2}{2p} \cdot q^2(1-b)^2 \right] .$$

В результате, для вероятности ошибки распознавания паттерна X^m имеем:

$$\mathrm{Pr}_{err} = N \exp\left[-\frac{N(1-2a)^2}{2p} \cdot q^2(1-b)^2 \right] . \tag{13}$$

С ростом N вероятность ошибки распознавания стремится к нулю, если p растет медленнее, чем

$$p_c = \frac{N(1-2a)^2}{2\ln N} \cdot q^2(1-b)^2 . \tag{14}$$

Оценку (14) можно рассматривать как *асимптотически достижимую емкость памяти* ПНС.

При $q = 1$ выражения (13) и (14) превращаются в известные результаты для модели Хопфилда (в этом случае $b = 0$). С ростом q экспоненциально уменьшается вероятность ошибки распознавания (13) — существенно растет *помехоустойчивость сети*. Одновременно, пропорционально q^2 растет

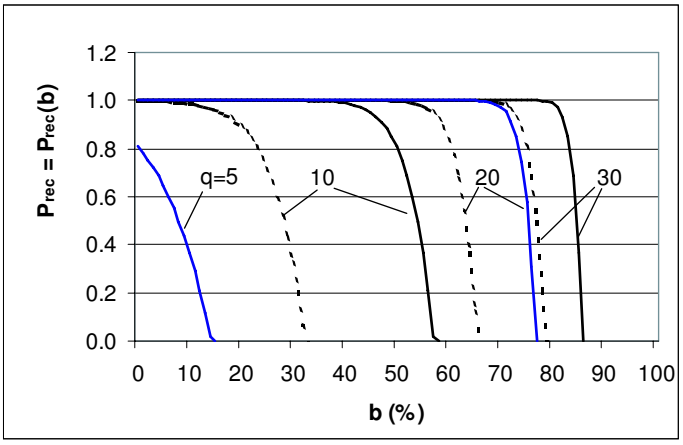

Рис. 1. Вероятность распознавания паттерна $\mathbf{P}_{rec} = 1 - \mathrm{Pr}_{err}$ как функция частотного шума b (в процентах) для различных значений $q = 5, 10, 20, 30$ при числе паттернов, вдвое большем числа нейронов, $\alpha = p/N = 2$. Здесь: сплошная линия — ПНС, штриховая линия — Поттс-стекольная нейросеть.

и емкость памяти (14). В отличие от модели Хопфилда, оказывается возможным эффективное запоминание большего, чем N, числа паттернов p.

На рис. 1, для различных значений q, сплошной линией показана зависимость вероятности правильного распознавания $\mathbf{P}_{rec} = 1 - \mathrm{Pr}_{err}$ от величины частотного шума $\mathbf{b} = b \cdot 100\%,\ b \in [0, 1]$ при числе паттернов, вдвое большем числа нейронов ($p = 2N$) и $a = 0$. Мы видим, что для $q = 20$ практически со стопроцентной вероятностью будет правильно восстановлен любой паттерн, зашумленный не более, чем на 70%, а для $q = 30$ — любой паттерн, зашумленный не более, чем на 85%. Вообще, если уровень шума b меньше критического значения $b_c = 1 - 2/q\sqrt{p/N}$, ПНС практически со стопроцентной вероятностью восстановит зашумленный паттерн. Если же $b > b_c$, вероятность правильного распознавания паттерна стремится к нулю. Машинные эксперименты подтверждают эти оценки.

Аналогично можно оценить емкость памяти Поттс-стекольной нейросети. Проводя вычисления, подобные тем, что делались выше, для рандомизированного набора паттернов получим:

$$\mathrm{Pr}_{err} = N \exp\left[-\frac{N}{2p}\frac{q(q-1)}{2}(1-\bar{b})^2\right], \quad \bar{b} = \frac{q}{q-1}\, b\,, \tag{15}$$

$$p_c = \frac{N}{2\ln N}\frac{q(q-1)}{2}\,(1-\bar{b})^2\,. \tag{16}$$

При $q = 2$ эти выражения переходят в известные оценки для модели Хопфилда. В то же время, при $q \gg 1$ емкость памяти Поттс-стекольной нейросети в два раза меньше емкости памяти ПНС (ср. (16) с (14) при $a = 0$). Эта двойка становится существенной, когда речь заходит о помехоустойчивости сети — для вероятности неправильного распознавания эта двойка попадает в показатель экспоненты (см. (13) и (15)), что ведет к заметному снижению помехоустойчивости Поттс-стекольной модели по сравнению с ПНС (особенно, в области умеренных значений $q \sim 10$). Это хорошо видно на рис. 1, где для Поттс-стекольной модели штриховой линией показана зависимость вероятности правильного распознавания $\mathbf{P}_{rec}$ от величины частотного шума $\mathbf{b}$ при тех же условиях, что и для ПНС (сплошная линия).

В заключение заметим, что в основе хороших емкостных характеристик обеих рассмотренных нами моделей лежит одно и тоже — высокая степень фильтрации внутренних шумов как на стадии передачи сигналов по межсвязям, так и в процессе выполнения решающего правила. Это приводит к уменьшению дисперсии внутреннего шума на величину порядка q^2, что и обеспечивает успех. В других q-нарных моделях ассоциативной памяти такая фильтрация полностью отсутствует. Более того, с ростом q все больше дробится шкала, градуирующая различные состояния нейронов (см. (1)), и все меньше становится разница между двумя их различными состояниями. В то же время, дисперсия шума остается практически неизменной. Это и приводит к уменьшению емкости памяти этих моделей по сравнению с моделью Хопфилда.

Работа выполнялась при финансовой поддержке РФФИ (гранты 02-01-00457 и 01-01-00090) и программы «Интеллектуальные компьютерные системы» (проект 2.45).

Литература

1. *Rieger H.* Storing an extensive number of grey-toned patterns in a neural network using multistate neurons // *J. Phys. A*, v. 23, pp. L1273–L1279 (1990).
2. *Bolle D., van Mourik J.* Capacity of diluted multi-state neural networks // *J. Phys. A*, v. 27, pp. 1151–1162 (1994).
3. *Bolle D., Shim G. M., Vinck B., Zagrebnov V. A.* // *J. Stat. Phys.*, v. 74, p. 565 (1987).
4. *Noest A. J.* // *Europhys. Letters*, v. 6, p. 469 (1988).
5. *Noest A. J.* Discrete-state phasor neural networks. // *Phys. Rev. A*, v. 38, pp. 2196–2199 (1988).
6. *Cook J.* The mean-field theory of a Q-state neural network model // *J. Phys. A*, v. 22, pp. 2000–2012 (1989).
7. *Gerl F., Bauer K., Krey U.* // *Z. Phys. B*, v. 88, p. 339 (1992).
8. *Nakamura Y., Torii K., Munaka T.* Neural-network model composed of multidimensional spin neurona // *Phys. Rev. B*, v. 51, pp. 1538–1546 (1995)
9. *Kanter I.* Potts-glass models of neural networks. // *Phys. Rev. A*, v. 37, pp. 2739–2742 (1988).
10. *Vogt H., Zippelius A.* Invariant recognition in Potts glass neural networks // *J. Physics A*, v. 25, pp. 2209–2226 (1992).
11. *Bolle D., Dupont P., van Mourik J.* Stability properties of Potts neural networks with biased patterns and low loading // *Journal of Physics A*, v. 24, pp. 1065–1081 (1991).
12. *Bolle D., Dupont P., Huyghebaert J.* Thermodynamics properties of the q-state Potts-glass neural network // *Phys. Rew. A*, v. 45, pp. 4194–4197 (1992).
13. *Bolle D., Jongen G., Shim G. M.* Q-Ising neural network dynamics: A comparative review of various architectures // *Cond-mat* /9907390 (1999).
14. *Крыжановский Б. В., Микаэлян А. Л.* О распознающей способности нейросети на нейронах с параметрическим преобразованием частот //*ДАН (мат.-физ.)*, т. 65(2), с. 286–288 (2002).
15. *Fonarev A., Kryzhanovsky B. V. et al.* Parametric dynamic neural network recognition power // Optical Memory and Neural Networks, v. 10(4), pp. 31–48 (2001).
16. *Крыжановский Б. В., Литинский Л. Б.* Векторные модели ассоциативной памяти // *Теоретическая и математическая физика*, 2003 (в печати)
17. *Chernov N.* A mesure of asymptotic efficiency for tests of hypothesis based on the sum of observations // *Ann. Math. Statistics*, v. 23, pp. 493–507 (1952).
18. *Kryzhanovsky B. V., Mikaelyan A. L., Koshelev V. N., Fonarev A.* On recognition error bound for associative Hopfield memory // Optical Memory and Neural Networks, v. 9(4), pp. 267–276 (2000).

19. *Potts R.B.* // *Proc. Camb. Phil. Soc.*, v. 48, p. 106 (1952).

20. *Wu F. Y.* The Potts model // *Review of Modern Physics*, v. 54(1), pp. 235–268 (1982).

21. *Бэкстер Р.* Точно решаемые модели статистической физики. – М.: Мир, 1985.

22. *Amit D., Gutfreund H., Sompolinsky H.* Storing Infinite Numbers of Patterns in a Spin-Glass Model of Neural Networks // *Phys. Rev. Lett.*, v. 55, pp. 1530–1533 (1985); *See also*: Information storage in neural networks with low levels of activity // *Phys. Rev. A*, v. 35, pp. 2293–2303 (1987).

23. *Bloembergen N.* Nonlinear optics. 1966.

24. *Крамер Г.* Математические методы статистики. – М.: Мир, 1975. – 206 с.

Борис Владимирович КРЫЖАНОВСКИЙ, заместитель директора Института оптико-нейронных технологий РАН (Москва), доктор физико-математических наук. Область научных интересов: нейронные сети, квантовая механика, нелинейная оптика. Имеет более 80 научных публикаций.

Леонид Борисович ЛИТИНСКИЙ, старший научный сотрудник, кандидат физико-математических наук, заведующий сектором Динамических нейронных сетей в Институте оптико-нейронных технологий РАН, Москва. Область научных интересов: ассоциативные нейронные сети, распознавание образов, многоэкстремальные задачи. Имеет более 30 научных публикаций.

Н. Г. МАКАРЕНКО
Институт математики, Алма-Ата, Казахстан
E-mail: makarenko@math.kz

ЭМБЕДОЛОГИЯ И НЕЙРОПРОГНОЗ

Аннотация

Лекция представляет собой попытку понятно рассказать о методах алгоритмического моделирования — *Эмбедологии* или реконструкции модели из наблюдаемых скалярных временных рядов. Топологическое вложение ряда в евклидово пространство подходящей размерности приводит к схеме нелинейного локального или глобального прогноза. Последний сводится к задаче наилучшей аппроксимации нелинейной функции от многих переменных. Оптимальным аппроксиматором, во многих случаях, является нейронная сеть.

N. MAKARENKO
Institute of Mathematics, Kazakhstan, Alma-Ata
E-mail: makarenko@math.kz

EMBEDOLOGY AND FORECASTING BY NEURAL NETWORKS

Abstract

This lecture represents an attemt of comprehensible introduction into algorithmic modelling methods — Embedology reconstruction of a model from observed scalar time series. A topological time series embedding in Euclidian space of appropriate dimension in a scheme of nonlinear local or global forecast results. The last is reduced to a task of best approximation of nonlinear multivariable function. In such cases a neural network is often an optimal approximator.

УДК 004.032.26 (06) Нейронные сети

Введение

> Подходите к вашим задачам с правильного конца и начинайте с ответов. Тогда в один прекрасный день вы, возможно, найдете правильный вопрос.
>
> *Р. ван Гулик*
> *«Отшельник в журавлиных перьях»*

Непостижимая эффективность математики в естественных науках[1] [2] обязана главным образом моделям, заданным дифференциальными уравнениями. Аналитика, однако, избыточна в смысле определений: она содержит случаи, которым не соответствует никакая реальность и не дает нам средств, с помощью которых мы могли бы различить *действительное* и *возможное*. Для того чтобы сделать это, приходится обращаться к опыту и те, кто возмущается такими экскурсами за пределы «фундаментальной науки» — явные лицемеры! Почему же тогда не попытаться построить или точнее реконструировать модель прямо из наблюдений? В истории науки уже были попытки получить модель непосредственно из данных. Наиболее удачной из них считается гелиоцентрическая модель эллиптического движения планет. Три знаменитых закона были получены *Иоганном Кеплером* из Вюртенберга (1571–1630 гг.) в результате чудовищных по объему ручных вычислений, проделанных над наблюдениями пражского астронома *Тихо Браге* (1546–1601 гг.) Модель Кеплера — не только замечательный пример получения *явных знаний из таблиц данных*. Она имела замечательные предсказательные возможности, реализованные Кеплером в так называемых *Рудольфовых таблицах* — наперед рассчитанных эфемеридах нескольких планет.

На практике наблюдения чаще всего представлены скалярными временными рядами. Можно предположить, что отсчеты ряда являются *нелинейной проекцией*[1] движения фазовой точки некоторой динамической

[1] Если ссылка на примечание представляет собой число, заключенное в круглые скобки, например, (3), то она обозначает номер примечания, помещенного в конце данного текущего раздела. Ссылка в квадратных скобках, например, [3], обозначает номер дополнения в конце лекции. Ссылка в виде числа, не заключенного в скобки — обычное подстраничное примечание. — Прим. ред.

[2] См. статью *Е. Вигнера*, с тем же названием в сборнике [59].

системы, продуцирующей ряд — нечто вроде теней на стене пещеры, если следовать известной аллегории *Платона* [3] [2].

В таком случае:

- *Что можно сказать о неизвестной системе на основе созерцания динамики ее тени?*
- *Можно ли по проекции восстановить образ системы и, если да, то в каком смысле?*

На первый взгляд кажется, что эти вопросы бессмысленны. Действительно, располагая, например, решением трехмерной системы уравнений лишь для одной, скажем $x(t)$, координаты, мы не имеем никакой информации о функциях $y(t)$ и $z(t)$. Однако, в 1980 году *Паккард* с коллегами [17] предположил, что для *топологической идентификации* [3] такой системы достаточно измерить одновременно три *произвольных* переменных, связанных с движением системы. Для этого они должны быть независимыми в некотором (*операционном*) смысле. Такими «координатами» могут быть, например, значение единственной переменной x и ее первых двух производных. Тройка $(x, \dot{x}, \ddot{x})$ заменяет оригинал $(x(t), y(t), z(t))$ так, что построенный в новых «координатах» фазовый портрет системы оказывается похожим на оригинал *с точностью до гримасы клоуна* — непрерывных, но нелинейных деформаций. Эта эвристическая идея была формализована в 1981 году в знаменитой статье *Ф. Такенса* [20] в форме теоремы о *типичном вложении* временного ряда в R^n. Так возник новый способ построения модели из наблюдаемого сигнала (или реализации), который тут же инициировал совершенно новую область численных методов топологической динамики — *эмбедологию* [4] [4]. С ее помощью наконец-то удалось подойти к проблеме моделирования динамики с «правильного конца»: модель начиналась с «ответа» — наблюдений, а затем ставился «правильный» вопрос: что их продуцирует?

Эмбедология позволила сразу же получить нелинейный многомерный вариант традиционного авторегрессионного прогноза временных рядов. Поиск предиктора свелся к проблеме аппроксимации нелинейной функции

[3] *Платон* в одном из своих знаменитых диалогов рассказал притчу об узниках, находящихся в пещере и прикованных там так, что они могут наблюдать только тени от проходящих людей и проносимых предметов. В связи с этим задается вопрос — могут ли узники понять, что происходит снаружи, наблюдая за этими тенями? И *Платон* отвечает на этот вопрос утвердительно.

[4] Название данной области [20] происходит от английского слова *embedology*, произведенного, в свою очередь, от слова *embedding* (вложение).

нескольких переменных. Для глобального варианта прогноза таким предиктором является нейронная сеть по своему *modus operandi* [5]. Кулинарная книга *Эмбедологии* предлагает рецепты острых «ресторанных» блюд, содержащих компоненты из дифференциальной топологии, фрактальной геометрии и эргодической теории гладких динамических систем. Они выглядят экзотикой с точки зрения практика-экспериментатора, привыкшего к меню традиционной «линейной» столовой, а *Пособия для Чайников* по приготовлению «блюд» эмбедологии нередко приводят в уныние. Поэтому, настоящая *Лекция* предлагает вместо рецептов — принципы, по которым готовятся основные деликатесы, включая нейропрогноз.

Автором намеренно был выбран стиль, являющийся компромиссом между лекцией и конспектом из различных источников. Лекционный стиль позволяет не слишком заботиться о строгости изложения, прибегая в особенно туманных местах к уловке *св. Августина*: *«Что мне, если кто не понимает!»* Форма конспекта, с другой стороны, избавляет от необходимости достичь определенной цели на заданном числе страниц. Издержки при этом сводятся к тому, что какие-либо авторские претензии на систематическую и безусловную завершенность будут по меньшей мере самообманом.

Автор искренне благодарен своим юным коллегам *Оле Круглун*, *Кате Данилкиной*, *Ерболу Куандыкову* и *Светлане Ким*, помощь и терпение которых позволило закончить рукопись.

Отображения, функции и типичность

> В отблесках пространств и отображений очертания мысленных структур, слов и, быть может, даже чисел становятся нечеткими, приобретая, что несомненно и крайне существенно, цвет морской волны.
>
> *А. Лиепиньш*
> *«Топологические пространства и их отображения»*

[5] Способ действия (лат.). — (Прим. ред.)

Мы будем использовать следующие стандартные обозначения:

- Запись $x \in X$ означает, что x принадлежит X.
- Запись $\{x|d(a,x) < r\}$ читается: множество таких x, для которых расстояние от точки a меньше r.
- Запись $A \subset B$ $(A \subseteq B)$ означает, что A содержится в B (содержится или совпадает с B).
- *Объединением* множеств A и B называют множество $C = A \cup B$, состоящее из элементов $c \in C$, каждый из которых принадлежит A или (и) B.
- *Пересечением* A и B называют множество $C = A \cap B$, каждый элемент c которого входит и в A, и в B.

Соответствие между различными множествами удобно описывать функциями или отображениями: $f(*)$ или $f : X \to Y$. Подмножество $\{f(x)\} \in Y$ — это область значений функции, а $\{x\} \in X$ — ее область определения.

Метрическое пространство $(\mathbf{M}, d)$ — это множество $\mathbf{M}$ вместе с вещественнозначной функцией d, которая удовлетворяет аксиомам метрики. Для $\mathbf{M} \equiv R^n$, где R^n — n-мерное евклидово пространство (обозначается как $n = \dim M$), часто используют L_2-метрику

$$d(x,y) = \Big[\sum_i (x_i - y_i)^2\Big]^{1/2}.$$

Пусть $(\mathbf{M}, d)$ — метрическое пространство и $r > 0$ — действительное число. Тогда *открытым* шаром в точке $a \in \mathbf{M}$ называется подмножество $B(a,r) \subset \mathbf{M}$: $B(a,r) = \{x \in \mathbf{M} \,|\, d(a,x) < r\}$. Шар называется *замкнутым*, если $d(a,x) \leqslant r$.

Отображение $f : A \to B$ называют *сюръекцией* или *отображением на*, если образ всего A совпадает с B, т. е. каждый элемент из B является образом по крайней мере одного элемента из A. Если $A = B$, то элемент x, удовлетворяющий условию $x = f(x)$ $(f : A \to A)$, называют *неподвижной точкой* отображения f. Отображение $\varphi : A \to B$ называют *инъективным* или *взаимно однозначным*, если $\forall y \in B$ существует не более одного элемента $x \in A$ такого, что $y = \varphi(x)$. Тогда $\varphi(x) = \varphi(y)$ влечет $x = y$. Если $B = R^n$, то f называют *функцией*: впрочем, термины «функция» и «отображение» обычно считают синонимами. Обратную к $f(x)$ функцию будем обозначать $f^{-1}(x)$. Отображение, являющееся инъективным и сюръективным, называется *биекцией*. Теперь сформулируем условие непрерывности:

отображение $f : X \to Y$ *непрерывно* в $x \in X$, если для любой окрестности V, открытой в Y и содержащей точку $f(x)$, найдется такая открытая в X окрестность $U(x)$, что $f(U) \subset V$. Для метрических пространств это определение совпадает с классическим, записанным на $(\varepsilon-\delta)$-языке. Отображение φ открытого множества U из R^n в R^m называется *гладким*, если оно имеет непрерывные частные производные всех порядков: говорят, что φ имеет класс C^r, если это справедливо для производных до порядка r включительно. В более общем случае, когда область определения f не открыта, отображение $\varphi : X \to R^n$ называют гладким, если для каждой точки $x \in X$ существует открытая окрестность $U \subset R^n$ и гладкое отображение $F : U \to R^n$ такое, что $F = f$ на $U \cap X$ (рис. 1).

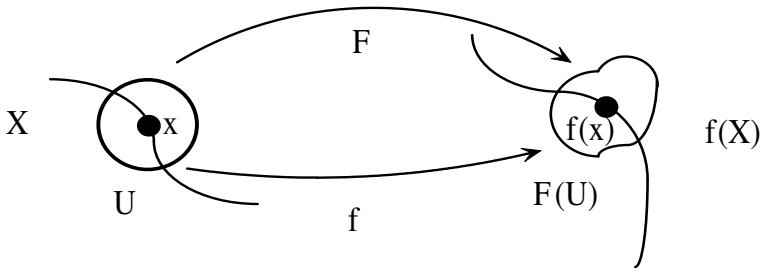

Рис. 1. Определение гладкого отображения

Известно важное свойство гладких функций: если дифференциал такой функции *в точке* x имеет некоторое свойство, то сама функция будет иметь такое же свойство по меньшей мере *в окрестности* точки x. Например, если производная $f'(x_0)$ функции f положительна, $f'(x_0) > 0$, то в малой окрестности точки x_0 функция f возрастает [(1)]. Это свойство называют обычно *принципом общего положения* или *принципом (догмой) линеаризации*, поскольку дифференциал функции является линейной частью отображения [(2)]. Принципу общего положения удовлетворяет не каждая функция, поэтому важно сделать математически точными утверждения о выполнимости некоторого свойства для «почти каждой» функции [(3)]. Вообще говоря, любые свойства математических объектов можно классифицировать как *обычные (usual)* и *необычные (unusual)*. Стандартный прием выразить распространенность некоторого свойства, выполняющегося на подмножестве $A \subseteq B$, сводятся к тому, чтобы показать, что обычные свойства соответствуют подмножеству с большей мерой. Однако, пространство функций

является бесконечномерным и поэтому здесь невозможно прибегнуть к замене выражения *почти каждая* на утверждение *с вероятностью единица*. Другой способ сделать это основан на топологии.

Рассмотрим два множества A и B таких, что $A \subset B$. Говорят, что A *плотно* в B, если:

- каждая точка B либо принадлежит A, либо является точкой накопления A, (т. е. каждая точка x, принадлежащая или не принадлежащая A, содержит в своей окрестности по меньшей мере одну точку $y \neq x \mid y \in B$), или
- каждое открытое подмножество точек B содержит точку A.

Пример 1. Множество нецелых чисел плотно и открыто в R, а множество рациональных чисел плотно, но не открыто на вещественной оси.

Кажется разумным связать «обычные» свойства с множеством A, которое плотно в B. Однако, такой путь не приводит к результату, совместимому с интуицией: в некоторых случаях дополнение A в B тоже плотное подмножество в B.

Пример 2. Подмножество рациональных чисел плотно в R, но таким же является и его дополнение — подмножество иррациональных чисел. Однако, интуитивно ясно, что иррациональное число на R — более обычный объект. Чтобы адекватно отслеживать такие ситуации, было введено понятие *остаточного множества (residual set)* [4]. Подмножество A *остаточное* в B, если оно является пересечением конечного или счетного числа открытых и плотных подмножеств в B. Любое счетное пересечение остаточных множеств плотно, и любое «респектабельное» [6] пространство содержит остаточное множество. Такие пространства называют *пространствами Бэра*. Например, любое полное метрическое пространство является бэровским. В пространстве Бэра дополнение остаточного множества не может быть остаточным: в противном случае пустое множество пришлось бы считать плотным.

Пример 3. Пусть $Q = \{q_1, q_2, \ldots q_n, \ldots\}$ — подмножество рациональных чисел в R, а $U_n = R - \{q_n\}$ открытое плотное подмножество. Тогда иррациональные числа — это счетное пересечение всех U_n и, следовательно, остаточное подмножество в R; с другой стороны очевидно, что Q не является таковым.

[6] Т. е. «хорошее» пространство, в котором есть все для работы — например, полное, метрическое. Или банахово, с нормой и метрикой.

Современная метрическая точка зрения на типичность заключается в идее *превалентности*: свойство превалентно [5] для отображений из некоторого гладкого конечнопараметрического семейства, если множество значений параметра, соответствующих отображениям, для которых свойство не выполняется, имеет лебегову меру нуль.

Ради простоты мы будем понимать выражение «с точностью до предположения о типичности» в следующем не очень строгом смысле. Свойство типично для функции f, если оно выполняется в каждой точке соответствующего пространства и открытой окрестности этой точки. Кроме того, если оно не выполняется, например, в точке p, всегда можно найти достаточно близкую к p точку, в которой это свойство выполняется. Аналогом является трактовка: свойство *типично*, если оно справедливо для f, либо, если это не так, становится справедливым для $f + \delta f$, где порядок возмущения δf заранее оговаривается.

Примечания

1. Более точно, если дифференциал функции f в точке x_0 строго возрастает, то и сама функция f строго возрастает в окрестности точки x_0.

2. Тривиальный пример информативности линеаризации: заменим в уравнении $f(x, y) = 0$ функцию f на df. Уравнение $f'_x dx + f'_y dy = df(x_0, y_0) = 0$ — линейное и условие $f'_y(x_0, y_0) = 0$ эквивалентно существованию его решения. Но именно тогда $y = \varphi(x)$ существует и является решением исходного уравнения.

3. Примером могут служить утверждения: почти каждая интегрируемая функция $f : [0, 1] \to R$ удовлетворяет условию

$$\int_0^1 f(x)dx \neq 0$$

   или почти каждая непрерывная функция $f : [0, 1] \to R$ нигде не дифференцируема.

4. Иногда его называют *густым* множеством.

5. Точнее, пусть $\mathbf{V}$ — полное метрическое пространство. Запись $S + b$ означает трансляцию множества $S \subset B$ на вектор $b \subset B$. Говорят,

что мера μ *трансверсальна* борелеву множеству $S \subset B$, если *(i)* существует компактное множество $A \subset B$, для которого $0 < \mu(A) < \infty$ и *(ii)* $\mu(S + b) = 0,\ \forall\ b \in B$. Борелево множество $S \subset B$ называется *недостаточным* (*shy*), если существует мера, трансверсальная S. Дополнение к множеству *shy* называют *превалентным*. Все превалентные множества плотны.

Путеводитель по литературе. Краткую сводку определений, касающихся функций и отображений, можно найти в карманном справочнике [1]. Необходимые детали содержатся в учебниках [2, 3] и популярном курсе [4]. Доступное введение в идеи типичности можно найти в книге [13] и приложении к монографии [12]. Строгое определение превалентности приведено в [14].

Многообразия

Облик их близок к облику просветленных стихиалей, но форма не струистая, как у тех, и лишенная способности телесного взаимопроникновения.

Д. Андреев
«Роза Мира»

Рассмотрим группу преобразований плоскости R^2

$$x' = ax + by + c\,, \quad y' = px + qy + s\,, \quad ||A|| = \begin{pmatrix} a & b \\ p & q \end{pmatrix}\,,$$

заданную матрицей $\|A\|$. Если $\|A\|$ — ортогональна, то обычную геометрию можно определить как совокупность тех свойств, которые сохраняются (инвариантны) относительно действия этой группы. Например, заведомо сохраняются расстояния между точками и углы между прямыми. *Аффинная геометрия* получается, если ортогональность матрицы $\|A\|$ заменить на более скромное требование $\det A \neq 0$. Следующим обобщением будет группа, заданная в виде:

$$x' = f(x, y),\ y' = g(x, y)\,,$$

где f и g — нелинейные, но C^r-гладкие функции. Если же такими свойствами обладают обратные преобразования (f^{-1}, g^{-1}) [7], мы получим *геометрию гладких многообразий*. Это совокупность свойств, инвариантных относительно растяжений и деформаций, т. е. *гомеоморфизмов* (1). В этой геометрии куб, шар и конус неразличимы. Здесь тело человека можно превращать в тела жирафа, зайца или верблюда — все они здесь «топологические синонимы».

Рассмотрим сферу S^2 и бесконечное число ее копий, не различимых с точностью до гомеоморфизмов. Весь этот класс эквивалентных объектов называют *многообразием*. Привычная сфера лишь типичный представитель этого класса, его конкретная реализация в евклидовом пространстве R^3. Визуализация сферы, к которой мы привыкли, это *поверхность уровня* для функции $f(x, y, z) = x^2 + y^2 + z^2$, т. е. f^{-1}.

Важно подчеркнуть, что форма вовсе не обязательный атрибут многообразия, а просто его геометрическая визуализация (2). Примером многообразия, *негомеоморфного* S^2, является тор T^2: действительно, не существует непрерывных деформаций, переводящих тор в сферу. Гомеоморфная копия T^2 может выглядеть совсем не похожей на «бублик». В качестве примера возьмем единичный квадрат

$$X = \{(x, y) \,|\, 0 \leqslant x, y \leqslant 1\}.$$

Эта запись означает, что X — множество пар точек x и y таких, что любая из них имеет координаты, заключенные в отрезке $[0, 1]$. Теперь «склеим» правую сторону x с левой (это запишем, как $(0, y) \sim (1, y)$).

Мы не делаем этого в действительности, а считаем, что указанные пары сторон совпадают, т. е. они «отождествлены». После этого отождествим верхнюю и нижнюю окружности полученного цилиндра. То, что получилось в итоге — это и есть копия тора (3). Сферу также можно получить «склейкой» круга или квадрата, у последнего надо отождествить пары смежных сторон — вместо клея можно представить себе «замки-молнии». Нетривиальная склейка «по диагонали»: $(0, y) \sim (1, y)$ и $(x, 0) \sim (1 - x, 1)$ приводит к экзотическому преобразованию — «бутылке Клейна». Чтобы выполнить второе отождествление, нам необходимо попасть в R^4!

Гомеоморфизмы позволяют «выпрямлять» куски многообразия на области в R^n: небольшой участок сферы хорошо моделируется плоской областью. Эта идея давно использована в картографии: любой известный

[7] Их существование гарантирует нам группа.

способ построения карт — это проекция кусков сферы S^2 на плоские области R^2. Почти очевидно, что невозможно получить непрерывное и взаимно однозначное отображение всей сферы на один связный кусок R^2. Поэтому сферу разбивают на отдельные куски, которые частично перекрываются, и для каждого из них строится карта. Перекрытия позволяют согласовать отдельные карты географического атласа при их «склейке» в сферу.

Согласование — это установление взаимно однозначных соотношений между разными картами и, тем самым, разными областями S^2. Даже для окружности недостаточно одной карты. Действительно, можно попытаться использовать уравнение $x^2 + y^2 = 1$, чтобы картографировать S^1 одним связным куском R^1. Однако, небольшое размышление показывает, что такая процедура возможна лишь для верхней и нижней полуплоскости отдельно. Можно использовать полярные координаты, но и в этом случае одного «куска» оси φ недостаточно: когда точка окружности делает один оборот, φ скачком меняется на 2π. При этом нарушается непрерывность гомеоморфизма. Идею гомеоморфизма часто используют для определения многообразия: многообразием называют топологическое пространство локально гомеоморфное R^m. Так, например, двойной конус на рис. 2 (слева) не является многообразием. Действительно, окрестность общей вершины двух конусов не похожа на фрагмент плоскости, она двухлистная. Напротив, сфера (рис. 2, справа) — локально похожа на фрагмент евклидовой плоскости.

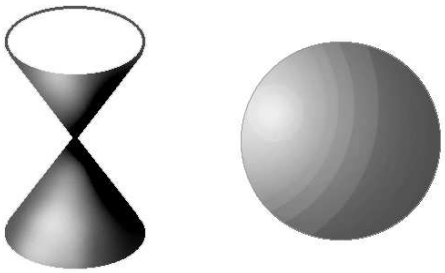

Рис. 2. Примеры топологических пространств: слева (два конуса с общей вершиной) — не многообразие; справа (сфера) — многообразие

В этом смысле гомеоморфизмы являются клеем для изготовления многообразий, которые можно считать теперь объектами, склеенными из от-

дельных кусков R^n. Изготовление многообразий таким способом сохраняет близость точек: близкие точки многообразия отображаются в близкие точки плоской карты. На карте можно выбрать подходящую систему координат и тогда понятие близости обретает смысл расстояния. Само многообразие не наделено координатами, поэтому понятие близости здесь носит несколько призрачный оттенок. Ситуация становится более определенной, если ввести понятие *окрестности*.

Окрестность точки x многообразия можно определить, указав выделенный класс подмножеств (их называют *открытыми*), все точки которых «достаточно близки к любой точке». Класс замкнут относительно операций объединения и пересечения. А именно, пересечение *конечного* числа открытых множеств и объединение *любого* их числа также принадлежит указанному классу. Именно он определяет *топологию* многообразия. Если в нем всего два объекта: все множество и пустое множество, то топология *тривиальна*. Объявив открытыми все точки множества, мы получим *дискретную топологию*. Важно, *что* объявить открытым!

Пусть многообразие — обычное R^n с евклидовой метрикой. Открытые окрестности, *индуцирующие метрическую топологию*, — это открытые шары: $\sum x_i^2 < \rho$. Не все множества позволяют ввести топологию, например, два конуса с общей вершиной на рис. 2. Окрестность общей точки не допускает «выпрямление» на плоскость. Аналог этой ситуации — множество $xy = 0$ в R^2. Попытка ввести карту, т. е. построить гомеоморфизм на R^1, ведет к нелепости: наше множество — координатные оси; и если мы отобразим открытое на оси x множество, например, $a < x < b$ в R^1, считая, что точка $(0, 0)$ принадлежит (a, b), то в R^1 не останется места для точек, содержащих $(0, 0)$ по оси y. С другой стороны, нельзя объявить открытыми множества

$$(a < x < b\,;\ y = 0) \quad \text{и} \quad (c < y < d\,;\ x = 0)\,,$$

так как их пересечение — точка $(0, 0)$ должна быть открытой. Но ее образ в R^1 — отдельная точка, т. е. замкнутое пространство! Такие множества не являются многообразиями, поскольку для них не выполняется *аксиома отделимости Хаусдорфа*, справедливая для многообразий: любая пара точек может быть разделена непересекающимися окрестностями.

Пусть M — многообразие, U — подмножество в M, а Φ — биекция U в R^n. Тогда *n-мерной картой «с» на M* называют пару (U, Φ). Здесь U — *область карты*, а Φ — *координатное отображение*. Иными словами, мы

просто приписываем каждой точке $p \in U$ n чисел:

$$(\Phi_1(p), \Phi_2(p), \ldots, \Phi_n(p)) = (x_1, x_2, \ldots, x_n)\,,$$

которые называются координатами p в карте $c = (U, \Phi)$. Отображение Φ^{-1} называют *параметризацией* U.

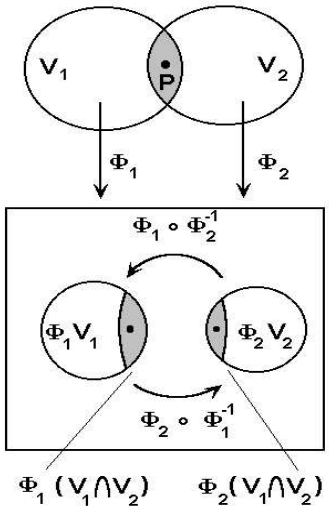

Рис. 3. Согласование карт на многообразии

Пусть точку p содержат две пересекающиеся окрестности V_1 и V_2 (см. рис. 3). Имеем две карты: $c = (V_1, \Phi_1)$ и $c' = (V_2, \Phi_2)$. Эти карты C^r-согласованы, если: $\Phi_1(V_1 \bigcap V_2)$ и $\Phi_2(V_1 \bigcap V_2)$ открыты в R^n и функции

$$\Phi_1 \circ \Phi_2^{-1} : \Phi_2(V_1 \bigcap V_2) \to \Phi_1(V_1 \bigcap V_2)$$
$$\Phi_2 \circ \Phi_1^{-1} : \Phi_1(V_1 \bigcap V_2) \to \Phi_2(V_1 \bigcap V_2)$$

являются C^r-гладкими функциями, т. е. C^r-диффеоморфизмами. Полный набор согласованных карт для всего многообразия называют C^r-*атласом*, а само многообразие C^r-дифференцируемым.

Отображение $\varphi : X \to Y$ двух гладких многообразий или двух произвольных множеств в R^n называют *диффеоморфизмом*, если φ и φ^{-1} —

гладкие гомеоморфизмы (см. рис. 4). Окружность в R^2 диффеоморфна эллипсу, но не треугольнику, которому она лишь гомеоморфна. Диффеоморфизмы определяют класс эквивалентных диффеоморфных пространств.

Введем на многообразии некоторые понятия. *Кривая на* M, проходящая через точку p — это отображение $\Im \equiv [0,1] \to M$, сопоставляющее каждой $\lambda \in \Im$ точку $c(\lambda) \in M$, так что $c(0) = p$. Ясно, что c — это параметризация: две кривые различны, даже если их образы совпадают, но одинаковым точкам p в M соответствуют разные λ.

Функция на M — это отображение $M \to R^1$, сопоставляющее каждой точке $p \in M$ вещественное число. Поскольку точка p всегда содержится в некоторой окрестности V, карта окрестности (V, Φ) превращает $f : M \to R^1$ в функцию, заданную на $R^n : f \circ \Phi^{-1} : R^n \to R^1$. Если она дифференцируема на R^n, то говорят, что она дифференцируема на M.

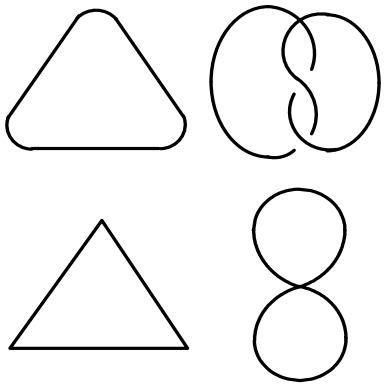

Рис. 4. Две фигуры вверху диффеоморфны; нижняя пара — нет

Определим касательный вектор к $p \in M$. Хотелось бы обойтись без координат, т. е. дать внутреннее определение. Идея стрелки, «приклеенной» к $p \in M$, дает подходящее представление, недостатком которого является необходимость иметь некоторое объемлющее пространство, в которое выходит конец стрелки.

Пусть через точку $p \in M$ проходит кривая $\Im \to M : \lambda \to c(\lambda)$ и $c(0) = p$. Будем изучать смещение точки p с помощью произвольной скалярной функции f: мы просто используем в качестве аргумента этой функции точки, принадлежащие кривой. Интуитивно ясно, что в инфинитези-

мальной окрестности p изменение функции $f(c(\lambda)) - f(c(0))$ не зависит от самой функции — это просто сдвиг p и больше ничего! Но сдвиг — это оператор, который выражает общее свойство всех кривых, проходящих через p, а именно, то что все они имеют одно и то же изменение для произвольных функций, моделирующих смещение p. Этот оператор $d/d\lambda$ называют *касательным вектором к p на M*. Он определяет класс эквивалентности кривых (т. е. то общее, что они имеют), проходящих через точку p.

Примечания

1. Точнее, *гомеоморфизмом* называется отображение f, для которого f и f^{-1} непрерывны. Пожалуй, наиболее оригинальное применение этого свойства — непрерывности гомеоморфизмов — нашел французский художник *Оноре Домье* (1808–1879 гг.). В серии карикатур он небольшими деформациями преобразовал физиономию *Луи Филиппа* в грушу! Королевский суд не смог указать, на каком именно рисунке кончается искусство и начинается оскорбление Его Величества! Рассмотрим два интервала: замкнутый — $[0, 1]$ и открытый — $(0, 1)$. Если выбросить точку 0 из первого, получится полуоткрытый интервал $(0, 1]$, который интуитивно состоит из одного куска. Но если выбросить любую точку из $(0, 1)$ — мы получим два куска. Поэтому эти два интервала не гомеоморфны.

2. *Имея дело с формами, мы приобретаем здоровое неуважение к их авторитету* ... — очень точная метафора топологии, принадлежащая *М. Шубу* [8].

3. Тор гомеоморфен следующему подпространству в R^3:

$$(x, y, z) \in R^3 \mid (\sqrt{x^2 + y^2} - 2)^2 + z^2 = 1\,.$$

Путеводитель по литературе. С основными понятиями топологии можно познакомиться в популярной книге [4], которую хорошо дополняет брошюра [5]. Строгие топологические определения можно найти в учебниках [3, 6–8, 10–11].

[8] *Майкл Шуб* (*Michael Shub*) — это тополог, работающий в области динамических систем. См. его страничку URL: http://www.research.ibm.com/people/s/shub/
Цитата взята из его письма *Ральфу Фруберту* (*Ralf Fruberg*), написанного в 1969 году; приводится по книге *М. Хирша* [6] (гл. 9, с. 244).

Погружения и вложения

> . . . и по виду их и по устроению их казалось, будто колесо находилось в колесе. . .
>
> *Книга Пророка Иезекиля*
> *«Гл. 1, ст. 16»*

Пусть $f : X \to Y$ гладкое отображение двух многообразий и $\dim X < \dim Y$. Этому отображению можно поставить в соответствие отображение касательных пространств $df : T_x(X) \to T_{f(x)}(Y)$, «приклеенных» к точкам x и $f(x)$, соответственно [1]. Его называют *дифференциалом* отображения f и обозначают df. Понятно, что в нашем примере df не может быть сюръекцией, потому что оно не в состоянии «накрыть» лишние координаты в Y. Самое большее, на что мы можем претендовать, это на свойство инъекции df. В этом случае, f называют *иммерсией* или *погружением* X в Y. Например, окружность $S^1 : r(t) = (\sin t, \cos t)$ можно погрузить в R^2 так, что ее образом будет «восьмерка»: $r(t) = (\sin t, 2 \sin 2t)$. При этом взаимно-однозначное соответствие точек образа и прообраза, конечно, отсутствует (рис. 5), но каждому вектору скорости $\dot{r}(t)$ на S^1 соответствует единственный касательный вектор «восьмерки».

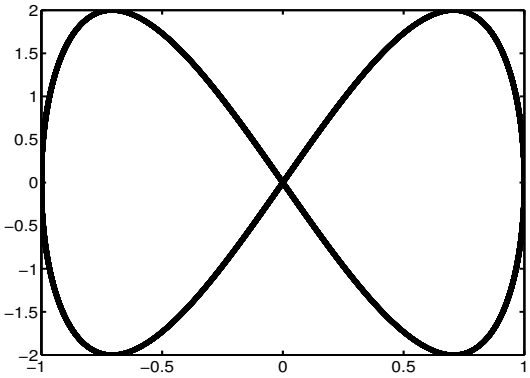

Рис. 5. Погружение окружности в плоскость

Можно изменить параметризацию, используя замену координат

$$t \to 2\operatorname{arctg} u : \tilde{r}(u) = (\sin(2\operatorname{arctg} u), 2\sin(4\operatorname{arctg} u))$$

и тогда кривая пройдет через начало координат только один раз, однако, взаимнооднозначное соответствие между кривой и окружностью отсутствует — у кривой появились свободные концы.

Каноническая иммерсия получается, если евклидово пространство R^k погрузить в R^m $(k < m)$. При этом точка $x \in R^k$ *линейно* отображается в точку $f(x) \in R^m$:

$$(x_1, x_2, \ldots, x_k) \to (x_1, x_2, \ldots, x_k, \underbrace{0, \ldots, 0}_{m-k}) .$$

В этом случае образ иммерсии — подпространство в R^m. К сожалению, для многообразий, которые являются только локально евклидовыми, это не так. Например, в случае «восьмерки» окрестность точки самопересечения не является открытой в R^2. Если использовать параметризацию:

$$\hat{r}(v) = (-\sin(2\operatorname{arctg} v), 2\sin(4\operatorname{arctg} v)) ,$$

образы $\tilde{r}(u)$ и $\hat{r}(v)$ совпадают, однако, композиция $r \circ \hat{r}^{-1} : R \to R$ не является непрерывным отображением.

Отображение

$$\varphi(t) = ((1 + e^{-t}) \cos t,\ (1 + e^{-t}) \sin t)$$

определяет иммерсию вещественной прямой R в плоскость R^2: образ $\varphi(e)$ накручивается на единичную окружность $x^2 + y^2 = 1$ (см. рис. 6).

В общем случае гладкое отображение F компактного гладкого многообразия A называют *иммерсией*, если дифференциал отображения $dF(x)$ является взаимно-однозначным в каждой точке A [2]. При этом безразлично, имеет ли само F такое свойство.

Дифференцируемое *вложение* A — это гладкий диффеоморфизм из A на собственный образ $F(A)$, т. е. гладкое взаимно-однозначное отображение, обратное для которого $F^{-1}(A)$ также гладко. Если A — компактное многообразие [3], то отображение F является вложением тогда и только тогда, когда F — взаимно-однозначная иммерсия [4]. Для *топологического вложения* диффеоморфизм заменяется на гомеоморфизм.

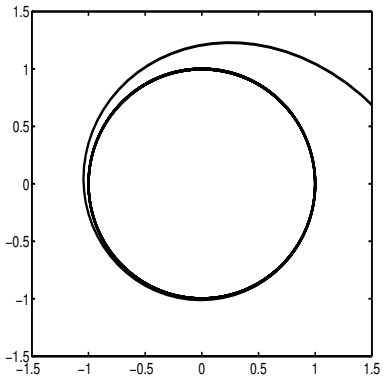

Рис. 6. Иммерсия прямой в плоскость

Знаменитая *теорема Уитни* показывает, что всякое компактное многообразие можно считать подмногообразием евклидова пространства достаточно большой размерности. Теорема Уитни гласит:

Пусть M — гладкое компактное k-мерное многообразие. Тогда существует вложение $F : M \to R^n$ для подходящего выбора размерности n, а именно $n = 2k + 1$.

Важное замечание. Вложение может быть реализовано *почти любой* функцией, удовлетворяющей упомянутым условиям. Это означает, что если A — гладкое многообразие размерности k, то множество отображений в R^{2k+1}, которые являются вложениями A, образуют остаточное (густое) множество в классе всех непрерывных отображений [5].

Условие теоремы Уитни для размерности R^n можно трактовать следующим образом [6]: для взаимной однозначности образа и прообраза отображения следует задать k координат; для такого же соответствия между касательными векторами необходимо еще k чисел. Наконец, единица гарантирует дополнительную степень свободы для того, чтобы избежать ситуацию с самопересечением [7]. Геометрические истоки условия Уитни восходят к *догме трансверсальности*, о которой рассказывается в следующем разделе. Существует важное обобщение теоремы Уитни на случай, когда A не является многообразием. Эта ситуация встречается, например, для *фрактальных аттракторов*, для которых не существует разумного определения топологической размерности. Вместо нее используют так называемую *box-*

размерность, которая обычно не является целым числом. В этом случае [20] для компактного подмножества $A \in R^k$ с box-размерностью d почти каждое гладкое отображение $F : R^k \to R^n$, где n — целое число, большее чем $2d$, является взаимно-однозначной иммерсией каждого компактного гладкого многообразия C, содержащего A.

Примечания

1. Объединение касательных пространств, построенных в каждой точке $x \in X$, т. е. $T_X = \bigcup T_x$ называют *касательным расслоением* X. Оно не является векторным пространством, так как невозможно ввести операцию сложения векторов из двух разных касательных пространств.

2. Дифференциал $dF(x)$ задается обычно матрицей Якоби. Поскольку $dF(x)$ — линейное отображение, иммерсия эквивалентна условию, что якобиан имеет полный ранг в касательном пространстве.

3. Множество A называется *компактным*, если оно ограничено и замкнуто. Ограниченность означает, что A содержится в шаре заданного радиуса, а замкнутость — что оно содержит все свои предельные точки.

4. Представим себе бесконечную нить, накрученную на тор: это инъективная иммерсия $R \to T^2$. Однако бесконечно удаленным точкам нити нельзя поставить в соответствие точки тора, следовательно инъективная иммерсия не является вложением без условия компактности.

5. Иными словами, если данное гладкое отображение есть вложение, его малое возмущение также является вложением. С другой стороны, любое гладкое отображение независимо от того, вложение это или нет, «лежит» произвольно близко от вложения.

6. Другая «физическая» интерпретация сводится к структуре фазового пространства: для идентификации динамической системы необходимо k координат, k импульсов и время.

7. Тор T^2 требует, согласно теореме Уитни, R^5 для вложения. Этот избыток размерностей гарантирует, что не только привычный «бублик»,

но и любые его диффеоморфные копии будут вложениями! Кроме того, условие $2k+1$ гарантирует, что любое гладкое отображение можно аппроксимировать вложением.

Путеводитель по литературе. Строгое изложение иммерсий и вложений содержится в [6–8]. Лучшее введение в эту область можно найти в книгах [10, 11] и сборнике [15], обзоре [16] и статье [9]. Наконец, фракталам и фрактальным аттракторам посвящена монография [23] и лекция [24].

Трансверсальность

> Помни, что в Искаженном Мире все правила ложны, в том числе и правило, перечисляющее исключения, в том числе и наше определение, подтверждающее правило.
>
> *Р. Шекли*
> *«Обмен разумов»*

В этом разделе речь пойдет о геометрической ипостаси типичности: о взаимном расположении геометрических тел в пространстве в рамках концепции «общего положения».

Две линии в R^3 в общем случае не пересекаются: в самом деле, достаточно слегка «пошевелить» две пересекающиеся прямые, чтобы достичь этого общего случая. Напротив, в R^2 такое «малое шевеление» не меняет ситуацию, так что пара прямых на плоскости, *как правило*, пересекается. Пара плоскостей в R^3 либо параллельна, либо пересекается вдоль линии — второй случай, очевидно, устойчив относительно «малых шевелений». Линия и плоскость имеют, *как правило*, общую точку — линия, параллельная плоскости не устойчива. Эти наблюдения легко обобщаются на R^n: пусть R^k и R^l — два подпространства в R^n. В общем случае они не пересекаются, если $(k+l) \leqslant n$; в противном случае они пересекаются так, что размерность их «общей части» равна $k+l-n$. На рис. 7 приведены примеры трансверсальных пересечений (вверху) и нетрансверсальных (внизу).

Важное замечание. Может случиться, что пересечение вообще отсутствует. Такая ситуация также считается трансверсальной [(1)]. Поскольку ло-

кальной версией многообразий являются касательные пространства, приведенные рассуждения легко обобщаются для взаимных конфигураций двух подмногообразий K и L в M. Говорят, что K и L пересекаются в M *трансверсально*, если для каждой $p \in K \bigcap L$

$$T_pK + T_pL = T_pM$$

и это записывают как $K \pitchfork L$. Иначе говоря, при трансверсальном пересечении сумма касательных пространств слагаемых равна полному касательному пространству многообразия [2].

Адаптируем это определение для отображений.

Пусть $Z \subset M$, $f : N \to M$ — гладкое отображение и $p \in N$ — точка. Говорят, что f — *трансверсально* Z *в* p ($f \pitchfork Z$), если либо $f(p) \notin Z$, либо $T_pf(T_pN) + T_{f(p)}Z = T_{f(p)}M$. Таким образом, условие трансверсальности необходимо для описания общей типичной ситуации. В локальном варианте его можно редуцировать в *принцип общего положения*.

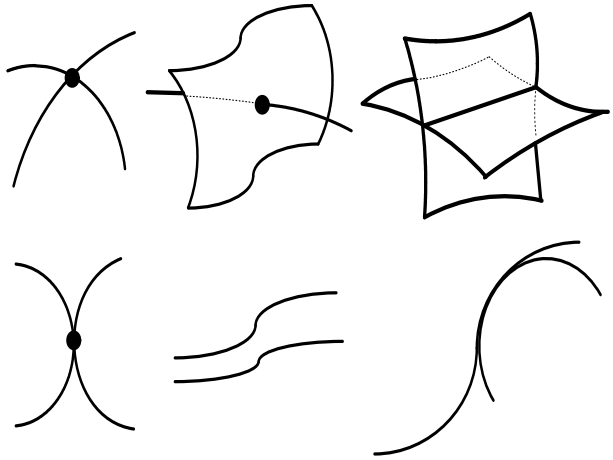

Рис. 7. Примеры взаимного расположения геометрических тел в R^3 (вверху) и R^2 (внизу): трансверсальными являются 2-й и 3-й пример в R^3

Рассмотрим теперь связь между трансверсальностью и условием теоремы Уитни. Пусть по-прежнему R^k и R^l — два подпространства в R^n, которые пересекаются трансверсально. Тогда верхняя часть рис. 7 учит нас, что

в случае трансверсальности размерности k, l, n удовлетворяют неравенству $(k + l - n) \geqslant 0$. Следовательно, при отсутствии трансверсального пересечения должно выполняться неравенство $k + l - n < 0$. Случай $k = l = d$ соответствует самопересечению или его отсутствию для одного геометрического тела, например, кривой. Тогда условие $n > 2d$ гарантирует нам, что самопересечение либо отсутствует, либо не является трансверсальным. В последнем случае оно легко ликвидируется малым шевелением. Самый простой способ удовлетворить этому неравенству — положить $n = 2d + 1$. Это и есть условие теоремы Уитни.

Примечания

1. Пусть $L \equiv R$, а $K \equiv R^2$, $M \equiv R^3$. Тогда в точке их трансверсального пересечения мы имеем тройку касательных векторов: два из них принадлежат плоскости, а третий — линии. Эта тройка — локальная модель R^3. Таким образом, трансверсальное пересечение моделирует (дает скелет) пространства, в котором происходит пересечение.

2. Здесь используется известный в математике и в жизни «принцип развесистой клюквы»: *развесистая клюква не обязательно является развесистой и не обязательно является клюквой*.

Путеводитель по литературе. Строгое изложение идей трансверсальности можно найти в учебниках [6–8, 10, 11]. Понятное для физиков введение содержится в [9, 16].

Эмбедология и теорема Такенса

> . . . Ничто так не ускользает от изображения в слове и в то же время ничто так настоятельно не требует передачи на суд людей, как некоторые вещи, существование которых недоказуемо, да и маловероятно.
>
> *Альберт Второй*
> *«Трактат о кристаллических духах», книга 1, раздел 28. Цит. по: Герман Гессе. «Игра в бисер».*

Напомним, что динамической системой в фазовом пространстве M называется однопараметрическая группа диффеоморфизмов $f^t(x) : M \to M$. Параметр группы t — это время, которое может принимать непрерывные или дискретные значения. В качестве M выступают обычные евклидовы пространства R^n или гладкие многообразия. Диффеоморфизмы f^t могут быть заданы в виде дискретных или непрерывных преобразований; групповое свойство означает:

$$f^{t_2}(f^{t_1}(x)) = f^{t_1+t_2}(x)\,.$$

При этом $f^0(x) = x$ — единица группы; если $x_1 = f^t(x_2)$, то $x_2 = f^{-t}(x_1)$.

Если динамическая система задана дифференциальным уравнением $\dot{x} = F(x)$, то в первом приближении

$$x(t+dt) = x + dx = x + Fdt = f^{dt}(x) = \\ = f^0(x) + \frac{df^0(x)}{dt}dt + \ldots = x + \frac{df^0(x)}{dt}dt\,.$$

Функцию $F(x) = \frac{df^t(x)}{dt}|_{t=0}$ называют *векторным полем*, которое задает касательную к траектории в каждой точке фазового пространства.

Траекторией или орбитой динамической системы называют последовательность точек

$$\ldots, f^{-3}(x), f^{-2}(x), f^{-1}(x), x, f(x), f^2(x), \ldots$$

Неподвижной точкой x_0 называется траектория, которая для всех t удовлетворяет условию: $f^t(x_0) = x_0$.

Периодической называется траектория, не являющаяся неподвижной точкой, и такая, что для некоторого P и произвольного t выполнено равенство $f^{t+P}(x) = f^t(x)$.

Точка p называется *ω–предельной точкой* для x, если $f^t(x) \to p$ при $t_i \to \infty$.

Пусть $A \in M$ и $f^t(A)$ — множество образов всех точек из A. Множество A *инвариантно* относительно f, если $f^t(A) = A$ для всех t.

Замкнутое инвариантное множество $A \subset M$ называется *притягивающим*, если для него существует окрестность U такая, что для всех $x \in U$, $f^t(x) \to A$ при $t \to \infty$. Инвариантное притягивающее множество, которое содержит все ω–предельные точки называют *аттрактором*.

Важное замечание. Если аттрактор достаточно велик и содержит только конечное число неподвижных точек и периодических орбит, то поведение фазовых кривых на аттракторе и рядом с ним является хаотическим [(1)]. Хаотический аттрактор ведет себя как гибкий, хотя не вполне управляемый, информационный процессор, обрабатывающий информацию о начальных данных. Фазовые траектории разбегаются в одних (неустойчивых) направлениях и сжимаются в других. Если сжатие преобладает (система диссипативна), то аттрактор копирует сам себя: сечение фазового потока приобретает самоподобную (странную) структуру канторова множества с дробной размерностью [(2)]. Информация порождается не только каскадом бифуркаций, приводящих к нарушению симметрии, но и последовательными итерациями, приводящими к все более тонкому разрешению геометрической структуры. Такая «аппаратурная реализация» может исполнять очень сложный функциональный репертуар, меняя поведение от относительно простого, квазипериодического, до стохастического. Наиболее важным свойством хаотического аттрактора является *существенная зависимость от начальных условий*. Говорят, что f^t обладает упомянутым свойством, если для точки $x \in M$ и некоторого числа $\delta > 0$ существует время $T > 0$ и такая точка y из окрестности x, что расстояние между образами $d(f^T(x), f^T(y)) > \delta$. Иными словами, погрешности начальных данных экспоненциально растут в фазовом потоке, так что начиная с некоторого момента времени, будущее состояние системы становится непредсказуемым.

Предположим теперь, что мы верим в существование динамической системы f^t, объясняющей наши наблюдения, но не имеем ее уравнений в явном виде. То, что нам доступно — это множество наблюдаемых функций $h(t) \equiv h(x(t)) : M \to R$, которые можно представить себе как нелинейные проекции фазовой траектории на вещественную ось. Обычно такими функциями являются экспериментальные временные ряды [(3)]. Мы будем предполагать, что они обладают непрерывностью [(4)] и может быть даже гладкостью, например, $h(t)$ может принадлежать классу $C^r(M, R)$. Во всяком случае, когда фазовая точка смещается под действием $f : x \to f(x)$, это движение должно индуцировать смещение в отсчетах временного ряда: $h(t) = h(x(t)) \to h(f(x(t)))$. Можно ли в этом случае восстановить динамику f^t в каком либо разумном смысле? Иными словами, какую информацию о неизвестной системе можно извлечь из наблюдений ее «тени», которую отбрасывают траектории? Или, в более общем смысле, можно ли восстановить модель непосредственно из экспериментальных данных? Впервые ответ на этот вопрос был дан в 1980 году в пионерской работе

группы *Паккарда и др.* [17]: *«Эвристическая идея метода реконструкции состоит в том, что для описания состояния трехмерной системы в любой момент времени достаточно измерения любых трех независимых величин, где термин "независимый" пока не определен формально, но будет определен операционно».* В качестве такой тройки в этой же работе было предложено использовать сам ряд $h(t)$ и две его производные. Полученные новые координаты $(h, \dot{h}, \ddot{h})$ позволяли восстановить фазовый портрет трехмерной модельной системы, заданной в явном виде. В качестве ряда было взято решение системы трех дифференциальных уравнений для одной из координат [(5)]. Несмотря на привлекательность самой идеи — реконструкция фазовой геометрии системы из временного ряда только одной координаты [(6)], математическая корректность самой процедуры вызывала сомнения. Действительно, согласно теореме Уитни, трехмерное многообразие невозможно даже погрузить в R^3 так, чтобы образ был компактным подмножеством с точностью до предположения о типичности [(7)]. Ситуация изменилась в 1981 году, когда *Такенс* [20] формализовал эвристику *Паккарда* в форме своей известной теоремы. Я приведу ее в строгой форме, а затем дам неформальные разъяснения.

Теорема Такенса. Пусть D^r — множество C^r-диффеоморфизмов $\{f^t\}$, $f^t : M \to M$, компактного d-мерного многообразия M и пусть $C^r(M, R)$ — множество наблюдаемых функций $h(t), h : M \to R$. Определим для $m \geqslant 2d + 1$ отображение запаздывающих координат $F_{f,h} : M \to R^m$:

$$F_{f,h}(\overrightarrow{x}) = (h(\overrightarrow{x}), h(f(\overrightarrow{x})), \ldots, h(f^{m-1}(\overrightarrow{x}))) .$$

Тогда множество (f^t, h), для которого $F_{f,h}$ является вложением, открыто и всюду плотно в $D^r(M) \times C^r(M, R)$ для $r \geqslant 1$.

Первая часть теоремы содержит наши *символы веры* [(8)], сводящиеся фактически к тому, что достаточно гладкие (т. е. класса $C^r(M, R)$) наблюдения — временной ряд $h(t)$ — продуцируются $C^r(M)$-гладкой динамической системой, которая имеет компактный d-мерный аттрактор.

Вторая часть теоремы учит нас, как, используя временной ряд $h(t)$, построить копию аттрактора в евклидовом пространстве R^m подходящей размерности $m = 2d + 1$ (см. рис. 8). Для этого следует сперва взять m отсчетов временного ряда, начиная с произвольного номера n:

$$h_n, h_{n+1}, \ldots, h_{n+m-1} .$$

Они рассматриваются как компоненты m-мерного вектора и образуют точку образа $z_n \in R^m$:

$$z_n = (h(f^n(\overrightarrow{x}_0)), h(f^{n+1}(\overrightarrow{x}_0)), \ldots, h(f^{n+m-1}(\overrightarrow{x}_0))) = \\ = (h(\overrightarrow{x}_n), h(f(\overrightarrow{x}_n)), \ldots, h(f^{m-1}(\overrightarrow{x}_n))) = F(\overrightarrow{x}_n)\,.$$

Следующая точка получается сдвигом m-мерного набора на один отсчет, а итеративное продолжение этой процедуры даст упорядоченную последовательность точек — копию истинной орбиты! Действительно, поскольку $F_{f,h}$ — гладкое и обратимое преобразование, можно определить отображение $\sigma = F \circ f \circ F^{-1}$ [9]. Так как точка z_n является образом F, применение σ в R^m дает:

$$\sigma(z_n) = F \circ f \circ F^{-1}(z_n) = F \circ f \circ F^{-1}(F(\overrightarrow{x}_n)) = \\ = F \circ f(\overrightarrow{x}_n) = F(\overrightarrow{x}_{n+1}) = z_{n+1}\,.$$

Заметим, что теорема справедлива для диффеоморфизмов f^t, имеющих лишь конечное число неподвижных точек и периодических орбит с периодом меньше m. В противном случае, для такой периодической точки нарушается даже условие иммерсии (погружения) $F_{f,h}$.

Теорема гарантирует, что копия будет *вложением*, т. е. сохранит свойство оригинала с точностью до диффеоморфизмов — непрерывных и дифференцируемых преобразований. Вспомним, что для (топологического) вложения можно использовать любую непрерывную функцию; сдвиг — самая простая функция из возможных. На рис. 9 приведен пример вложения в R^3 временного ряда чисел Вольфа (рис. 10).

Важное замечание. Именно это свойство позволяет оценивать многие необходимые динамические инварианты (размерность, ляпуновские показатели, энтропию) для неизвестной в явном виде системы по диффеоморфной копии ее аттрактора. Реконструкция находится в привычном R^m и вполне доступна для численных манипуляций! Более того, отображение $\sigma = F \circ f \circ F^{-1}$ [9] : $z_{n+1} = \sigma(z_n)$ лежит в основе современной схемы нелинейного многомерного прогноза. Конечно, z_{n+1} является вектором, однако уравнение можно переписать для проекции на первую координату:

$$h_{n+1} = \Phi(h_n, h_{n-1}, \ldots, h_{n-m+1})\,.$$

Таким образом, прогноз сводится к поиску наилучшей аппроксимации нелинейной, но непрерывной функции σ от m переменных. Отображение

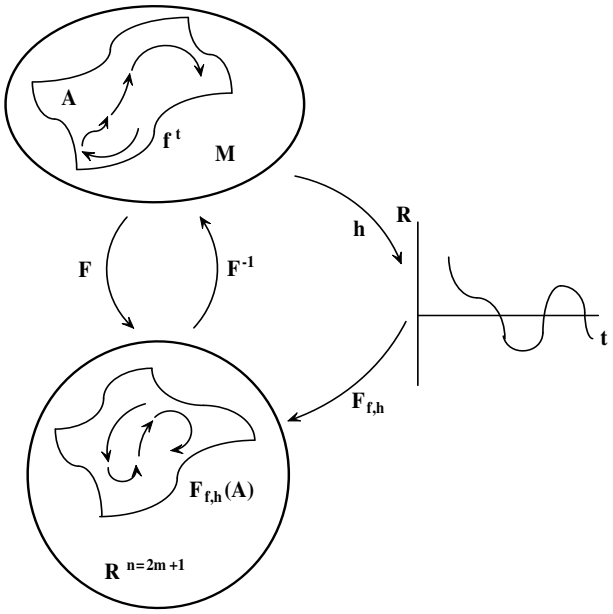

Рис. 8. Реконструкция копии аттрактора A из временного ряда $h(t)$ согласно теореме Такенса

F, которое реализует вложение, называют *отображением запаздывающих координат*, а пространство R^m называют *пространством вложения*. Хотя справедливость теоремы не зависит от величины сдвига (*лага*) между отсчетами — в приведенной выше формулировке он был выбран равным единице — на практике требуется выполнение условия «независимости» новых запаздывающих координат. В противном случае реконструкция «коллапсирует» к диагональной гиперплоскости в пространстве вложения [10]. Несмотря на простоту их получения, запаздывающие координаты Такенса в аналитическом смысле хуже варианта *Паккарда* с производными: они не позволяют получить простую форму уравнений движения [11].

Для практического применения теоремы Такенса нужна, прежде всего, оценка box-размерности аттрактора d. Для ее получения используют простую разновидность метода «проб и ошибок». Поскольку мы не знаем истинной размерности аттрактора, для оценки d воспользуемся его копией.

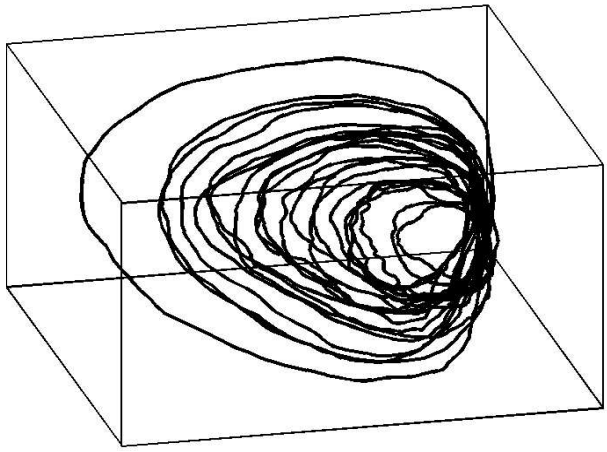

Рис. 9. Вложение временного ряда чисел Вольфа в R^3. Показан фазовый портрет, построенный в запаздывающих координатах; здесь $x, x + \text{лаг}, x + 3\text{лаг}$ играют роль x, y, z

.

А именно, начнем строить реконструкции аттрактора в R^d, для $d = 2, 3, \ldots$ Для каждой такой пробной копии будем вычислять некоторую меру, зависящую от размерности. Чаще всего для этих целей используют усредненное число пар ε-близких точек копии, подсчитанное в любой подходящей метрике. Мера меняется в зависимости от пробной размерности, но эта зависимость стабилизируется, как только мы достигаем «правильного» значения размерности. Используя затем масштабное поведение размерности от разрешения ε, легко оценить значение показателя этой зависимости, так называемую *корреляционную размерность* ν. Она и будет нижней оценкой размерности d аттрактора [12]. Практическое применение теоремы Такенса, т. е. реконструкция копии аттрактора из временных рядов с помощью топологических вложений, получило широкое распространение: весь комплекс необходимых для этого процедур, вместе с искусством получения численных оценок называют *эмбедологией*. Этот термин ввел *Тим Зауер и др.* в статье [16], синонимом является термин *«алгоритмическое моделирование»*, предложенное в [22].

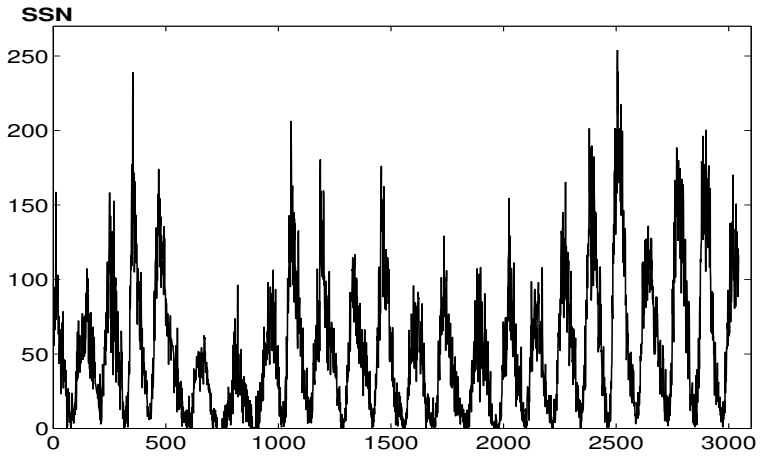

Рис. 10. Временной ряд чисел Вольфа: по вертикали — среднемесячные числа Вольфа (SSN — Sun Spot Numbers), по горизонтали — порядковые номера

Следует помнить, что корректность реконструкции методами эмбедологии существенно зависит от *Кредо идеального экспериментатора* [19] (см. с. 115). Самый болезненный для совести «эмбедолога» вопрос — это проблема существования аттрактора. Для того чтобы избежать в этом месте мучительных раздумий, обычно прибегают к нехитрой уловке, полагая, что . . . *все что происходит, если это действительно происходит, должно находиться на аттракторе. В противном случае, мы ничего бы и не видели!*

Примечания

1. Этот принцип не только не доказан, но даже и не формализован.
2. С физической точки зрения размерность множества — это число параметров, различающих точки этого множества. Для множества экзотической структуры это понятие может быть формализовано как размерность типичной плоскости, на которую множество можно взаимнооднозначно ортогонально спроектировать.

3. Часто используют прилагательное «наблюдаемая» как существительное: *наблюдаемая* $h(t)$. Более точно, ее называют *детерминированно порожденной* [18], если существует гладкая динамическая система $\dot{x} = F(x)$ в $M = R^n$, начальная точка $x_0 \in M$ и ограниченное решение $t \to x_0(t)$ вместе с функцией $h(t) : R^n \to R$ так, что $h(t_i) = h(x_0(i))$.

4. Обычно достаточно условия липшиц-непрерывности, т. е. $|h(t+\varepsilon) - h(t)| \leqslant C\varepsilon^H$, $H \in [0, 1]$.

5. Прием *Паккарда* сводится к следующему изменению координат [25]: для динамической системы

$$\frac{d\mathbf{x}}{dt} = \mathbf{F}(\mathbf{x})\,,$$

$\mathbf{x} = (x_1, x_2, x_3)$, определим вектор реконструкции $\mathbf{y}$ следующим образом:

$$\begin{aligned} y_1 &= x_1\,, \\ y_2 &= \dot{x_1} = \dot{y_1} = F_1\,, \\ y_3 &= \dot{y_2} = \ddot{x_1} = \dot{F_1} = \frac{\partial F_1}{\partial x_i}\frac{dx_i}{dt} = \mathbf{F}\nabla_1\,. \end{aligned}$$

В этом случае уравнения движения:

$$\begin{aligned} \dot{y_1} &= y_2\,, \\ \dot{y_2} &= y_3\,, \\ \dot{y_3} &= G(y_1, y_2, y_3) = (\mathbf{F}\nabla)^2 F_1 \end{aligned}$$

упрощаются, так как вместо трех отдельных функций от трех переменных мы имеем дело лишь с одной функцией G.

6. Статья [17] так и называлась: *«Геометрия из временных рядов»*.

7. Условие $n \geqslant 2d + 1$ можно редуцировать к $n = d$ только в случае *линейных* f^t и $h(t)$ [21]. Но это исключительный случай!

8. Условия теоремы Такенса резюмирует *Кредо идеального экспериментатора* [19], которое предполагает, что в фазовом пространстве M:

- существует гладкая динамическая система $f^t : M \to M$ с конечномерным аттрактором A;
- аттрактор имеет единственную инвариантную меру μ: она описывает частоту посещаемости различных частей аттрактора, т. е. время, которое проводит фазовая траектория в каждом фрагменте аттрактора;
- мера эргодична (частота посещаемости пропорциональна объему фрагмента) и инвариантна под действием f^t, т. е. $\mu(B) = \mu(f^t(B))$ для любого $B \subseteq A$;
- начальная точка фазовой траектории — типичная точка в смысле меры μ.

9. Это просто запись последовательных процедур, которые следует читать слева направо: сперва мы отображаем точку из R^m в фазовое пространство M с помощью F^{-1}, затем сдвигаем полученный образ в M с помощью отображения f^t и наконец, отображаем полученную фазовую точку с помощью F назад в R^m.

10. Это легко понять для вложения временного ряда в R^2: если $h_n \simeq h_{n+1}$ — пара координат (h_n, h_{n+1}) принадлежит диагонали!

11. В общем случае, запаздывающие координаты получают сдвигом на некоторый лаг τ. Следовательно, для трехмерной системы имеем:

$$\begin{aligned} y_1(t) &= x_1(t)\,, \\ y_2(t) &= x_1(t-\tau)\,, \\ y_3(t) &= x_1(t-2\tau)\,. \end{aligned}$$

Тогда уравнения движения имеют вид: $\dot{y}_i = H_i(\mathbf{y})$, для $i = 1, 2, 3$, но функции H_i, в отличие от дифференциальных координат Паккарда нельзя явно выразить через первоначальные функции $F_i(\mathbf{x})$ [25].

12. Пусть μ — борелева вероятностная мера с ограниченным носителем в R^n с метрикой $|*|$. Определим корреляционную функцию $C_d(\varepsilon)$, как взвешенную с помощью μ долю пар векторов $(x, y) \in R^d \subseteq R^n$, таких что $|x - y| \leqslant \varepsilon$. Тогда, *корреляционная размерность* ν меры μ определяется выражением:

$$\nu(\mu) = \sup_s \Big\{ \int\int |x-y|^{-s} d\mu(x) d\mu(y) = \int_0^\infty \varepsilon^{-s} dC_d(\varepsilon) < \infty \Big\}.$$

Масштабное поведение размерности следует из следующего построения. Разобьем компактную область в R^d, содержащую N точек, на $N(\varepsilon)$ непустых кубов, и пусть μ_i — число точек в i-ом кубе. Тогда с точностью до $O(1)$:

$$C_d(\varepsilon) \approx N^{-2} \sum_i \mu_i^2 = N(\varepsilon) N^{-2} \langle \mu^2 \rangle.$$

Здесь $\langle * \rangle$ — среднее по всем непустым боксам. Используя неравенство Шварца, получаем:

$$N(\varepsilon) N^{-2} \langle \mu^2 \rangle \approx N^{-2} N(\varepsilon) \langle \mu \rangle^2 = \\ = N^{-2} N^{-1}(\varepsilon) \Big[\sum \mu_i \Big]^2 = 1/N(\varepsilon) = \varepsilon^d .$$

Таким образом, в пределах малых ε:

$$C_d(\varepsilon) \propto \varepsilon^\nu; \quad \nu \equiv d = \lim \log C_d(\varepsilon) / \log \varepsilon.$$

Путеводитель по литературе. В основе эмбедологии лежат статьи [17–18]; историческая предтеча алгоритмического моделирования изложена в обзорной статье [22]. Понятное и строгое изложение теоремы Такенса дано в работе [21]. В [19] обсуждаются *Кредо идеального экспериментатора* (см. с. 115), впервые введенное в [18]. Фракталы и их связь с динамическими системами изложены в [23] и лекции [24]. Современное изложение теории гладких эргодических динамических систем можно найти в [25–33]. Разнообразные определения аттракторов приводятся в статье *Милнора* [34]. В обзорах [35–36] изложена техника алгоритмического моделирования, главные первоисточники содержатся в сборнике репринтов оригинальных работ [15]. Существует довольно много компьютерных программ для алгоритмического моделирования, наиболее полезными из них, по моему мнению, являются пакеты *Tisean* (Linux, Dos) [37] и *TSTOOL* (Windows MatLab) [38].

Прогноз как аппроксимация: сети, но не нейронные сети

> Если бы это было так, это бы еще ничего, а если бы ничего, оно бы так и было, но так как это не так, так оно и не этак! Такова логика вещей!
>
> *Л. Кэрролл*
> *«Алиса в Зазеркалье»*

Напомним прежде всего схему обычного авторегрессионного [1] линейного прогноза временных рядов. Обозначим наблюдаемый временной ряд (время дискретно) как $x(n)$ $(n = 1, 2, \ldots, N)$, а его предсказание — как $\hat{x}(n)$. Предположим, что наша последняя известная точка ряда $x(n-1)$. Тогда линейное авторегрессионное предсказание на один шаг вперед определяется выражением:

$$\hat{x}(n) = a_1 x(n-1) + \ldots + a_p x(n-p)\,,$$

где a_i $(i = 1, 2, \ldots, p)$ — весовые коэффициенты, а p — порядок авторегрессии [2]. Веса находятся из условия, чтобы ошибки предсказания:

$$w(n) = x(n) - \hat{x}(n) = x(n) - a_1 x(n-1) - \ldots - a_p x(n-p)$$

для $n = 1, \ldots, N-p$ были случайными некоррелированными величинами. В общем случае AR-процесс p-го порядка можно записать в виде:

$$\sum_{j=0}^{p} a_j x(n-j) = w(n)\,,$$

где $a_0 = 1, a_1, a_2, \ldots, a_p < 0$. Введем оператор сдвига на один отсчет: B ($Bx(n) = x(n-1)$). Тогда уравнение для ошибки предсказания принимает вид:

$$(1 - a_1 B - \ldots - a_p B^p)x(n) = Q_p(B)x(n) = w(n)$$

и его решением будет выражение, в котором заключена суть AR-модели: $x(n) = Q_p^{-1}(B)w(n)$. Мы получили просто *Черный Ящик*, который преобразует белый шум в отсчет временного ряда!

Для дальнейшего важно, что AR-прогноз основан на *линейной* комбинации прошлых значений. Наиболее очевидное обобщение заключается в

том, чтобы заменить линейную комбинацию на нелинейную функцию. Но именно этот вариант и предлагает эмбедология! Действительно, согласно теореме Такенса в пространстве вложения R^m существует универсальная динамическая модель:

$$x_{i+k} = \Phi(\mathbf{z}_i)\,,$$

где k — искомая координата вектора $\mathbf{z}_{i+1}$. Таким образом, задача прогноза в эмбедологии приобретает следующую форму: для данного временного ряда и полученной для него реконструкции в R^m, известны $(N-m+1)$ значений [9] векторов $\mathbf{z}_i = x_i, x_{i+1}, x_{i+2}, \ldots, x_{i+(m-1)}$.

Следовательно, для каждого индекса $i = 1, 2, .., N-m+1$ мы имеем значения функции: $x_{i+m} = \Phi(\mathbf{z}_i)$. Так, например, если $m = 3$, $N = 6$, мы получим три «примера»:

$$\begin{aligned} x_4 &= \Phi(x_3, x_2, x_1)\,,\\ x_5 &= \Phi(x_4, x_3, x_2)\,,\\ x_6 &= \Phi(x_5, x_4, x_3)\,. \end{aligned}$$

Следовательно, для прогноза необходимо найти оценки для x_{6+i}, начиная с $i = 1$, т. е. $x_7 = \Phi(x_6, x_5, x_4)$.

К сожалению, мы имеем лишь самую общую информацию о функции $\Phi(\mathbf{z}_i)$: эта функция непрерывна и, возможно, дифференцируема. Хуже того, она обычно определена не на всем пространстве векторов $\mathbf{z}_i$, а только на поверхности некоторой неизвестной размерности, меньше чем m. В этих условиях, задача аппроксимации Ф решается только на уровне технической строгости. Удивительно, что в ряде случаев результаты прогноза тем не менее вполне удовлетворительны.

Принято делить нелинейные методы прогноза на *локальные* и *глобальные*. В *локальных* методах, функция Ф аппроксимируется в локальной окрестности каждой фазовой точки. Поскольку сшивать эти отдельные приближения не требуется, локальный метод называют иногда *непараметрической регрессией*. Идея состоит в следующем. Для каждой точки $\mathbf{z}_i^0$ находят k ее ближайших соседей. После этого, поведение функции Ф аппроксимируется полиномом степени n в окрестности точки $\mathbf{z}_i^0$. Коэффициенты полинома находят методом наименьших квадратов из условия:

[9] Здесь, ради простоты, мы полагаем лаг $\tau = 1$; если это не так, число таких векторов $(N-(m+1)\tau)$.

$\sum_{i=1}^{k} |P_n(\mathbf{z}_i^0) - \Phi|^2 \to \min$ [3]. В *глобальных* методах динамика аппроксимируется сразу во всем **z**-пространстве. При использовании полиномов, минимизируется функционал *полной* ошибки:

$$\sum_{i=1}^{N} \left|P_n(\mathbf{z}_i^0) - \Phi\right|^2 \to \min .$$

На практике чаще используют глобальные методы с *локальными* свойствами — радиальные базисные функции и нейросетевые технологии. Рассмотрим теперь общие принципы аппроксимации непрерывной многомерной функции f**(x)** с помощью некоторой аппроксимирующей функции $F(\mathbf{w}, \mathbf{x})$ с фиксированным числом параметров **w**; **x**, **w** — вещественные векторы вида $\mathbf{x} = x_1, x_2, \ldots, x_n$ и $\mathbf{w} = w_1, w_2, \ldots, w_m$. При выборе F возникает задача нахождения параметров **w**, которые обеспечивают наилучшую возможную аппроксимацию функции f на основе конечного множества известных «примеров» [4]. Качество аппроксимации можно измерить с помощью некоторой функции ρ, измеряющей расстояние $\rho[f(\mathbf{x}), F$**(w,x)**$]$ между аппроксимацией F**(x,w)** и f**(**x**)**. Тогда, если $f(\mathbf{x})$ — непрерывная функция, определенная на множестве X, F**(w,x)** — аппроксимирующая функция, непрерывно зависящая от $\mathbf{w} \in P$ и X, задача аппроксимации заключается в поиске параметров $\mathbf{w}^*$, таких что

$$\rho[F(\mathbf{w}^*, \mathbf{x}), f(\mathbf{X})] < \rho[F(\mathbf{w}, \mathbf{x}), f(\mathbf{x}))]$$

для всех **w** из P.

Как правило, функции F выбираются из некоторого класса $A = \{F_i\}$, который является подмножеством некоторого общего множества функций $\Re \supset A$, содержащего и функцию f. Поэтому, полезно определить расстояние между функцией f и подмножеством A, как $\rho(f, A) = \inf_{F_i \in A} \|f - F_i\|$. Если точная нижняя граница inf достигается для некоторой $F_* \in A$, то именно эта функция и будет *наилучшей аппроксимацией* для f [5]. Теперь задача аппроксимации формулируется так: для $f \in \Re$ и $A \in \Re$ найти наилучшую аппроксимацию f функциями из A. Для решения задачи необходимо конечно, чтобы функции F_i, т. е. подмножество A было *всюду плотным* в $\Re$. В этом случае, какова бы ни была $f \in \Re$, в ее окрестности всегда найдутся подходящие «строительные блоки», из которых мы и построим ее лучшую аппроксимацию [6]. К счастью, существуют теоремы, которые помогают нам найти такие A. Например, из курса анализа (*теорема Вейерштрасса*) мы знаем, что любую непрерывную функцию, заданную в

ограниченном интервале, можно сколь угодно точно аппроксимировать полиномами. Следовательно, множество полиномов (т. е. подмножество A) всюду плотно в пространстве $\Re \equiv C[X]$ непрерывных функций. Самым известным обобщением этого утверждения является *теорема Стоуна*.

Теорема Стоуна. Пусть X компактное метрическое пространство, $C[X]$ — множество непрерывных функций, определенных на X и A подалгебра [(7)] в $C[X]$ со следующими свойствами:

- функция $f(x) = 1$ принадлежит A;
- для любых двух точек $x \neq y$ из X существует функция $f \in A$, такая что $f(x) \neq f(y)$.

Тогда A плотно в $C[X]$.

Теорема Стоуна значительно расширяет класс доступных строительных блоков: действительно, достаточно взять компактное подмножество функций, различающих точки и образующих подалгебру относительно «школьных» алгебраических операций и мы получим возможность аппроксимировать практически любую непрерывную функцию! Такими функциями могут быть, например, тригонометрические многочлены, радиальные базисные функции и др.

Почти любая схема аппроксимации может быть «отображена» на некоторый тип «сети» или графа, который можно условно считать «нейронной сетью». Такие сети являются просто графическим обозначением широкого класса алгоритмов. Чтобы продемонстрировать, как задача аппроксимации выглядит в «сетевой» формулировке, рассмотрим несколько вариантов для возможного выбора $F(\mathbf{w}, \mathbf{x})$:

- линейный случай: $F(\mathbf{w}, \mathbf{x}) = \mathbf{w} \cdot \mathbf{x}$, где $\mathbf{w}$ — матрица размера $m \times n$, $\mathbf{x}$ — n-мерный вектор; этот вариант соответствует сети без скрытых слоев;
- классическая схема аппроксимации с базисными функциями $\varphi_i(\mathbf{x})$, т. е. $F(\mathbf{w}, \mathbf{x}) = \mathbf{w} \cdot \varphi_i(\mathbf{x})$. Этот вариант соответствует сети с одним скрытым слоем; разложение по ортогональным полиномам или радиальным базисным функциям;
- схема со вложенными сигмоидами, которую мы рассмотрим в следующем разделе.

Примечания

1. *Регрессия* — это взгляд назад, на прошлое какого-то временного ряда, *авторегрессия* — это собственное прошлое. Идеи авторегрессионной (*AR*)-модели восходят к пионерской работе английского статистика *Дж. Юла* (1927 г.), посвященной исследованию вариаций солнечных пятен [39–40]. Предтечей *AR*-моделей была модель скользящего среднего (*Moving average — MA*), предложенная *Е. Слуцким*, в которой случайный процесс X_n моделировался как линейная комбинация белого шума R_n: $X_n = \sum_i C_i R_{n-i}$. *AR*-модель выражает будущие значение через линейные комбинации своего прошлого: $X_n = R_n + \sum_i B_i R_{n-i}$, где B_i — постоянные коэффициенты. Комбинация *AR*- и *MA*-моделей называется *ARMA*-моделью [40].

2. Для нахождения порядка авторегрессии p используют различные подходы. Чаще всего используются критерии *Акаике* [40]. Согласно первому из них — *Окончательной ошибке предсказания (ООП)*, p выбирается так, чтобы средняя дисперсия ошибки на каждом шаге предсказания была минимальна. Второй — *Информационный критерий Акаике (ИКА)* оценивает порядок модели посредством минимизации некоторой информационной функции. Если AR-процесс имеет гауссовские статистики, то ИКА$[p] = Nln(\rho_p) + 2p$, где ρ_p-оценка дисперсии белого шума, которая используется в качестве ошибки линейного предсказания. Следует упомянуть еще критерий *Риссанена — ДМО*, который основан на минимизации *длины минимального описания модели*: ДМО$[\,p\,] = N\ln(\rho_p) + p\,\ln N$.

3. Поскольку число соседей k должно быть больше числа коэффициентов полинома (это число в R^m растет как m^n), обычно используют полиномы степени $n \leqslant 2$. Дело еще и в том, что при чрезмерном увеличении числа ближайших соседей в их число попадают фазовые точки с далеких соседних траекторий реконструкции; динамический смысл локальной окрестности поэтому теряется.

4. Вообще говоря, следует различать три разные проблемы [45]:
   - задача выбора аппроксимации (или задача представления): какие классы функций $f(\mathbf{x})$ могут быть эффективно аппроксимированы и какими аппроксимирующими функциями $F(\mathbf{w}, \mathbf{x})$?
   - проблема выбора алгоритма поиска оптимальных значений параметров $\mathbf{w}$ для данного F;

- проблема эффективной реализации алгоритма в параллельном, по-возможности аналоговом аппаратном средстве.

5. Множество A называют *множеством существования*, если для каждой $f \in \Re$ существует *по меньшей мере* одна наилучшая аппроксимация. Наконец, множество A называют *чебышёвским*, если существует *только одна* такая аппроксимация [46].

6. Для полного комфорта крайне желательно, чтобы множество A было *чебышёвским*, тогда лучшая аппроксимация будет единственной. Однако, на практике приходиться ограничить себя более скромным требованием, A — просто *множество существования*. Соответствующие теоремы, говорят нам, что для этого A должно быть компактным [46].

7. *Алгебра* — это множество элементов E вместе со скалярным полем $\mathcal{F}$, замкнутое относительно операций сложения $(+)$ и умножения $(\times)$ между своими элементами, а также умножения $(\cdot)$ на скаляр из $\mathcal{F}$. Множество E вместе с $\mathcal{F}$, $(+)$ и $(\cdot)$ образует линейное пространство. Кроме того, для любых f, g, h из E и $a \in \mathcal{F}$,

$$f \times g \in E\,; \quad f + g \in E\,; \quad a \cdot f \in E\,;$$
$$f \times (g + h) = f \times g + f \times h\,; \quad f \times (g \times h) = (f \times g) \times h\,;$$
$$a \cdot (f \times g) = (a \cdot f) \times g = f \times (a \cdot g)\,.$$

   Подалгебра — это подмножество $S \subset E$, которое является линейным подпространством в E и замкнуто относительно операции умножения $(\times)$: если $f,\ g \in S$, то $f \times g \in S$.

Путеводитель по литературе. Исходные идеи AR-моделей великолепно описаны в учебнике [39]. В монографии [40] дано подробное описание техники этих моделей. Статья [41] чрезвычайно полезна в качестве введения в нелинейный многомерный локальный и глобальный прогноз. Пионерскую работу [42] *Фармера* и *Сидоровича* хорошо дополняют детали, приведенные в учебнике [43]. Некоторые интересные вещи содержит обзорная статья [44]. Замечательный обзор [45] и препринт [46], которым я и следовал, посвящен методам аппроксимации в применении к нейронным сетям.

Нейропрогноз: обучение отображению

> Она (Программа) способна не только *предвидеть, что случится*, если **Что-То** случится, но, кроме того точно предсказывает, **Что** случится, если **То** ни капельки не случится, то есть вовсе не произойдет.
>
> *Станислав Лем*
> *«Экстелопедия Вестранда»*

Вернемся вновь к основному соотношению для прогноза: $x_{i+m} = \Phi(\mathbf{z}_i)$. В правой части стоит $\Phi(\mathbf{z}_i)$ — функция многих переменных. Нельзя ли представить эту функцию как некоторую, возможно нелинейную, комбинацию более простых функций, содержащих, например, не более одной переменной? Простой пример показывает, что такие попытки не выглядят безнадежными. Рассмотрим функцию двух переменных $f(x, y) = xy$. Пусть $g(\star) = \exp(\star)$ и $h(\star) = \log(\star)$ — две вспомогательных функции. Тогда соотношение

$$f(x, y) = xy = \exp(\log(x) + \log(y)) = g(h(x) + h(y))$$

показывает, что функцию двух переменных можно выразить через *суперпозицию* функций только *одной* переменной. Вопрос о том, существуют ли вообще *истинные* функции многих переменных, составлял 13-ю проблему *Гильберта*: *можно ли произвольную непрерывную функцию n переменных получить с помощью операций сложения, умножения и суперпозиции непрерывных функций двух переменных?*(1) Ответ найденный *А. Н. Колмогоровым* был положительным:

Теорема Колмогорова. Существуют фиксированные возрастающие непрерывные функции $h_{pq}(x)$, определенные на $I = [0, 1]$ такие, что любая непрерывная функция f на $I^n = I \times I \times \ldots I$ может быть записана в форме [10]

$$f(x_1, x_2, \ldots, x_n) = \sum_{q=1}^{2n+1} g_q \Big[\sum_{p=1}^{n} h_{qp}(x_p) \Big] ,$$

[10] Заметим,что здесь речь идет о *точном* представлении функции, а не о ее *аппроксимации*.

где g_q — должным образом выбранные непрерывные функции одной переменной.

Эту теорему удалось улучшить. В частности, было показано, что функции g_q можно заменить на одну универсальную непрерывную функцию g, которую можно представить в виде абсолютно сходящегося ряда Фурье, а функции h_{pq} можно «расщепить» на произведения $l_p h_q$, где $(l_p, p = 1, 2, \ldots, n)$ — рациональные независимые числа, а h_q, $(q = 1, 2, \ldots, 2n + 1)$ — непрерывные неубывающие функции на $I = [0, 1]$. Тогда, представление любой непрерывной функции $f \in C[I^n]$ можно переписать в виде:

$$f(x_1, x_2, \ldots, x_n) = \sum_{q=1}^{2n+1} g\Big[\sum_{p=1}^{n} l_p h_q(x_p)\Big] .$$

Ради наглядности, запишем это представление для случая функции двух переменных:

$$f(x, y) = g(l_1 h_1(x) + l_1 h_2(x) + l_1 h_3(x) + l_1 h_4(x) + l_1 h_5(x) + \\ + l_2 h_1(y) + l_2 h_2(y) + l_2 h_3(y) + l_2 h_4(y) + l_2 h_5(y)) .$$

На рис. 11 приведена сетевая форма этого примера. Она похожа на многослойную нейронную сеть, которую, однако, почти невозможно реализовать на практике. Число «нейронов» растет здесь с увеличение числа переменных, но дело не в этом. Основная проблема заключается в том, что функции h_q в формуле Колмогорова *непрерывны*, но не являются *гладкими* [(2)]. Поскольку длина машинного кода функции обратно пропорциональна ее гладкости, мы получим очень плохую точность реализации колмогоровского представления.

Кроме того, *хорошее* сетевое представление должно иметь фиксированные элементы и изменяемые параметры. Сеть *Колмогорова* на рис. 11 не является сетью такого типа: форма g зависит от функции f, которую необходимо представить (функции h_q не зависят от f).

С другой стороны, давно известна не точная, а аппроксимационная схема представления функции многих переменных в форме суперпозиции функций от одной переменной:

$$F(\mathbf{w},\mathbf{x}) = \sigma(\sum_n w_n \sigma(\sum_i v_i \sigma(\ldots \sigma(\sum_j u_j x_j) \ldots)))) ,$$

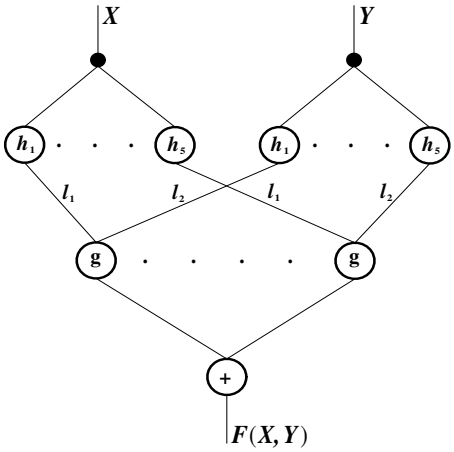

Рис. 11. Сеть, соответствующая *колмогоровскому* представлению функции двух переменных

которую называют обычно *схемой с обратным распространением*. Она соответствует многослойной нейронной сети с элементами, которые суммируют свои входы с весами $w, v, u, \ldots$ и затем выполняют *сигмоидальное* преобразование этой суммы [3]. Такая форма (схема вложенных нелинейных функций) необычна с точки зрения классической теории аппроксимации непрерывных функций [4]. Ее истоки лежат в том факте, что с помощью соотношения:

$$F(\mathbf{w},\mathbf{x}) = \sigma(\sum_n w_n \sigma(\sum_j u_j x_j))\,,$$

где σ — *линейная* пороговая функция, можно выразить любую *булеву* функцию [5]. В теории нейронных сетей класс функций:

$$F(\mathbf{w},\mathbf{x}) = \sum_{i=1}^{m} c_i \sigma(\mathbf{x}\cdot\mathbf{w} + \theta_i)\,,$$

которые соответствуют одному скрытому слою, принято рассматривать как аппроксимирующие функции для функций $f \in C[U \subset R^n]$. Действительно, функции σ образуют плотное множество, но к сожалению оно не является *множеством существования* [6], и следовательно нельзя гарантировать

единственности лучшей аппроксимации. Впрочем, это не так страшно для практических приложений.

Любопытно, что аппроксимационные свойства нейросети определяются не конкретным выбором функции активации, будь то сигмоида или что-то другое, а свойством нелинейности этой функции. *А. Н. Горбань* [48–50] доказал замечательную теорему, обобщающую теорему Стоуна на нейронные сети. Оказывается, если подалгебра E в теореме Стоуна замкнута относительно нелинейной операции, она остается всюду плотной в $C[X]$. Это эквивалентно утверждению об универсальных аппроксимационных свойствах любой нелинейности [49]: *с помощью линейных операций и каскадного соединения можно из произвольных нелинейных элементов получить требуемый результат с любой наперед заданной точностью!*

Нейросетевая схема аппроксимации функции многих переменных идеально подходит для реализации глобального [(7)] нелинейного прогноза. Сеть *учится отображению запаздывающих координат* $\Phi(\mathbf{z}_i) \to x_{i+m}$ на задачнике, составленном из прошлых значений временного ряда. Рассмотрим, в качестве примера, форму такого задачника. Пусть $N = 10, m = 4$ и лаг $\tau = 2$. Тогда таблица для обучения выглядит следующим образом:

$$\begin{array}{cccc|c} x_1 & x_3 & x_5 & x_7 & x_9 \\ x_2 & x_4 & x_6 & x_8 & x_{10} \\ x_3 & x_5 & x_7 & x_9 & \hat{x}_{11} \\ x_4 & x_6 & x_8 & x_{10} & \hat{x}_{12} \end{array}$$

В правом крайнем столбце стоят *ответы*. Поскольку лаг равен двум, одношаговый прогноз позволяет получить сразу два предсказанных значения: $\hat{x}_{11}$ и $\hat{x}_{12}$.

На рис. 12 приведен пример нейропрогноза ежемесячных значений временного ряда чисел Вольфа. Прогноз начинается с цикла № 20 и заканчивается серединой цикла № 24. Каждый цикл прогнозировался отдельно. Фактически, до середины текущего 23-го цикла делался эпигноз [11]. Параметры вложения $m = 7, \tau = 129$. Использовалась коррекция, описанная в Примечании 7.

[11] Здесь речь идет о том, что в прогнозе для тестирования используют уже известные значения ряда. Т. е. вы делаете вид, что их не знаете, и предсказываете, чтобы проверить метод. Мотивировка простая: для проверки корректности прогноза нескольких 11-летних циклов вперед можно и не дожить. Вот для такого варианта прогноза используют термин *эпигноз*. Иногда приходиться делать прогноз в прошлое, продолжая какие-то ряды за пределы интервала времени, когда они действительно есть. Такой прогноз называют *палегноз*.

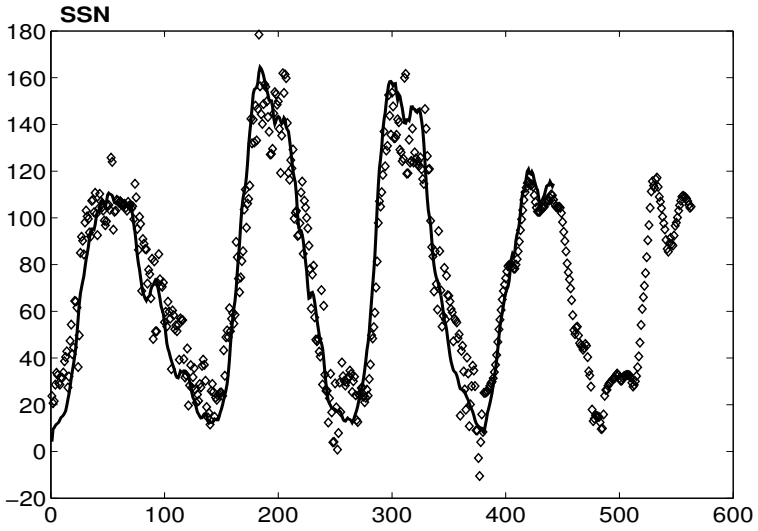

Рис. 12. Нейропрогноз временного ряда ежемесячных чисел Вольфа (SNN) с цикла № 20. Сплошная линия — реальные значения, ромбики — прогноз.

Важное замечание. Размерность вложения не слишком влияет на качество нейропрогноза: теорема Такенса требует лишь выполнения условия $m > 2d$. Правильный выбор лага τ очень существенен: большое τ позволяет получить долгосрочный прогноз «за один шаг» — на рис. 12 прогноз был сделан сразу на 129 точек. Это значительно уменьшает ошибку, которая растет с числом итераций в «одношаговом» варианте. Обычно выбор лага связан с характерным масштабом времени динамической системы, например с ее периодом. Проблемы возникают, когда в системе два или более таких масштабов. Лаг, связанный с одним из них, не позволяет получить оптимальное вложение, отслеживающее динамику на неучтенных масштабах. Выход был предложен *A. Mees* и др. [53] в форме так называемого *неоднородного вложения*. Идея заключается в замене запаздывающих векторов однородного вложения:

$$\mathbf{z}(t) = (x(t), x(t-\tau), \ldots, x(t-(m-1)\tau) \in R^m$$

на векторы:

$$\hat{\mathbf{z}}(t) = (x(t), x(t-l_1), x(t-l_2), \ldots, x(t-l_{m-1})\,,$$

где лаги $(l_1, \ldots, l_{m-1})$ — различные числа. Любая схема прогноза основана на минимизации функционала ошибки предсказания. Его нельзя использовать дважды: для нахождения модели (аппроксимирующей функции) и параметров вложения. Для получения второго, дополнительного функционала авторы [53] предложили использовать *принцип минимальной длины описания*, формализующий средневековое правило *Бритвы Оккама* [(8)]. Идея заключается в следующем. Предположим, что вам надо передать кому-либо экспериментальный временной ряд, измеренный с некоторой точностью. Вы можете послать исходные данные и в этом случае длина описания данных будет равна количеству бит, необходимых для бинарного кодирования данных. Однако можно поступить иначе: построить динамическую модель для данных, позволяющую прогнозировать ряд. Если вы договорились о классе модели, то вам достаточно передать ее параметры, начальное значение и допустимую ошибку предсказания. Тогда, полная длина описания будет состоять из описания самой модели (или данных, кодированных с помощью модели), начального значения и ошибки. Чем сложнее модель, тем длиннее ее описание, но тем короче оставшаяся часть сообщения. Необходимый нам функционал содержит как раз эти два слагаемых:

$$\begin{gathered}(\textit{Длина описания}) \approx (\textit{число данных})\times \\ \times \log(\textit{ошибка предсказания})+ \\ +(\textit{штраф за число и точность параметров модели})\,.\end{gathered}$$

Когда число параметров модели увеличивается, ошибка предсказания уменьшается: согласно *МДО*-принципу, оптимальная модель доставляет минимум этому функционалу в заданном классе моделей. Именно этот функционал и позволяет вычислить набор лагов для неоднородного вложения [(9)]. Наши эксперименты показали, что такое вложение в большинстве случаев существенно улучшает нейропрогноз.

Примечания

1. Если отказаться от условия непрерывности, получится тривиальный результат, связанный в сущности с многомерным обобщением известного примера *Кантора*, который показал, как построить отображение единичного квадрата в единичный отрезок: эти два множества *равномощны*.

2. Гладкость функции измеряется ее *гельдеровским показателем*: говорят, что функция $f(x)$ имеет локальный *показатель Гельдера* $\alpha \in [0,1]$ в точке x, если для достаточно малого числа $h > 0$ существует такая константа C, что $|f(x+h) - f(x)| \leqslant Ch^{\alpha}$. Значению $\alpha = 1$ соответствует существование одной производной. Пусть функция $f(x) \in C^r$ задана в единичном n-мерном кубе. Если все ее частные производные порядка r имеют показатель Гельдера α, то *сложность* функции можно измерить величиной [45] $\chi^{-1} = n/(r+\alpha)$. Таким образом, сложность функции прямо пропорциональна числу переменных и обратно пропорциональна суммарной гладкости функции и ее производных. Известно, что *не все* функции с показателем χ_0 могут быть представлены суперпозицией функций с характеристикой $\chi > \chi_0$. В частности, существуют C^r-функции n переменных, которые нельзя представить с помощью суперпозиции C^r-функций от *менее* чем n переменных. Таким образом, теорема Колмогорова — просто патологический случай!

3. Функция активации σ называется *сигмоидальной*, так как форма тех функций, которые выбираются на практике, обычно напоминают по форме деформированную букву S. Примером может служить широко распространенная *логистическая функция* $y = (1+\exp(-x))^{-1}$. Она имеет два преимущества. Во-первых, ее производная dy/dx очень просто выражается через y, что существенно упрощает численную минимизацию ошибки при коррекции весов. Во-вторых, она естественным образом применима в задачах *байесовской классификации* [57]. Формально говоря, сигмоидальными называют функции, удовлетворяющие двум условиям: $\lim_{t\to\infty}\sigma(t) = 1$ и $\lim_{t\to-\infty}\sigma(t) = 0$.

4. Недавно появилась работа [56], в которой предлагается нейронная сеть, согласованная с традициями теории аппроксимации.

5. Иными словами, отображение $S : \{0,1\}^N \to \{0,1\}$ может быть записано как *дизъюнкция конъюнкций*, что в переводе на язык пороговых элементов и есть приведенное в тексте выражение.

6. Существенное множество должно быть замкнутым. Этому условию удовлетворяют радиальные базисные функции, но не сигмоидные [46].

7. При этом полагают, что функция Ф *универсальна*, т. е. справедлива для всей истории временного ряда. Таким образом, предполагается *стационарность* ряда. Однако ряды, с которыми приходится иметь дело, обычно нарушают это условие. В этом случае полагают, что отображение $\Phi(\mathbf{z}_i, \mathbf{c})$ зависит от вектора *скрытых* или *бифуркационных* параметров $\mathbf{c}$: их изменение и управляет перестройкой динамики. Для того чтобы учесть влияние $\mathbf{c}$, разбивают ряд на фрагменты с приблизительно одинаковым поведением, полагая, что на каждом таком участке значение параметра фиксировано. После эпигноза каждого фрагмента полученную ошибку используют для настройки *корректора*, исправляющего расхождение. Его используют затем для коррекции истинного долгосрочного прогноза [52].

8. *Уильям Оккам* (*William Occam*, 1285–1349) английский монах-францисканец, философ, теолог-логик, основатель одной из форм *номинализма*. Логику, еще мальчиком изучал во францисканском монастыре, теологию — в Оксфордском университете. Оккам настаивал на строго рациональных оценках, различии между необходимым и случайным, очевидностью и степенью вероятности. Принцип, известный сейчас под наименованием *бритвы Оккама*, является вариантом средневекового правила «не следует допускать множественность без необходимости». Оккам использовал его для исключения многих сущностей, которые использовались для объяснения реальности схоластами. В оригинале принцип звучит так: "*Pluralitas non est ponenda sine necessitate*", т. е. *«Не следует умножать сущности без необходимости»*.

9. Кроме того, *МДО*-принцип можно использовать еще для нахождения параметров радиальных базисных функций и оптимального числа нейронов сети [56].

Путеводитель по литературе. С *теоремой Колмогорова* и ее историей можно познакомиться по работе [47]. Эта теорема и нейросетевое обобщение *теоремы Стоуна* приведены в статьях *А. Н. Горбаня* [48–50]. Методы нелинейного прогноза обсуждаются в книгах [30, 43] и в сборнике *«Пределы предсказуемости»* (см. ссылку [58]). Много полезных замечаний об ошибках предсказания можно найти в замечательном обзоре [51]. Коррекция прогноза была предложена в статье [52], а неоднородное вложение —

в работе [53]. Принцип минимальной длины описания обсуждается в лекциях *С.А.Шумского* [54] и *Риссанена* [55]. Его применение к оптимизации нейронной сети предложено в [56]. *Байесовы* аргументы в пользу логистической функции обсуждаются в препринте [57].

Пределы предсказуемости: хаотическая динамика и эффект Эдипа в мягких системах

> Что было, то и будет; и что делалось, то и будет делаться, и нет ничего нового под солнцем.
>
> *Книга Екклесиаста «Гл. 1 ст. 9»*

Мы начнем с простого, но поучительного примера прогноза хаотической системы. Фазовым пространством здесь является единичный отрезок $M = [0, 1]$, а динамика определяется дискретным кусочно-линейным отображением $x(n+1) = 2x(n), \pmod 1$. Таким образом, каждая последующая координата фазовой точки получается удвоением предыдущего значения. Если полученное число больше 1, то эта единица отбрасывается — это и означает символ $\pmod 1$ [(1)]. Легко найти общее решение $x(n) = 2^n x(0)$. Воспользуемся тем, что каждое число из интервала $[0, 1]$ можно записать в форме двоичной дроби: $x(0) = 0, a_1 a_2 a_3 \ldots a_n \ldots$, где a_i — принимают значения 0 либо 1 [(2)]. Умножение на 2 в двоичной арифметике эквивалентно переносу запятой вправо на одну позицию и отбрасыванию единицы $\pmod 1$. Поэтому такое отображение часто называют *сдвигом Бернулли*. Его график приведен на рис. 13.

Рассмотрим некоторое начальное значение, например 0.10111001, заданное с точностью до восьми знаков. Первая цифра после запятой (1) означает, что начальная точка находится в правой половине единичного интервала ($0,5 < x(0) < 1$); вторая цифра (0) означает, что начальная точка находится в интервале $0,5 < x(0) < 0,75$; и т. д. Следовательно, каждая итерация (а это вариант интегрирования для уравнения!) уточняет «адрес» начальной точки. Выпишем первые три шага итерации для некоторого на-

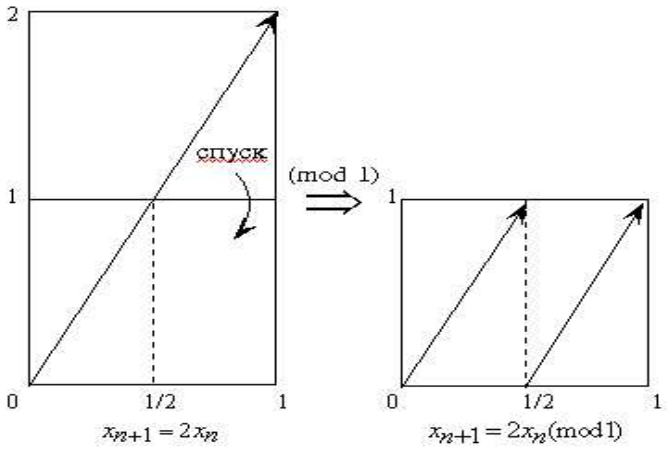

Рис. 13. Сдвиг Бернулли

чального значения:

$$x(0) = 0.10111001\,,$$
$$x(1) = 0.0111001\,,$$
$$x(2) = 0.111001\,,$$
$$x(3) = 0.11001\,.$$

Мы видим, что каждый последующий шаг динамики не что иное, как *уточнение начального условия!* Предположим теперь, что точность начальных данных ограничена — в нашем примере восемью разрядами. Иными словами, мы не знаем, какой символ стоит в девятом разряде: 0 или 1. Тогда через восемь последовательных итераций эта девятая цифра окажется сразу после запятой, и мы не будем знать, справа или слева от середины интервала окажется фазовая точка. Самое удивительное в том, что наша простая модель не содержит каких либо случайных членов — она полностью детерминирована! Итак, чему же учит сдвиг Бернулли?

- Любой прогноз сводится к уточнению начальных данных, точность которых всегда ограничена!
- Некоторые детерминированные системы имеют ограниченный горизонт предсказуемости [3]. Такие системы называют *хаотическими*.

Рис. 13 подсказывает причину хаотичности. Две ε-близкие начальные точки на горизонтальной оси уже после первой итерации окажутся разделенными расстоянием 2ε (так как коэффициент наклона прямых на рис. 13 равен двум), после второй — $2^2\varepsilon$ и т. д. Через конечное число итераций расстояние между ними достигнет размеров фазового пространства $(= 1)$. Число необходимых для этого итераций и определяют границу прогноза. Для измерения скорости «разбегания» близких траекторий в общем случае используют так называемые *показатели Ляпунова*. Пусть отображение задано в виде: $x(n+1) = f(x(n)), M = [0,1]$. Тогда две близкие точки $x(0)$ и $(x(0)+\varepsilon)$ после N итераций окажутся на расстоянии:

$$\varepsilon e^{N\lambda(x_0)} \approx |f^N(x_0+\varepsilon) - f^N(x_0)|$$

друг от друга. В пределе $\epsilon \to 0, N \to \infty$ получим:

$$\lambda(x_0) = \lim_{N\to\infty}\lim_{\varepsilon\to 0}\frac{1}{N}\ln\Big|\frac{f^N(x_0+\varepsilon) - f^N(x_0)}{\varepsilon}\Big| = \\ = \lim_{N\to\infty}\frac{1}{N}\ln\Big|\frac{df^N(x_0)}{dx_0}\Big|,$$

где $\lambda(x_0)$ — коэффициент растяжения (или сжатия) расстояния между двумя точками за одну итерацию. Его называют *локальным ляпуновским показателем* [4].

Используя это понятие, нетрудно оценить среднее время предсказуемости T_m нашего отображения. Очевидно, наша способность предсказывать кончается, когда расстояние между близкими точками достигнет размеров фазового пространства M — этот размер у нас равен 1, т. е. $\varepsilon \times \exp(\lambda T_m) \sim 1$. Отсюда следует, что $T_m \propto (1/\lambda) ln(\varepsilon^{-1})$ [5]. Эту величину часто называют *горизонтом предсказуемости*. Поскольку точность начальных данных всегда конечна, поведение систем с хаотической динамикой всегда *ограниченно предсказуемо*. Следует заметить, что не всегда следует полагаться на оценки ляпуновских показателей. Для реконструкций, построенных из временных рядов со значительной шумовой компонентой, невозможно получить даже оценку самого простого максимального положительного показателя. В этом случае более полезны оценки горизонта предсказуемости, основанные на эвристических соображениях.

В заключение упомянем еще об одном интересном классе систем, для которых *будущее предвидимо, но не предсказуемо!* Это так *называемые мягкие* системы, существенным элементом которых является человеческий

фактор. Такие системы приходится рассматривать, например, в социологии, демографии и экономике. Представим себе такую ситуацию: в некотором условном государстве футуролог, астролог или эмбедолог предсказывает государственный переворот, инициированный военными. Правительство, получив эту информацию, тут же использует ее. Например, подвергает аресту всех заговорщиков. Сценарий предсказанного Будущего тут же меняется! Или, в другой ситуации, узнав о предсказанном вам благоприятном событии, вы «самоосуществили» его, предприняв ряд шагов в нужном направлении, вещь, которую Вы никогда бы не сделали, не зная предсказания! [(6)]. Обобщение подобных примеров резюмировал *И. В. Бестужев-Лада* в форме следующего постулата [58]:

Постулат 1. *О принципиальной невозможности безусловного предсказания будущих состояний объектов, поддающихся видоизменению действиями на основе решений, принимаемых с учетом подобного прогноза*. Изменения сценария Будущей Реальности действиями, учитывающими предсказание этой Реальности в мягких системах, называют *эффектом Эдипа* [(7)]. Казалось бы, что предсказание поведения таких систем заведомо обречено на провал. Однако неумение или нежелание управлять социальными или экономическими явлениями, которые *в принципе* поддаются управлению, позволяет считать их *квазистихийными* или *квазиестественными*, а следовательно и применять к ним методы прогноза хаотических природных систем. В этом состоит суть **Постулата 2** И. В. Бестужева-Лады [58]. Оба постулата кажутся весьма правдоподобными, но, к сожалению, чисто эвристическими. Поэтому следует быть очень осмотрительным, применяя методы эмбедологии и прогноза к мягким системам — полученный результат вполне может оказаться *квазистихийным*. В прогностике особенно полезен известный совет *Козьмы Пруткова*: *«Имея в виду какое-либо предприятие, помысли сперва, точно ли оно тебе удастся»*.

Примечания

1. Арифметика вычетов по модулю единица используется для того, чтобы избежать потери фазовых точек при эволюции системы. Так, например, точка с координатой $1,3$ не покидает M, так как $1,3 = 0,3$ по модулю единица.

2. Для того чтобы обеспечить единственность разложения, условимся выбирать разложение кончающееся нулями для дробей, имеющих два двоичных представления.

3. Этот эффект связан не с видом уравнения, а с типом нелинейности в них.

4. Для *сдвига Бернулли* $\lambda(x) = \ln 2$. Для многомерных систем каждой пространственной координате соответствует свой показатель. Знак $\lambda(x) > 0$ означает разбегание траекторий, $\lambda(x) < 0$ — сжатие. Аттрактору — неподвижной точке в R^3, соответствуют, например, три отрицательных показателя — $(-, -, -)$, предельному циклу — $(0, -, -)$.

5. Величина ε, которая задает точность начального состояния, влияет на горизонт предсказуемости лишь логарифмически. Для многомерной системы, с размерностью аттрактора d и максимальным *положительным* ляпуновским показателем λ_m^+, существует более точная формула для оценки горизонта [29,39]:

$$T_m = \frac{\Delta I}{d\lambda_m^+},$$

где ΔI — неопределенность начальных данных. Если мы имеем дело с реконструкцией аттрактора из временного ряда, содержащего N равновероятных отсчетов, то $\Delta I = \log_2 N$.

6. Вообще говоря, трудно представить себе наше существование, если бы, например, каждый человек точно знал все этапы своей будущей реальности. Теоретик легального марксизма, русский экономист и философ *Петр Бернгардович Струве* (1870–1944) считал, например, прогноз в социальной области вещью весьма безнравственной: *«Я считаю чудовищной идею научной этики — разве мыслимо такое, чтобы то, что есть, могло обосновывать то, что должно быть».*

7. *Эдип*, в греческой мифологии — сын фиванского царя *Лая* и его жены *Иокасты.* Лаю была предсказана *Аполлоном* смерть от собственного сына, поэтому новорожденного Эдипа с проколотыми булавкой сухожилиями — имя *Эдип* — означает *с опухшими ногами* — отнесли в горы. Его усыновил бездетный царь Коринфа — *Полиб.* Юношей Эдип узнал от дельфийского оракула, что ему предстоит убить своего отца и жениться на матери. Считая настоящим отцом Полиба, Эдип, пытаясь изменить предсказание, ушел из Коринфа. Путешествуя, он встретил и убил в ссоре мужчину, который и был его родным отцом.

Не зная этого, Эдип пришел в Фивы и избавил город от чудовища *Сфинкс*. Благодарные жители сделали его царем и женили на вдовой Иокасте. Конец истории о человеке, пытавшимся изменить предсказанную судьбу очень грустный. Узнав через некоторое время всю правду, Иокаста повесилась на своем поясе, Эдип — выколол себе глаза пряжкой этого пояса и был изгнан из Фив.

Путеводитель по литературе. *Сдвиг Бернулли* и другие дискретные преобразования, обладающие свойством детерминированного хаоса, обсуждаются в [23, 32–33]. Свойства ляпуновских показателей описаны в [26–30, 33]. На web-страничке [53] можно найти много полезных статей, связанных с пределами предсказуемости. Наконец, сборник [58] содержит хорошие обзоры на русском языке по нелинейному прогнозу.

Литература

1. *Фор Р., Кофман А., Дени-Папен М.* Современная математика. – М.: Мир, 1966. – 271 с.
2. *Колмогоров А. Н., Фомин С. В.* Элементы теории функций и функционального анализа. – М.: Наука, 1989. – 496 с.
3. *Келли Дж. Л.* Общая топология. – М.: Наука, 1968. – 432 с.
4. *Стинрод Н., Чинн У.* Первые понятия топологии. – М.: Мир, 1967. – 224 с.
5. *Прасолов В. В.* Наглядная топология. – М.: МЦНМО, 1995. – 111 с.
6. *Хирш М.* Дифференциальная топология. – М.: Мир , 1979. – 280 с.
7. *Милнор Дж., Уоллес А.* Дифференциальная топология. – Платон, 1997. – 277 с.
8. *Мищенко А. С., Фоменко А. Т.* Дифференциальная геометрия и топология. – М.: МГУ, 1980. – 439 с.
9. *Макаренко Н. Г.* Многообразия, вложения,погружения и трансверсальность // В сб. *Проблемы Солнечной активности*. – Ленинград: ФТИ, 1991. – с. 13–27.
10. *Chillingworth D.* Differential topology with a view to applications. – Pitman Press, 1976. – 291 pp.
11. *Guillemin V., Pollak A.* Differential topology. – Prentice-Hall, Inc, Englewood Cliffs, New Jersey, 1974. – 219 pp.
12. *Рихтмайер Р.* Принципы современной математической физики. – М.: Мир, 1984 – т.2 – 486 с.

13. *Ильяшенко Ю. С., Ли В.* Нелокальные бифуркации. – М.: МЦНМО-ЧеРо, 1999. – 415 с.

14. *Hunt B. R., Sauer T., Yorke J. A.* Prevalence: A translation-invariant “almost every” on infinite-dimensional spaces // *Bull. Amer. Math. Soc., (New Series)*. – 1992. – v. 27 – pp. 217–238.

15. *Ott E., Sauer T., Yorke J.A.* Coping with chaos: Analysis of chaotic data and the exploitation of chaotic systems. – John Wiley and Sons, 1994. – 432 pp.

16. *Sauer T., Yorke J. A., Casdagli M.* Embedology // *J. Statist. Phys*.– 1991. – v. 65. – pp. 579–616.

17. *Packard N. H., Crutchfield J. P., Farmer J. D., Shaw R. S.* Geometry from a time series // *Phys. Rev. Lett.* – 1980. – v. 45. – pp. 712–716.

18. *Takens F.* Nonlinear dynamics and turbulence // Ed. *Barenblatt G. J., Jooss G., Joseph D. D.* – N. Y. Pitman. – 1983. – pp. 314–333.

19. *Афраймович В. С., Рейман А. М.* Размерности и энтропии в многомерных системах // *Нелинейные волны. Динамика и эволюция*. – М.: Наука, 1989. – с. 238–262.

20. *Takens F.* Detecting strange attractors in turbulence // *Lecture Notes in Math*. – 1981. – v. 898. – pp. 366–381.

21. *Noakes L.* The Takens embedding Theorem // *Inter. J. Bifurcation and Chaos*. – 1991. – v. 1. – pp. 867–872.

22. *Rapp P. E., Schah T. I., Mees A. I.* Models of knowing and the investigation of dynamical systems // *Physica D*. – 1999. – v. 132. – pp. 133–149.

23. *Кроновер Р. М.* Фракталы и хаос в динамических системах. – М.: ПОСТМАРКЕТ, 2000. – 350 с.

24. *Макаренко Н. Г.* Фракталы, аттракторы, нейронные сети и все такое // *Лекции по нейроинформатике*ю – М.: МИФИ, 2002. – ч. 2. – с. 121–169.

25. *Gilmore R.* Topological analysis of chaotic dynamical systems // *Rev. Mod. Phys*. – 1998. – v. 70 – pp. 1456–1529.

26. *Ruelle D.* Chaotic evolution and strange attractors. The statistical analysis of time series for deterministic nonlinear systems. – Cambridge Univ. Press., 1989. – 96 p.

27. *Abarbanel H. D. I., Rabinovich M. I., Sushchic M. M.* Introduction to nonlinear dynamics for physicists. – World Scientific. – 1993. – v. 53 – 158 p.

28. *Eckmann J. P., Ruelle D.* Ergodic theory of chaos and strange attractors // *Rev. Mod. Phys*. – 1985. – v. 57. – pp. 617–656.

29. *Schaw R.* Strange attractors, chaotic behavior, and information flow // *Z. Naturforsch*. – 1981. – v. 36a. – pp. 80–112.

30. *Малинецкий Г. Г., Потапов А. Б.* Современные проблемы нелинейной динамики. – М.: Эдиториал УРСС, 2000. – 335 с.

31. *Малинецкий Г. Г.* Хаос. Структуры. Вычислительный эксперимент: Введение в нелинейную динамику. – М.: Эдиториал УРСС, 2002. – 254 с.

32. *Данилов Ю. А.* Лекции по нелинейной динамике – М.: ПОСТМАРКЕТ, 2001. – 184 с.

33. *Шустер Г.* Детерминированный хаос – М.: Мир, 1988. – 240 с.

34. *Milnor J.* On the concept of attractor // *Commun. Math. Phys*.– 1985. – v. 99. – pp. 177–195.

35. *Schreiber T.* Interdisciplinary application of nonlinear time series methods // *Phys. Rep*. – 1999. – v. 308. – No. 2
URL: http://xyz.lanl.gov/chao-dyn/980700

36. *Parlizt U.* Nonlinear time series analysis
URL: http://WWWDPI.Physik.Uni-Goettingen.DE/~ulli/pub95.html

37. *TISEAN* Nonlinear time series analysis
URL: http://WWW.mpipks-dresden.mpg.de/ tisean/

38. *TSTOOL* Software package for nonlinear time series analysis
URL: http://WWW.physik3.gwdg.de/tstool/

39. *Тутубалин В. Н.* Теория вероятности и случайные процессы. – М.: МГУ, 1992. – с. 342–353.

40. *Марпл-мл. С. Л.* Цифровой спектральный анализ и его приложения. – М.: Мир, 1990. – 584 с.

41. *Serio C.* Towards long-term prediction // *Il Nuovo Cimento*. – 1992. – v. 107B. – pp. 681-701.

42. *Farmer J. D., Sidorovich J. J.* Predicting chaotic time series // *Phys. Rev. Lett*. – 1987. – v. 59. – pp. 845–848.

43. *Малинецкий Г. Г.* Синергетика, предсказуемость и детерминированный хаос: Пределы предсказуемости. – М.: ЦентрКом, 1997. – 247 с.

44. *Малинецкий Г. Г., Курдюмов С. П.* Нелинейная динамика и проблемы прогноза // *Вестник Российской Академии наук*. – 2001. – т. 71, № 3. – с. 210–232.

45. *Poggio T., Girosi F.* A Theory of networks for approximation and learning // *MIT AI Lab. Technical Report*. – 1989. – Memo No 1140. – Paper No 31.
URL: http://citeseer.nj.nec.com/poggio89theory.html

46. *Girosi F., Poggio T.* Networks and the best approximation property // *MIT AI Lab. Technical Report*. – 1989. – Memo No 1164. – Paper No 45.
URL: http://www.ai.mit.edu/people/poggio

47. *Арнольд В. И.* О представлении функций нескольких переменных суперпозицией функций меньшего числа переменных // *Математическое просвещение*. – 1958. –. вып. 3. – с. 41–61.

48. *Горбань А. Н.* Возможности нейронных сетей // В сб. *Нейроинформатика*. – Новосибирск: Наука, 1998. – с. 18–45.

49. *Горбань А.Н.* Обобщенная аппроксимационная теорема и вычислительные возможности нейронной сети // *Сибирский журнал вычислит. математики*. – 1998. – т. 1. – с. 11–24.

50. *Gorban A. N.* Approximation of continuous functions of several variables by an arbitraqry nonlinear continuous function of one variable, linear functions, and their superpositions // *Appl. Math. Lett*. – 1998. – v. 11. – pp. 45–49.

51. *Smith L. A.* The Maintenance of uncertainty
URL: http://www.maths.ox.ac.uk/~lenny/papers.html

52. *Judd K., Small M.* Towards long-term prediction // *Physica D*. – 2000. – v. 136. – pp. 31–44.

53. *Judd K., Small M., Mees A.* Achieving good nonlinear models: Keep it simple, vary the embedding, and get the dynamics right //In: *Nonlinear Dynamics and Statistics*, edited by *A. I. Mees*. – 2001. – Birkhäuser, Boston. – pp. 65–80.

54. *Шумский С. А.* Байесова регуляризация обучения // *Лекции по нейроинформатике*. – М.: МИФИ, 2002. –ч. 2. –с. 30–93.

55. *Rissanen J.* Lectures on statistical modelling theory
URL: http://www.cs.tut.fi/ rissanen

56. *Small M., Tse C. K.* Minimum description length neural networks for time series prediction
URL: http://www.eie.polyu.edu.hk/ensmall/pubs.html

57. *Jordan M. I.* Why the logistic function? A tutorial discussion on probabilities and neural networks
URL: ftp://psyche.mit.edu/pub/jordan/uai.ps

58. *Бестужев-Лада И. В.* Будущее предвидимо, но не предсказуемо: Эффект Эдипа в социальном прогнозировании // В сб.: *Пределы предсказуемости*. – М.: ЦентрКом, 1997. – 247 с.

59. *Вигнер Е.* Этюды о симметрии. – М.: Мир, 1971. – 320 с.

Дополнения [12]

Дополнение 1

Р: Мне не очень понятен термин *«нелинейная проекция»*. Я знаю про центральную проекцию, параллельную проекцию, ортогональную проекцию, но все они порождаются проектирующими лучами-прямыми, а откуда же берется нелинейность? Это что — некая «искажающая» проекция? Если да, то как и почему искажающая?

А: Да, вы правы! Это *искажающая* проекция. Только в обобщенном смысле. Например, ортогональная проекция $R^3 \to R^2$ делается просто: из тройки координат x_1, x_2, x_3 удаляем, скажем x_3 — вот вам и проекция на плоскость. Временной ряд считают *нелинейной проекцией* фазовой траектории динамической системы $f^k(x) : M \to M$, потому что последовательность скалярных отсчетов h_n, $n = 1, 2, \ldots$ порождается наблюдаемой функцией, которая и играет роль «проектора» $h : M \to R$ как:

$$h_1 = h(x),\ h_2 = h(f(x)),\ h_3 = h(f^2(x)),\ \ldots,\ h_m = h(f^{m-1}(x)) \ldots$$

В общем случае h — нелинейная функция. Конечно она «искажает»! Впрочем, в простейшем варианте это могут быть искажения, вызванные возмущениями. Ну например, качается физический маятник, а мы регистрируем луч лазера отраженный от его поверхности. Поставим на пути отраженного луча аквариум, пусть в нем плавает карась и «баламутит» воду. Принимаемый сигнал искажается. Совсем другой пример — нас интересует теория внутреннего строения звезд. В простейшем варианте светимость звезды зависит от массы и радиуса. Именно эти величины входят, например, в уравнения. Но мы наблюдает всего лишь блеск (светимость) как функцию времени. Конечно она зависит от упомянутых величин, но с помощью нелинейной функции или даже суперпозиции таких функций. Вот это-то и есть нелинейная проекция! Важно чтобы функция h была достаточно гладкой. Поскольку она зависит от времени сложным образом $h(x(t))$, обычно требуют *липшиц-непрерывности* относительно x [19] и (или) существования двух производных по t.

[12] Между редактором (**Р**) лекций Школы-семинара и автором (**А**) данной лекции несколько раз возникала дискуссия, касавшаяся принципиальных моментов, связанных с темой лекции. Отдельные фрагменты этой дискуссии, позволяющие ярче высветить упомянутые моменты, приводятся здесь в качестве дополнений к лекции.

Дополнение 2

Р: Мне тоже нравится приведенная *платоновская аллегория* — она хорошо иллюстрирует процесс восприятия окружающего мира через наблюдение за ним. Но ведь если прочитать соответствующий фрагмент «Диалогов» целиком [13], то дальше-то там выводы, как Вы помните, не столь благостные, как можно заключить из примечания 3 на странице 88. Когда узника выводят, наконец, «на свет божий» и он видит то, что раньше наблюдал только через тени, то ему становится ясно, насколько же беден был весь его предыдущий «теневой» опыт!

А: Ну конечно, именно это и имелось в виду (см. выше Дополнение 1 о нелинейной проекции): бедность теневого опыта — это восстановление образа по тени «с точностью до». В случае эмбедологии — с точностью до топологии. Круг узник распознает как овал или эллипс, но он не должен увидеть восьмерку!

Р: А если имеет дело нелинейная проекция, как чаще всего и будет? Пусть, например, перед пещерой стоит тот же самый аквариум с карасем и свет в пещеру проходит только через него, через этот аквариум. Будут ли здесь основания считать, что мы сможем «правильно» реконструировать действительность по ее теням? Возможны ли такие нелинейные преобразования, при которых нельзя будет говорить даже и о реконструкции с точностью до топологии? Другими словами, *где пределы применимости* методов эмбедологии?

А: Эти пределы конечно, есть. Именно их и определяет *Кредо идеального экспериментатора* (см. с. 115). Ну например, если тень дает не динамическая система с аттрактором. Или если вы не выполнили условие вложения $2d + 1$. Оно, правда, часто избыточно.

Например, тень от листа бумаги — двумерная. Но если лист повернуть ребром, то тень одномерная. Однако ситуаций, когда лист оказывается ребром к экрану, меньше, чем всех остальных его положений.

Бублик обычно дает тень с дыркой. Его можно повернуть так, что получится прямоугольник — тоже ребром. Но опять же — таких ситуаций мало.

Теперь про то, когда не получается с точностью до топологии. Нелинейная проекция (наши наблюдения) может не быть непрерывной, точнее *липшиц-непрерывной* проекцией. Ну, например, это функция типа $1/\log x$. Если доопределить ее в нуле условием $\log 0 = 0$, она будет непрерывной, но

[13] См., например, с. 296–299 в книге: *Платон*. Собрание сочинений в 4-х томах. Том 3. – М.: Мысль, 1994. – (Серия «Философское наследие»).

условие Липшица нарушится. Или это фрактальная функция типа триады Коха. Тогда ничего не получится.

Хороший пример нелинейной проекции, когда все получается. В комнате по сложной траектории летает муха. Типичная муха. Вы записываете хорошим микрофонам звук ее жужжания. Вот это ваш скалярный ряд или «детерминированно порожденная наблюдаемая», как сказал бы *Такенс*. Если использовать эмбедологию, *можно восстановить* ее пространственную траекторию — с точностью до топологии, вложив фонограмму полета в R^m.

Р: «. . . восстановить пространственную траекторию "с точностью до топологии"». Что это значит? Платоновский узник вместо круга может увидеть эллипс, т. е. «несколько не то», что есть «на самом деле». На примере с мухой — чем же придется заплатить за «услуги» эмбедологии?

А: Предположим, что муха описывает в комнате пространственную спираль. Если пробная размерность вложения выбрана неверно, скажем R^2, мы получим наложение нескольких окружностей. Можно проверить, все ли в порядке, т. е. имеем ли мы дело с вложением. Для этого достаточно выбрать на полученной копии три точки и проверить их взаимное расположение для реконструкции в R^3. Если расстояния между центральной точкой и ее соседями скачком увеличиваются (ложные соседи), то, следовательно, размерность была выбрана неверно. Аналогичные эксперименты могли бы делать платоновские узники, проверяя справедливость своих реконструкций. Совсем другое дело «Кредо». Это символ веры, т. е. *догма*. Его почти бессмысленно проверять. Конечно, хорошо было бы дополнить Лекцию некоторым списком *Предубеждений* против символов веры. Они более полезны для практики, чем *Ожидания*. Это, впрочем, относится к любой новой концепции. Мне давно хотелось это сделать: в конце концов всегда легко найти то, что ищешь. . . Но очень трудно увидеть место, где споткнешься.

Дополнение 3

Р: Хотя бы пару слов надо сказать, что такое *топологическая идентификация*. «Просто» идентификация — вещь достаточно известная, но про топологическую идентификацию такого не скажешь.

А: Имеется в виду, что аттрактор оригинальной системы и аттрактор копии будут гомеоморфны — топологически неразличимы. Иными словами, копия и оригинал могут быть отображены друг на друга непрерывными обратимыми преобразованиями. Термин выделен в тексте лекции курсивом,

потому что цитируется из оригинальной статьи *Паккарда* и др. [16]. Кстати, в теореме Такенса, аттрактор и его копия *диффеоморфны*, т. е. могут быть отображены друг в друга обратимыми *дифференцируемыми* преобразованиями.

Дополнение 4

P: Один из первых вопросов, возникающий при начальном знакомстве с *эмбедологией* — о месте данного направления среди уже существующих. Чем, кроме используемого аппарата, рассматриваемый класс задач отличается от традиционной задачи *идентификации*, понимаемой достаточно широко — как получение математической модели объекта по данным наблюдений за ним? Не будет ли введение нового термина тем самым «умножением сущностей» против которого выступал *У. Оккам* (см. с. 131)? Можно, конечно, и некую подобласть знаний снабдить именем собственным, но тогда неплохо бы указать и область, в которую эта подобласть входит.

A: Нет, не будет. Обычно, под идентификацией [14] понимают формирование модели путем анализа данных «вход–выход». Таким образом, идентификация предполагает модель типа «черного ящика». Эмбедология же предполагает построение модели (из тени — образ) вложением тени (временного ряда) в евклидово пространство. Еще одно отличие — полученная реконструкция всегда *многомерна*.

P: Так куда же все-таки отнести эмбедологию? Я не пытаюсь ее «принизить», мне хотелось бы просто понять ее место среди других научных областей. Ваша аргументация на тему «эмбедология — это не идентификация» меня пока не убедила. Формулируя свой вопрос, я хотел сослаться на того же самого *Л. Льюнга*, да потом как-то забыл в спешке. А ведь он пишет буквально следующее: «Решение задачи построения математических моделей динамических систем по данным наблюдений за их поведением составляет предмет теории идентификации. . . » А в предисловии высказывается еще интереснее: «Идентификация систем представляет собой ту часть теории автоматического регулирования и анализа временных рядов. . . » и т. д. Чуть ли не дословно то же самое говорится в соответствующих главах известного справочника под ред. *А. А. Красовского* по теории автоматического управления [15]. И обратите внимание, нигде никаких упо-

[14] См., например: *Льюнг Л.* Идентификация систем: Теория для пользователя: Пер. с англ. под ред. *Я. З. Цыпкина*. – М.: Наука, 1991. – 432 с.

[15] Справочник по теории автоматического управления / Под ред. *А. А. Красовского*. – М.: Наука, 1987. – 712 с.

минаний о парах «вход-выход», вообще об управлениях, о «черном ящике» и т. п. Совсем уж «черный ящик» будет, если мы станем искать модель в виде нейросети, которая представляет собой, как известно, композицию «неизвестно чего», т. е. набор элементов и связей между ними, не имеющих «предметной» интерпретации.

Но ведь типичная ситуация-то как раз не такая. Возьмем в качестве примера такую область как механика полета. Типичная задача идентификации в ней выглядит так. Есть модель движения летательного аппарата в виде системы дифференциальных уравнений, полученных из законов механики (и никакого «черного ящика»!) В этой системе есть неизвестные параметры (функции, на самом деле) — коэффициенты аэродинамических сил и моментов, надо найти их значения. В летном эксперименте записывается поведение самолета (значения его фазовых переменных по времени) в ответ на некоторое начальное возмущение (обычно это ступенчатое отклонение соответствующего органа управления), и из этих данных «извлекаются» значения коэффициентов.

Или возьмем полет неуправляемой ракеты, по регистрации которого опять же надо восстановить модель движения — есть только выходы (а на роль входов могут претендовать разве что начальные условия движения), внешние воздействия (например, турбулентность атмосферы) мы ни контролировать, ни измерять не можем.

А довод о том, что получаемая в эмбедологии реконструкция всегда многомерна мне непонятен — точно так же можно при желании проинтерпретировать и «классику».

А: Отчасти я с вами согласен. Но не во всем. Во всех ваших примерах из наблюдений строится не истинная (в некотором обобщенном смысле) модель, а некоторая правдоподобная логическая схема. От таких моделей требуется лишь «функциональное» сходство с оригиналом. В эмбедологии требуется сходство решений модели и копии с точностью до диффеоморфизмов. Она ведь претендует на восстановление геометрии фазового пространства! Разумеется, при большом желании можно попытаться поместить Слона в чайник и подобрать для новой области подходящие рамки из теории идентификации. Но, быть может, более уместно найти для Слона более достойное вместилище? Можно поступить следующим образом. В математике есть такое старое понятие, как «число Коши», введенное, кажется, *А. Лебегом*, которое относится к способу определения некоторого понятия. Его (понятие) можно просто декларировать. Но можно, согласно

О. Коши, поступить иначе. А именно, привести рецепт получения данной величины, а затем рецепт принять за определение. Скажем определением классической окрошки может служить способ ее приготовления: «Возьмите огурец, вареный картофель, яйцо, колбасу и т. д. Мелко покрошите. Залейте квасом. То, что получится и называется окрошкой». Можно воспользоваться аналогичным приемом и попробовать перечислить, что же позволяет *делать* эмбедология.

1. Из временного ряда (тени) восстановить (с точностью до топологии) и символов веры («кредо») фазовый портрет (интегральные траектории) неизвестной системы, порождающей ряд.
2. Оценить (по копии) размерность аттрактора, т. е. число неизвестных уравнений первого порядка.
3. Оценить (по копии) стохастичность неизвестной системы (ляпуновские показатели и энтропию). Ряды, которые генерируются хаотическими системами, проходят *все* тесты на случайность. И только методы эмбедологии позволяют обнаружить там детерминизм.
4. Восстановить (и это делается, правда не всегда) сами уравнения.
5. Реализовать нелинейную схему прогноза.
6. Обнаружить и оценить взаимодействие двух систем и даже определить его направленность. И это в том случае, когда коэффициент линейной корреляции близок к нулю!

Не мало, не правда ли, если в руках только скалярный временной ряд? Правда еще и Символы веры — Кредо Идеального экспериментатора (см. с. 115). В самом примитивном варианте, эмбедологией можно назвать набор алгоритмов, позволяющий из «тени» системы построить ее диффеоморфную модель.

Р: Я же ведь не пытаюсь доказать, что эмбедология — слабая область с не нужными никому результатами. Не пытаюсь я и утверждать, что эмбедология есть ветвь *традиционной* идентификации. Но ведь если появляется некая новая область — это же не «кошка, гуляющая сама по себе», она имеет не только различия с другими научными областями, но и сходства! И мне-то представляется, что такое сходство у эмбедологии есть с идентификацией, если ее трактовать в духе того определения из Льюнга, что я цитировал. Ведь заметьте, там речи не идет о линейности-нелинейности, применяемых методах и прочем — это уже вещи вторичные, хотя и безусловно важные. И, на мой взгляд, *сходство* эмбедологии и теории идентификации — в решаемой задаче, а *различие* (коренное!) — в используемых подходах и методах.

Тем эмбедология, на мой взгляд, и интересна — взяли старую почтенную задачу (а ей лет уж столько, сколько и науке) и подошли к ней так, как никто пока не подходил, получив за счет этого качественный рывок.

A: Я уже почти согласен с тем, что между идентификацией и эмбедологией очень много общего. И из этой общности, наверное, еще не все извлечено. Мне кажется вопрос, куда отнести эмбедологию, это больше вопрос *семантики*. Истинный смысл любого определения следует искать не в его формальном определении, а употреблении. Ведь именно так Платон в «Диалогах» пытался найти истинное значение некоторых слов.

Существовала, кстати, точка зрения, известная в средние века как *номинализм* (ее активно защищал, в частности, *У. Оккам*, а в новейшее время — английский математик *Г. Х. Харди* (1877–1947)), согласно которой общие имена (т. е. имена, связанные не с конкретными предметами, а с их классами) являются чисто словесными, произвольными знаками, а не относятся к объективным сущностям. Великолепный пример номинализма содержится в прелестном диалоге *Алисы* и *Шалтай-Болтая* из известной книги *Льюиса Кэррола* [16]:

— Когда *я* беру слово, оно означает то, что я хочу, не больше и не меньше — сказал Шалтай-Болтай презрительно.

— Вопрос в том, подчинится ли оно вам, — сказала Алиса.

— Вопрос в том, кто из нас здесь хозяин, — сказал Шалтай-Болтай. — Вот в чем вопрос!

P: «. . . между идентификацией и эмбедологией очень много общего. . . » На этом обсуждение «взаимоотношений» идентификации и эмбедологии можно пока и завершить.

Действительно, между ними много общего в решаемых задачах, но не в применяемых методах!

В то же время, считать эмбедологию «всего лишь» разделом теории идентификации было бы, по-видимому, несправедливо.

В самом деле, обратимся опять к упоминавшейся книге *Л. Льюнга* (с. 15): «Формирование моделей на основе результатов наблюдений и исследование их свойств — вот, по существу, основное содержание науки. Модели (“гипотезы”, “законы природы”, “парадигмы” и т. п.) могут быть

[16] См. *Кэрролл Л.* Приключения Алисы в стране чудес. Сквозь зеркало и что там увидела Алиса, или Алиса в Зазеркалье: пер. с англ. *Н. М. Демуровой, С. Я. Маршака, Д. Г. Орловской* и *О. А. Седаковой*. – СПб.: Кристалл, 1999. – 431 с. – (Серия «Библиотека мировой литературы»).

более или менее формализованными, но все обладают той главной особенностью, что связывают наблюдения в некую общую картину». А далее идет цитата, уже приводившаяся выше.

Если это высказывание понимать буквально, то получим такую широкую трактовку теории идентификации, что практически все науки надо будет считать ее разделами! А это уже явный «перебор», из него видно, что расширяя «область влияния» той или иной научной области, надо и меру знать. . .

. . . **А:** Вообще-то еще академик *В. А. Фок* говорил, что основная задача науки «упорядочить наши впечатления».

Николай Григорьевич МАКАРЕНКО, в. н. с, к. ф.-м. н., руководитель группы в Лаборатории компьютерного моделирования (Институт математики, Алма-Ата, Казахстан). Область научных интересов: фрактальная геометрия, вычислительная топология, алгоритмическое моделирование, детерминированный хаос, нейронные сети, физика Солнца. Имеет более 50 научных публикаций.

С. А. ТЕРЕХОВ
Снежинский физико-технический институт, г. Снежинск;
ООО НейрОК, г. Москва,
E-mail: alife@narod.ru

ВВЕДЕНИЕ В БАЙЕСОВЫ СЕТИ

Аннотация

Байесовы сети представляют собой графовые модели вероятностных и причинно-следственных отношений между переменными в статистическом информационном моделировании. В байесовых сетях могут органически сочетаться эмпирические частоты появления различных значений переменных, субъективные оценки «ожиданий» и теоретические представления о математических вероятностях тех или иных следствий из априорной информации. Это является важным практическим преимуществом и отличает байесовы сети от других методик информационного моделирования. В обзорной лекции изложено краткое введение в методы вычисления вероятностей и вывода статистических суждений в байесовых сетях.

S. A. TEREKHOFF
Snezhinsk Institute of Physics and Technology (SFTI), Snezhinsk;
NeurOK LLC, Moscow,
E-mail: alife@narod.ru

BAYESIAN NETWORKS PRIMER

Abstract

Bayesian networks represent probabilistic graph models of casual relations between variables in statistical information modelling. Bayesian networks consolidate empirical frequencies of events, subjective beliefs and theoretical representations of mathematical probabilities. It is the important practical advantage which distinguishes Bayesian networks from other techniques of information modelling. This survey lecture presents a brief introduction to calculation of probabilities and statistical inference methods in Bayesian networks.

УДК 004.032.26 (06) Нейронные сети

Вероятностное представление знаний в машине

Наблюдаемые события редко могут быть описаны как прямые следствия строго детерминированных причин. На практике широко применяется вероятностное описание явлений. Обоснований тому несколько: и наличие неустранимых погрешностей в процессе экспериментирования и наблюдений, и невозможность полного описания структурных сложностей изучаемой системы [26], и неопределенности вследствие конечности объема наблюдений.

На пути вероятностного моделирования встречаются определенные сложности, которые (если отвлечься от чисто теоретических проблем) можно условно разделить на две группы:

- технические (вычислительная сложность, «комбинаторные взрывы» и т. п.);
- идейные (наличие неопределенности, сложности при постановке задачи в терминах вероятностей, недостаточность статистического материала).

Для иллюстрации еще одной из «идейных» сложностей рассмотрим простой пример из области вероятностного прогнозирования [5]. Требуется оценить вероятность положительного исхода в каждой из трех ситуаций:

- Знатная леди утверждает, что она может отличить на вкус, был ли чай налит в сливки или наоборот — сливки в чай. Ей удалось это проделать 10 раз в течение бала.
- Азартный игрок утверждает, что он может предсказать, орлом или решкой выпадет монета (которую вы ему дадите). Он смог выиграть такое пари уже 10 раз за этот вечер, ни разу не проиграв!
- Эксперт в классической музыке заявляет, что он в состоянии различить творения Гайдна и Моцарта лишь по одной странице партитуры. Он уверенно проделал это 10 раз в музыкальной библиотеке.

Удивительная особенность — во всех трех случаях мы формально имеем одинаковые экспериментальные свидетельства в пользу высказанных утверждений — в каждом случае они достоверно подтверждены 10 раз. Однако мы с восхищением и удивлением отнесемся к способностям леди, весьма скептически воспримем заявления бравого игрока, и совершенно естественно согласимся с доводами музыкального эксперта. Наши субъективные оценки вероятности этих трех ситуаций весьма отличаются. И, несмотря на то, что мы имеем дело с повторяющимися событиями, весьма

непросто совместить их с классическими положениями теории вероятностей.

Особенно затруднительно получить формулировку, понятную вычислительной машине.

Другая сторона идейных трудностей возникает при практической необходимости вероятностного прогнозирования событий, к которым не вполне применимы классические представления о статистической повторяемости. Представим себе серию экспериментов с бросанием кубика, сделанного из сахара, на влажную поверхность стола. Вероятности исходов последующих испытаний зависят от относительной частоты исходов предыдущих испытаний, при этом исследуемая система каждый раз необратимо изменяется в результате каждого эксперимента. Этим свойством обладают многие биологические и социальные системы, что делает их вероятностное моделирование классическими методами крайне проблематичным.

Часть из указанных проблем решается в вероятностных *байесовых сетях*, которые представляют собой графовые модели причинно-следственных отношений между случайными переменными. В байесовых сетях могут органически сочетаться эмпирические частоты появления различных значений переменных, субъективные оценки «ожиданий» и теоретические представления о математических вероятностях тех или иных следствий из априорной информации. Это является важным практическим преимуществом и отличает байесовы сети от других методик информационного моделирования.

Байесовы сети широко применяются в таких областях, как медицина, стратегическое планирование, финансы и экономика.

Изложение свойств байесовых сетей, которые являются основным предметом этой лекции, будет построено следующим образом. Вначале рассматривается общий байесов подход к моделированию процесса вывода суждений в условиях неопределенности. Далее излагаются алгоритмы вычисления вероятностей сложных событий в байесовых сетях. Затем обсуждаются подходы к обучению параметров таких сетей. В завершение лекции предлагается эффективная практическая методика эмпирического вывода на основе вероятностных деревьев, и рассматриваются приложения байесовых методик. В приложении приведен краткий обзор ресурсов Интернет по данной тематике.

Неопределенность и неполнота информации

Экспертные системы и формальная логика

Попробуем проследить за способом работы эксперта в некоторой определенной области. Примерами экспертов являются врач, проводящий обследование, финансист, изучающий условия предоставления ссуды, либо пилот, управляющий самолетом.

Действия эксперта могут условно быть представлены в виде повторяющейся последовательности из трех этапов:

- получение информации о состоянии окружающего мира;
- принятие решения относительно выбора некоторых действий, по поводу которых у эксперта имеются определенные ожидания последствий;
- приобретение опыта путем сопоставления результатов действий и ожиданий и возврат к первому этапу.

Приобретенные новый опыт и информация о мире позволяют эксперту сообразно действовать в будущем.

Попытки компьютерного моделирования действий эксперта привели в конце 60-х годов к появлению экспертных систем (ЭС) [6], которые чаще всего основывались на продукционных правилах типа «ЕСЛИ условие, ТО факт или действие». Будущее подобных систем связывалось при этом с заменой экспертов их моделями. Однако после первых успехов обнажились проблемы, и первой среди них — серьезные затруднения при попытках работы с нечеткой, недоопределенной информацией.

Следующие поколения ЭС претерпели кардинальные изменения:

- вместо моделирования эксперта моделируется предметная область;
- вместо попыток учета неопределенности в правилах — использование классической теории вероятностей и теории принятия решений;
- вместо попыток замены эксперта — оказание ему помощи.

В конце 80-х годов были предложены обобщения ЭС в виде байесовых [1] сетей [21] и была показана практическая возможность вычислений вероятностных выводов даже для сетей больших размеров.

[1] По-видимому, первые работы по использованию графовых моделей в статистическом моделировании относятся еще к 20-м годам прошлого века. (*Wright S.* Correlation and causation // *Journal of Agricultural Research*, **20**, 557, 1921).

Вернемся к трехэтапному описанию профессиональных действий эксперта. Сейчас нас будет интересовать вопрос, как наблюдения эксперта, т. е. получение им информации о внешнем мире, изменяют его ожидания по поводу ненаблюдаемых событий?

Особенности вывода суждений в условиях неопределенности

Суть приобретаемого знания в условиях неопределенности состоит в понимании, влияет ли полученная информация на наши ожидания относительно других событий. Основная причина трудностей при использовании систем, основанных на правилах, состоит в учете «сторонних», «косвенных» последствий наблюдаемых событий. Проиллюстрируем это на уже успевшем стать классическим примере [15].

Шерлок Холмс вышел из дома утром и заметил, что трава вокруг влажная. Он рассудил [2]: «Я думаю, что ночью был дождь. Следовательно, трава возле дома моего соседа, доктора Ватсона, вероятно, также влажная». Таким образом, информация о состоянии травы у дома Холмса *повлияла на его ожидания* относительно влажности травы у дома Ватсона. Но предположим, что Холмс проверил состояние сборника дождевой воды и обнаружил, что тот — сухой. В результате Холмс вынужден изменить ход своих рассуждений, и состояние травы возле его дома *перестает* влиять на ожидания по поводу травы у соседа.

Теперь рассмотрим две возможные причины, почему трава у дома Холмса оказалась влажной. Помимо дождя, Холмс мог просто забыть выключить поливальную установку накануне. Допустим, на следующее утро Холмс снова обнаруживает, что трава влажная. Это повышает его субъективные вероятности и для прошедшего дождя, и по поводу забытой дождевальной установки. Затем Холмс обнаруживает, что трава у дома Ватсона также влажная и заключает, что ночью был дождь.

Следующий шаг рассуждений практически невозможно воспроизвести в системах, основанных на правилах, однако он абсолютно естественен для человека: *влажность травы у дома Холмса объясняется дождем, и следовательно нет оснований продолжать ожидать, что была забыта включенной поливальная машина. Следовательно, возросшая, было, субъек-*

[2] Мы сознательно обужаем мир возможных причин и способов рассуждений великого сыщика до элементарного их набора, что не мешает, однако, понять, почему и они весьма трудно программируются в машине.

тивная вероятность относительно забытой поливальной машины уменьшается до (практически) исходного значения, имевшего место до выхода Холмса из дома. Такой способ рассуждения можно назвать «попутное объяснение», «контекстное объяснение» или «редукция причины» (explaining away).

Важная особенность «попутного объяснения» состоит в изменении отношений зависимости между событиями по мере поступления информации. До выхода из дома Холмса факты дождя и работы поливальной установки были независимы. После получения информации о траве у дома они стали зависимыми. Далее, когда появилась информации о влажности травы у дома Ватсона, состояние зависимости вновь изменилось.

Эту ситуацию удобно описать при помощи графа, узлы которого представляют события (или переменные), а пара узлов (A, B) связывается направленным ребром, если информация об A может служить причиной для B. В этом случае узел A будет родителем для B, который, в свою очередь, называется узлом-потомком по отношению к A.

История с травой у Холмса и Ватсона представлена на рис. 1.

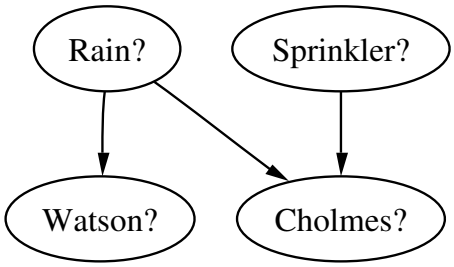

Рис. 1. Граф рассуждений Шерлока Холмса

Граф на рис. 1 может быть отнесен к семейству *байесовых сетей*. В данном примере переменные в узлах могут принимать только булевы значения 1 или 0 (да/нет). В общем случае переменная может принимать одно из набора [3] взаимоисключающих состояний. Из графа на рис. 1 можно сделать несколько полезных выводов о зависимости и независимости переменных.

[3] В традиционной постановке байесовы сети не предназначены для оперирования с непрерывным набором состояний (например, с действительным числом на заданном отрезке). Для представления действительных чисел в некоторых приложениях можно провести разбиение отрезка на сегменты и рассматривать дискретный набор их центров.

Например, если известно, что ночью не было дождя, то информация о состоянии травы у дома Ватсона не оказывает влияния на ожидания по поводу состояния травы у дома Холмса.

В середине 80-х годов были подробно проанализированы способы, которыми влияние информации распространяется между переменными в байесовой сети [21]. Будем считать, что две переменные разделены, если новые сведения о значении одной из них не оказывают влияния на ожидания по поводу другой. Если состояние переменной известно, мы будем называть такую переменную конкретизированной. В байесовой сети возможны три типа отношений между переменными: *последовательные* соединения (рис. 2a), *дивергентные* соединения (рис. 2b), *конвергентные* соединения (рис. 2c).

Ситуация на рис. 2c требует, по-видимому, дополнительных пояснений — как возникает зависимость между предками конвергентного узла, когда становится известным значение потомка [4]. Для простоты рассмотрим пример, когда узел A имеет всего двух предков — B и C. Пусть эти две переменные отвечают за выпадение орла и решки при независимом бросании двух разных монет, а переменная A — логический индикатор, который «загорается», когда обе монеты оказались в одинаковом состоянии (например, обе — решки). Теперь легко понять, что если значение индикаторной переменной стало *известным*, то значения B и C стали *зависимыми* — знание одного из них полностью определяет оставшееся.

Общее свойство (условной) независимости переменных — узлов в байесовой сети получило название d-разделения (d-separation).

Определение (d-разделимость). Две переменные A и B в байесовой сети являются d-разделенными, если на каждом пути, соединяющем эти две вершины на графе, найдется промежуточная переменная V, такая что:

- соединение с V последовательное или дивергентное, и значение V известно, либо
- соединение конвергентное, и нет свидетельств ни о значении V, ни о каждом из ее потомков.

Так, в сети задачи Шерлока Холмса (рис. 1) переменные «Полив?» и «Трава у дома Ватсона?» являются d-разделенными. Граф содержит на пути между этими переменными конвергентное соединение с переменной «Трава у дома Холмса?», причем ее значение достоверно известно (трава — влажная).

[4] Зависимости, возникшие таким способом, называют *индуцированными*.

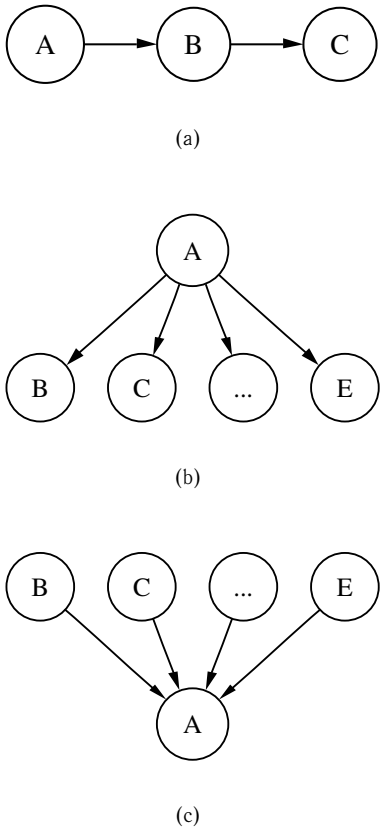

Рис. 2. Три типа отношений между переменными

(a) *Последовательное соединение*. Влияние информации может распространяться от A к C и обратно, пока значение B не конкретизировано. **(b)** *Дивергентное соединение*. Влияние может распространяться между потомками узла A, пока его значение не конкретизировано. **(c)** *Конвергентное соединение*. Если об A ничего не известно, кроме того, что может быть выведено из информации о его предках $B, C, \ldots, E$, то эти переменные предки являются разделенными. При уточнении A открывается канал взаимного влияния между его предками.

Свойство d-разделимости соответствует особенностям логики эксперта-человека, поэтому крайне желательно, чтобы в рассуждениях машин относительно двух d-разделенных переменных новая информация об одной из них не изменяла степень детерминированности второй переменной. Формально, для переменных A и C, независимых при условии B, имеет место соотношение $P(A \mid B) = P(A \mid B, C)$.

Отметим, что интуитивное восприятие условной зависимости и независимости иногда, даже в простых случаях, оказывается затрудненным, так как сложно из всех исходов событий мысленно выделить только те события, в которых значение обусловливающей переменной определено, и далее в рассуждения оперировать только ими.

Вот простой пример, поясняющий эту трудность: в некотором сообществе мужчины среднего возраста и молодые женщины оказались материально более обеспеченными, чем остальные люди. Тогда *при условии фиксированного повышенного уровня обеспеченности* пол и возраст человека оказываются условно *зависимыми* друг от друга!

Еще один классический пример, связанный с особенностями условных вероятностей. Рассмотрим некоторый колледж, охотно принимающий на обучение сообразительных и спортивных молодых людей (и тех, кто обладает обоими замечательными качествами!). Разумно считать, что среди *всех* молодых людей студенческого возраста спортивные и интеллектуальные показатели независимы. Теперь если вернуться к множеству зачисленных в колледж, то легко видеть, что высокая сообразительность эффективно *понижает* вероятность спортивности и наоборот, так как каждого из этих свойств *по-отдельности* достаточно для приема в колледж. Таким образом, спортивность и умственные показатели *оказались* зависимыми при условии обучения в колледже.

Исчисление вероятностей и байесовы сети

Вероятности прогнозируемых значений отдельных переменных

Рассмотрим вначале вопрос о вычислении вероятностей интересующих нас значений переменных в обученной байесовой сети при условии, что в нашем распоряжении имеется некоторая информация о значениях (части) других переменных.

Это — задача статистического вывода суждений. В простейших случаях

она может быть решена точно с использованием соотношений для условных вероятностей. Понятие условной вероятности $P(A \mid B) = x$ составляет основу байесова подхода к анализу неопределенностей. Приведенная формула означает «при условии, что произошло B (и всего остального, что не имеет отношения к A), вероятность возникновения A равна x». Совместная вероятность наступления событий A и B дается формулой полной вероятности:

$$P(A, B) = P(A \mid B) \cdot P(B)\,.$$

Если в нашем распоряжении имеется информация о зависимых переменных (*следствиях*), а суть исследования состоит в определении сравнительных вероятностей исходных переменных (*причин*), то на помощь приходит теорема Байеса.

Пусть имеется условная вероятность $P(A \mid B)$ наступления некоторого события A при условии, что наступило событие B. Теорема Байеса дает решение для *обратной* задачи — какова вероятность наступления более раннего события B, если известно что более позднее событие A наступило. Более точно, пусть $A_1, \ldots, A_n$ — набор (полная группа) несовместных взаимоисключающих событий (или альтернативных гипотез). Тогда *апостериорная* вероятность $P(A_j \mid B)$ каждого их событий A_j при условии, что произошло событие B, выражается через априорную вероятность $P(A_j)$:

$$P(A_j \mid B) = \frac{P(A_j) \cdot P(B \mid A_j)}{P(B) = \sum_{j=1}^{n} P(A_j) \cdot P(B \mid A_j)}\,.$$

Обратная вероятность $P(B \mid A_j)$ называется *правдоподобием* (likelihood), а знаменатель P(B) в формуле Байеса — *свидетельством* (evidence) [5]. Совместная вероятность является наиболее полным статистическим описанием наблюдаемых данных. Совместное распределение представляется функцией многих переменных — по числу исследуемых переменных в задаче. В общем случае это описание требует задания вероятностей всех

[5] Термин “evidence” не получил общеупотребительного аналога в отечественной литературе, и его чаще называют просто «знаменателем в формуле Байеса». Важность evidence проявляется при сравнительном теоретическом анализе различных моделей, статистически объясняющих наблюдаемые данные. Более предпочтительными являются модели, имеющие наибольшее значение evidence. К сожалению, на практике вычисление evidence часто затруднено, так как требует суммирования по всем реализациям параметров моделей. Популярным способом приближенного оценивания evidence являются специализированные методы Монте-Карло.

допустимых конфигураций значений всех переменных, что мало применимо даже в случае нескольких десятков булевых переменных. В байесовых сетях, в условиях, когда имеется дополнительная информация о степени зависимости или независимости признаков, эта функция факторизуется на функции меньшего числа переменных:

$$P(A_1, \ldots, A_n) = \prod_j P\big[A_j \mid pa(A_j)\big] .$$

Здесь $pa(A_j)$ — состояния всех переменных-предков для переменной A_j. Это выражение носит название *цепного правила* для полной вероятности.

Важно, что обусловливание происходит всей совокупностью переменных-предков A_j — в противном случае будет потеряна информация об эффектах совместного влияния этих переменных.

Таким образом, байесова сеть состоит из следующих понятий и компонент:

- множество случайных переменных и направленных связей между переменными;
- каждая переменная может принимать одно из конечного множества взаимоисключающих значений;
- переменные вместе со связями образуют ориентированный граф без циклов;
- каждой переменной-потомку A с переменными-предками $B_1, \ldots, B_n$ приписывается таблица условных вероятностей $P(A \mid B_1, \ldots, B_n)$.

Если переменная A не содержит предков на графе, то вместо условных вероятностей (автоматически) используются безусловные вероятности $P(A)$. Требование отсутствия (ориентированных) петель является существенным — для графов с петлями в цепочках условных вероятностей в общем случае нет корректной схемы проведения вычислений — вследствие бесконечной рекурсии.

На практике нам необходимы распределения интересующих нас переменных, взятые по отдельности. Они могут быть получены из соотношения для полной вероятности при помощи *маргинализации* — суммирования по реализациям всех переменных, кроме выбранных.

Приведем пример точных вычислений в простой байесовой сети, моделирующей задачу Шерлока Холмса. Напомним обозначения и смысл переменных в сети (рис. 1): R — был ли дождь, S — включена ли поливальная

установка, C — влажная ли трава у дома Холмса, и W — влажная ли трава у дома Ватсона.

Все четыре переменные принимают булевы значения 0 — ложь, (f) или 1 — истина (t). Совместная вероятность $P(R,S,C,W)$, таким образом, дается совокупной таблицей из 16 чисел. Таблица вероятностей нормирована, так что

$$\sum_{R=\{t,f\},\ldots,W=\{t,f\}} P(R,S,C,W) = 1\,.$$

Зная совместное распределение, легко найти любые интересующие нас условные и частичные распределения. Например, вероятность того, что ночью не было дождя при условии, что трава у дома Ватсона — влажная, дается простым вычислением:

$$P(R=f \mid W=t) = \frac{P(R=f, W=t)}{P(W=t)} =$$

$$= \frac{\sum_{S=\{t,f\},C=\{t,f\}} P(R=f,S,C,W=t)}{\sum_{R=\{t,f\},S=\{t,f\},C=\{t,f\}} P(R,S,C,W=t)}\,.$$

Из теоремы об умножении вероятностей [7, с. 45] полная вероятность представляется цепочкой условных вероятностей:

$$P(R,S,C,W) = P(R)\cdot P(S \mid R)\cdot P(C \mid R,S)\cdot P(W \mid R,S,C)\,.$$

В описанной ранее байесовой сети ориентированные ребра графа отражают суть те обусловливания вероятностей, которые реально имеют место в задаче. Поэтому формула для полной вероятности существенно упрощается:

$$P(R,S,C,W) = P(R)\cdot P(S)\cdot P(C \mid R,S)\cdot P(W \mid R)\,.$$

Порядок следования переменных в соотношении для полной вероятности, вообще говоря, может быть любым. Однако на практике целесообразно выбирать такой порядок, при котором условные вероятности максимально редуцируются. Это происходит, если начинать с переменных-«причин», постепенно переходя к «следствиям». При этом полезно представлять себе некоторую «историю», согласно которой причины влияют на следствия [6].

[6] Существуют методики частичной автоматизации выбора порядка переменных и синтеза топологии байесовой сети, основанные на большом объеме экспериментальных данных — примеров связей между переменными в данной задаче.

ТАБЛИЦА 1. Априорные условные вероятности в байесовой сети

$P(R=t)$	$P(R=f)$
0.7	0.3

$P(S=t)$	$P(S=f)$
0.2	0.8

R	$P(W=t \mid R)$	$P(W=f \mid R)$
t	0.8	0.2
f	0.1	0.9

R	S	$P(R=t)$	$P(R=f)$
t	t	0.9	0.1
t	f	0.8	0.2
f	t	0.7	0.3
f	f	0.1	0.9

В этом смысле, исключая связь между двумя переменными в сети, мы удаляем зависимость от этой переменной в конкретном выражении для условной вероятности, делая переменные условно независимыми [7].

В байесовой сети для описания полной вероятности требуется указание таблиц локальных условных вероятностей для всех дочерних переменных в зависимости от комбинации значений их предков на графе.

Эта таблица (по случайности!) также содержит 16 чисел, однако вследствие разбиения задачи на мелкие подзадачи с малым числом переменных в каждой, нагрузка на эксперта по их оцениванию значительно снижена.

Перейдем к вычислениям. Найдем сначала полные вероятности двух событий: трава у дома Холмса оказалась влажной, и у дома Ватсона наблюдается то же самое.

$$P(C=t) = \sum_{\substack{R=\{t,f\}\\S=\{t,f\}}} P(R)\cdot P(S)\cdot P(C=t \mid R,S) =$$

$$= 0.3\cdot 0.2\cdot 0.9 + 0.3\cdot 0.8\cdot 0.8 + 0.7\cdot 0.2\cdot 0.7 + 0.7\cdot 0.8\cdot 0.1 = 0.4\,.$$

[7] В литературе часто встречается утверждение «отсутствие ребра в байесовой сети означает независимость между соответствующими переменными». Понимать его следует как условную независимость этих переменных *при отсутствии прочей информации*. Однако, как отмечено в тексте, эти две переменных могут оказаться зависимыми, если они являются предками конвергентного узла, *при наличии информации* о значении в этом узле.

Аналогично,

$$P(W = t) = \sum_{R=\{t,f\}} P(R) \cdot P(W = t \mid R) = 0.31\,.$$

Теперь, если мы достоверно знаем, что ночью был дождь, то эта информация изменит (увеличит) вероятности наблюдения влаги на траве:

$$P(C = t \mid R = t) = \sum_{S=\{t,f\}} P(S) \cdot P(C = t \mid R = t, S) = 0.82\,,$$
$$P(W = t \mid R = t) = 0.8\,.$$

С другой стороны, пусть Холмсу известно, что трава у его дома влажная. Каковы вероятности того, что был дождь, и что дело — в поливальной установке? Вычисления дают:

$$P(R = t \mid C = t) = \frac{P(R = t, C = t)}{P = (C = t)} = \frac{0.054 + 0.192}{0.4} = 0.615\,.$$

$$P(S = t \mid C = t) = \frac{P(S = t, C = t)}{P = (C = t)} = \frac{0.054 + 0.098}{0.4} = 0.38\,.$$

Числитель этой формулы получен, как обычно, путем маргинализации — суммирования по значениям переменных S и W, причем суммирование по последней фактически тривиально из-за нормированности матрицы $P(W \mid R)$.

Значения апостериорных вероятностей и для дождя, и для поливальной машины выше соответствующих величин 0.3 и 0.2 для априорных вероятностей, как и следовало ожидать. Если теперь Холмс обнаружит, что трава у дома Ватсона влажная, то вероятности снова изменятся:

$$P(R = t \mid C = t, W = t) = \frac{P(R = t, C = t, W = t)}{P(C = t, W = t)} =$$
$$= \frac{0.8 \cdot 0.3 \cdot (0.18 + 0.64)}{0.8 \cdot (0.054 + 0.192) + 0.1 \cdot (0.098 + 0.056)} = \frac{0.1968}{0.2122} = 0.9274\,,$$
$$P(S = t \mid C = t, W = t) = 0.2498\,.$$

Ясно, что вероятность дождя возросла вследствие дополнительной информации о влаге на траве у дома Ватсона. Так как высокая вероятность дождя

объясняет влажность травы у дома самого Холмса, то объяснений при помощи другой причины (т. е. включенной поливальной установки) больше не требуется и вероятность ее понижается почти до исходного значения 0.2. Это и есть пример «попутного объяснения», о котором говорилось в предыдущих параграфах.

В общем случае при росте числа переменных в сети, задача точного нахождения вероятностей в сети является крайне вычислительно сложной [8]. Это вызвано комбинаторным сочетанием значений переменных в суммах при вычислении маргиналов от совместного распределения вероятностей, а также потенциальным наличием нескольких путей, связывающих пару переменных на графе. На практике часто используются приближенные методы для оценок комбинаторных сумм, например вариационные методы и многочисленные вариации методов Монте-Карло. Одним из таких приближенных подходов является метод выборок из латинских гиперкубов [9].

Выборочное оценивание вероятностей на латинских гиперкубах

Выборки из латинских гиперкубов начали широко использоваться после удачных решений в области планирования эксперимента, где их применение позволяет уменьшить взаимную зависимость факторов без увеличения числа экспериментов [19].

Представим вначале, что нам требуется разместить 8 фигур на 64 клетках шахматной доски так, чтобы в одномерных проекциях на вертикаль и горизонталь фигуры были распределены максимально равномерно. Решение этой классической задачи дается 8 ладьями, расположенными так, что, в соответствии с шахматными правилами, ни одна из них не бьет другую. Любое из 8! решений годится для приведенных условий. Однако, если мы добавим требование, чтобы двумерное распределение на доске было как можно ближе к равномерному, не все решения окажутся пригодными. В частности, решение, в котором все ладьи выстроились на одной диагонали, существенно недооценивает области двух противоположных углов.

Более удовлетворительное решение можно получить, многократно меняя местами строки для пар наудачу выбираемых ладей. При этом отбираются конфигурации, удовлетворяющие некоторым специальным услови-

[8] Эта задача относится к классу NP-полных. (*Cooper G. F.* The computational complexity of probabilistic inference using Bayesian belief networks // *Artificial Intelligence*. 1990. – **42**, (2–3). – pp. 393–405).

[9] LHS — Latin Hypercube Sampling.

ям — например, распределения, в которых минимальное расстояние между всеми парами ладей максимально, либо конфигурации с ортогональными [10] столбцами. Такой поиск соответствует перестановкам во множестве чисел $1, 2, \ldots, 8$, которые нумеруют строки для упорядоченных по столбцам фигур. Все эти размещения называются *латинскими квадратами*.

Латинские гиперкубы представляют собой прямое обобщение этого примера на случай N точек в пространстве D измерений. Все конфигурации получаются из базовой матрицы размещений размерности $(N \times D)$, в которой в каждом столбце последовательно расположены числа $1, 2, \ldots, N$.

Таким образом, метод латинских гиперкубов позволяет путем целочисленного комбинаторного перебора получить множество из N многомерных векторов, распределение которых, по построению, может обладать полезными дополнительными свойствами, в сравнении с полностью случайными векторами.

В применении к оцениванию условных вероятностей таким важным свойством может являться использование каждого из допустимых значений переменных-предков хотя бы один раз в выборке.

Вновь обратимся к численному примеру [4]. Пусть рассматривается байесова сеть из трех узлов с одним конвергентным соединением.

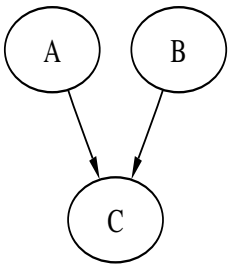

Рис. 3. Архитектура простейшей сети

В отличие от ранее рассмотренных примеров, переменные в сети могут принимать более чем два значения. Для проведения анализа вероятностей построим матрицу латинского гиперкуба (100×3), соответствующую 100 испытаниям. Эта матрица может выглядеть, например, так:

[10] Имеется в виду ортогональность столбцов при переходе к нумерации $-N/2 \ldots + N/2$.

ТАБЛИЦА 2. Архитектура и таблицы вероятностей в простейшей байесовой сети

A	$P(A)$
a_1	0.20
a_2	0.35
a_3	0.45

B	$P(B)$
a_1	0.4
a_2	0.6

	a_1		a_2		a_3	
$P(C \mid A, B)$	b_1	b_2	b_1	b_2	b_1	b_2
c_1	0.01	0.04	0.28	0.06	0.18	0.82
c_2	0.68	0.93	0.52	0.12	0.50	0.10
c_3	0.31	0.03	0.20	0.82	0.32	0.08

ТАБЛИЦА 3. Фрагмент латинского гиперкуба (100×3)

N	A	B	C
1	25	13	74
2	14	91	7
...	...	...	...
39	47	32	56
...	...	...	...
100	69	4	84

Числа от 1 до 100 соответствуют разбиению естественного вероятностного интервала $[0 \dots 1]$ на 100 равных частей. Пусть в i-м испытании для переменной A элемент матрицы гиперкуба равен LHS_iA. Выборка значений переменных далее проводится по следующим формулам:

$$\begin{cases} a_1\,, & \text{если} \quad LHS_iA \leqslant \big[P(a_1) \times 100)\big] \\ a_2\,, & \text{если} \quad \big[P(a_1) \times 100\big] \leqslant LHS_iA \leqslant \big[(P(a_1) + P(a_2)) \times 100\big] \\ a_1\,, & \text{в противном случае} \end{cases}$$

Аналогично, элемент LHS_iB служит для генерации значения переменной B. Пусть для генерации используется 39-я строка гиперкуба в табл. 3., в нашем случае $\{47, 32, 56\}$. Тогда переменная A принимает значение a_2, а B, соответственно, b_1.

Для получения C обратимся к матрице условных вероятностей $P(C \mid A, B)$. Комбинации (a_2, b_1) отвечает 3-й столбец, розыгрыш в котором по $LHS_iC = 56$ дает для C значение c_2. По этой же схеме разыгрываются значения всех переменных для всех 100 случаев. В итоге для переменной C накапливается выборочная гистограмма — оценка фактического распределения вероятности различных значений при заданных предположениях об остальных переменных.

В случае большего числа переменных и более сложной топологии сети, вычисления проводятся по цепочке, от предков к потомкам. К розыгрышу очередной переменной можно приступать, когда значения всех ее предков в данном примере выборки уже установлены [11].

Сложность вычислений пропорциональна числу переменных и размеру выборки, т. е. определяется размером гиперкуба, а не комбинаторным числом сочетаний значений переменных в сети.

Отметим, что метод латинского гиперкуба применим и для моделирования непрерывных распределений. В этом случае для розыгрыша безусловных вероятностей используется известный метод обратных функций распределения [24], а для интерпретации апостериорных вероятностей необходимо применить один из методов сглаживания плотности, заданной на дискретном множестве точек.

Замечание о субъективных вероятностях и ожиданиях

Исчисление вероятностей формально не требует, чтобы использованные вероятности базировались на теоретических выводах или представляли собой пределы эмпирических частот. Числовые значения в байесовых сетях могут быть также и субъективными, личностными, оценками ожиданий экспертов [12] по поводу возможности осуществления событий. У разных лиц степень ожидания (надежды или боязни — по *Лапласу*) события может

[11] Эта схема вычислений пригодна для любых ориентированных графов без направленных петель.

[12] Последовательное искоренение субъективных «вероятностей» имеет уходящие глубоко корнями в прошлое традиции в русской литературе о теории вероятностей. Изданная недавно энциклопедия «Вероятность и математическая статистика» (БРЭ, 1999) цитирует статью выдающегося математика *А. Н. Колмогорова* из энциклопедии 1951 года относительно субъективного понимания вероятности: «Оно приводит к абсурдному утверждению, что из чистого незнания, анализируя лишь одни субъективные состояния нашей большей или меньшей уверенности, мы сможем сделать какие-либо определенные заключения относительно внешнего мира» (ВИМС, с. 97).

быть разной, это зависит от индивидуального объема априорной информации и индивидуального опыта. Можно ли придать этим значениям ожиданий смысл вероятностей — полемический вопрос, выходящий далеко за рамки данной лекции.

Предложен оригинальный способ количественной оценки субъективных ожиданий. Эксперту, чьи ожидания измеряются, предлагается сделать выбор в игре с четко статистически определенной вероятностью альтернативы — поставить некоторую сумму на ожидаемое событие, либо сделать такую же ставку на событие с теоретически известной вероятностью (например, извлечение шара определенного цвета из урны с известным содержанием шаров двух цветов). Смена выбора происходит при выравнивании степени ожидания эксперта и теоретической вероятности. Теперь об ожидании эксперта можно (с небольшой натяжкой) говорить как о вероятности, коль скоро оно численно равно теоретической вероятности некоторого другого статистического события.

Использование субъективных ожиданий в байесовых сетях является единственной альтернативой на практике, если необходим учет мнения экспертов (например, врачей или социологов) о возможности наступления события, к которому неприменимо понятие повторяемости, а также невозможно его описание в терминах совокупности элементарных событий.

Более подробно вопрос о предмете теории вероятностей с точки зрения отечественной школы изложен, например, в [7], байесов подход к понятию вероятности представлен в [14].

Синтез и обучение байесовых сетей

В предыдущем разделе был подробно рассмотрен вопрос статистического вывода суждения о вероятностях различных значений переменных, которые в ряде случаев можно наблюдать на практике. Тем самым, решена задача вероятностного прогнозирования с использованием имеющихся представлений об априорных и условных вероятностях в байесовой сети.

При прогнозировании использовался байесов подход — априорные вероятности для распределения (части) переменных в сети в свете наблюдаемой информации приводят к оценкам апостериорных распределений, которые и служат для прогноза.

В этом разделе дается краткий обзор подходов к построению самих байесовых сетей: синтезу архитектуры и оценке параметров вероятност-

ных матриц для выбранной архитектуры. При обучении мы также будем опираться на последовательную байесову логику, в которой значения параметров также считаются случайными, а вместо поиска универсального единственного значения для каждого параметра оценивается его распределение. При этом, априорные предположения об этом распределении уточняются при использовании экспериментальных данных.

Проблема автоматического выбора архитектуры связей сети заметно сложнее задачи оценивания параметров. Дополнительные сложности возникают при использовании для обучения экспериментальных данных, содержащих пропуски. В самом сложном варианте должны вероятностно моделироваться и изменения архитектуры, и параметры условных вероятностей, и значения пропущенных элементов в таблицах наблюдений [13].

Синтез сети на основе априорной информации

Как уже отмечалось, вероятности значений переменных могут быть как физическими (основанными на данных), так и байесовыми (субъективными, основанными на индивидуальном опыте). В минимальном варианте полезная байесова сеть может быть построена с использованием только априорной информации (экспертных ожиданий).

Для синтеза сети необходимо выполнить следующие действия:

- сформулировать проблему в терминах вероятностей значений целевых переменных;
- выбрать понятийное пространство задачи, определить переменные, имеющие отношение к целевым переменным, описать возможные значения этих переменных;
- выбрать на основе опыта и имеющейся информации априорные вероятности значений переменных;
- описать отношения «причина-следствие» (как косвенные, так и прямые) в виде ориентированных ребер графа, разместив в узлах переменные задачи;
- для каждого узла графа, имеющего входные ребра указать оценки вероятностей различных значений переменной этого узла в зависимости от комбинации значений переменных-предков на графе.

[13] Эта задача, очевидно, является некорректно поставленной, как и всякая задача обучения. Описанный уровень сложности отражает необходимость глубокой регуляризации для согласования распределений такого большого числа ненаблюдаемых переменных.

Эта процедура аналогична действиям инженера по знаниям при построении экспертной системы в некоторой предметной области. Отношения зависимости, априорные и условные вероятности соответствуют фактам и правилам в базе знаний ЭС.

Построенная априорная байесова сеть формально готова к использованию. Вероятностные вычисления в ней проводятся с использованием уже описанной процедуры маргинализации полной вероятности.

Дальнейшее улучшение качества прогнозирования может быть достигнуто путем обучения байесовой сети на имеющихся экспериментальных данных. Обучение традиционно разделяется на две составляющие — выбор эффективной топологии сети, включая, возможно, добавление новых узлов, соответствующих скрытым переменным, и настройка параметров условных распределений для значений переменных в узлах.

Обучение байесовых сетей на экспериментальных данных

Если структура связей в сети зафиксирована, то обучение состоит в выборе свободных параметров распределений условных вероятностей $P[x_j \mid pa(x_j)]$. Для случая дискретных значений переменных x_j, в основном рассматриваемого в лекции, неизвестными параметрами являются значения матричных элементов распределений (см., например, табл. 1).

Максимально правдоподобными оценками этих матричных элементов могут служить экспериментальные частоты реализации соответствующих наборов значений переменных. Вспомним, однако, ситуацию с любительницей чая со сливками и профессиональным музыкантом, описанную во вводной части лекции. Статистические частоты, абсолютизируемые в методе максимального правдоподобия, не учитывают различия в экспертном опыте для разных экспериментальных ситуаций.

Для учета априорных знаний эксперта в вероятностном моделировании обратимся к байесову подходу к обучению. Рассмотрим сначала оба подхода — классический и байесов на примере простой задачи [14] обучения однопараметрической модели.

Пусть в нашем распоряжении имеется монета, свойства которой нам достоверно не известны. Проведем N экспериментов по подбрасыванию

[14] Описание этой задачи, восходящей еще к классикам, можно в том или ином виде найти во многих работах по теории вероятности. Но задача эта столь показательна, что комментарий к ней кажется просто необходимым при иллюстрации особенностей байесова подхода.

монеты и обозначим число выпавших «орлов» как h, а «решек», соответственно, $t = N - h$. Совокупность всех N наблюдений обозначим символом D. Зададимся теперь целью оценить вероятность различных исходов в следующем, $N + 1$ испытании.

В классическом подходе моделируемая система описывается одним фиксированным параметром — вероятностью выпадения «орла» θ, единственное («правильное») значение которого нужно оценить из эксперимента. При использовании метода максимального правдоподобия эта наилучшая оценка, очевидно, равна относительной частоте появления орлов в N экспериментах.

В байесовом подходе значение параметра θ само является *случайной* величиной, распределение которой используется при прогнозировании исхода следующего бросания монеты. Априорная (учитывающая наш опыт ξ до проведения испытаний) плотность распределения θ есть $p(\theta \mid \xi)$. После проведения серии экспериментов наши представления об этом распределении изменятся, в соответствии с теоремой Байеса:

$$p(\theta \mid D, \xi) = \frac{p(D \mid \theta, \xi) \cdot p(\theta \mid \xi)}{p(D \mid \xi) \triangleq \int p(D \mid \theta, \xi) \cdot p(\theta \mid \xi) d\theta} .$$

Функция правдоподобия, разумеется, одна и та же и в классическом, и в байесовом подходе, и равна биномиальному распределению:

$$p(D \mid \theta, \xi) \sim \theta^h \cdot (1 - \theta)^t$$

Прогноз вероятности будущего эксперимента дается формулой суммирования:

$$p(x_{N=1} = H \mid D, \xi) = \int p(x_{N+1} = H \mid \theta, \xi) \cdot p(\theta \mid D, \xi) d\theta =$$

$$= \int \theta \cdot p(\theta \mid d, \xi) d\theta = \langle \theta \rangle_{p(\theta|D,\xi)} .$$

Для получения интерпретируемого результата в замкнутом виде ограничим выбор априорных распределений классом β-*распределений* [14]:

$$p(\theta \mid \xi) = \beta(\theta \mid \alpha_h, \alpha_t) \triangleq \frac{\Gamma(\alpha_h + \alpha_t)}{\Gamma(\alpha_h) \cdot \Gamma(\alpha_t)} \theta^{\alpha_h - 1} \cdot (1 - \theta)^{\alpha_t - 1} .$$

Произведение двух биномиальных распределений вновь дает биномиальный закон, и это проясняет суть использования β-распределения в качестве априорного [15].

$$p(x_{N+1} = H \mid D, \xi) = \frac{\alpha_h + h}{\alpha_h + \alpha + t + h + t}\,.$$

Идея состоит в формализации опыта с бросанием монет путем добавления «искусственных» (полученных в гипотетических предыдущих экспериментах) отсчетов α_h «орлов» и α_t «решек» в экспериментальную серию. Чем больше мы добавим в экспериментальную выборку этих априорных наблюдений, тем меньше наша оценка вероятности $N+1$ испытания будет чувствовать возможные аномальные «выбросы» во множестве D. Поэтому байесово обучение иногда называют *обучением с априорной регуляризацией*.

Задача с одним параметром напрямую обобщается на обучение многопараметрической байесовой сети. Вместо бросания монеты генерируется случайный вектор, составленный из всех параметров сети, при этом исходы значений разыгрываются в соответствии с распределением Дирихле (обобщающем биномиальное распределение на случай более чем двух исходов). Теперь исходом испытания будет реализация значений вектора переменных в байесовой сети.

Если $D = \{D_1, \ldots, D_k, \ldots, D_S\}$ — множество обучающих примеров (каждый элемент D_k является вектором значений всех переменных сети в k-м примере), не содержащее пропущенных значений, то классический вариант обучения состоит в максимизации правдоподобия данных, как функции матричных элементов:

$$L = \frac{1}{N \cdot S} \sum_{j=1}^{N} \sum_{k=1}^{S} \log\bigl[P(x_j \mid pa(x_j), D_k)\bigr]\,.$$

Легко видеть, что обучение в этом подходе состоит в подсчете статистики реализаций векторов ситуаций для каждого матричного элемента в таблицах условных вероятностей. Максимально правдоподобными (наименее противоречащими экспериментальным данным) будут значения вероятностей, равные нормированным экспериментальным частотам.

[15] Априорные распределения, которые в итоге приводят к апостериорным распределениям из того же класса, называют *сопряженными* (conjugate priors).

Классическая схема максимального правдоподобия весьма проста, но ее основным недостатком в многомерном случае являются нулевые оценки вероятностей сочетаний значений переменных, не встретившихся в выборке. Между тем, из-за комбинаторного роста числа таких сочетаний в матрицах условных вероятностей будет все больше нулевых значений, а остальные значения будут сильно зашумлены из-за малого числа отсчетов. В байесовом подходе нулевые значения заменятся априорными вероятностями, которые позволят улучшить оценки прогнозируемых значений, но зато будут нивелированы экспериментальные свидетельства.

Таким образом, оба подхода имеют свои недостатки и преимущества, что и дает пищу нескончаемым дискуссиям сторонников классических и байесовых вероятностей. Практика требует примирения, но в настоящее время равновесие, по-видимому, смещено в сторону классического подхода [16].

Сложная задача автоматического синтеза топологии связей сети также обычно рассматривается на основе и классического, и байесового подходов. После внесения пробного изменения в топологию сети, два структурных варианта сравниваются путем оценки значения знаменателя формулы Байеса (evidence) или правдоподобия всей совокупности обучающих примеров. Далее архитектуры с лучшими значениями ценности отбираются в соответствии с принятой схемой рандомизированного или детерминированного поиска.

Обсудим здесь упрощенный вариант задачи синтеза топологии вероятностной сети для случая бинарного дерева, аппроксимирующего распределение условной вероятности для нескольких независимых и одной зависимой переменной. Обучение таких регрессионных деревьев может использоваться для сравнительного анализа значимости независимых переменных при их отборе для байесовой сетевой модели.

Вероятностные деревья

В этом разделе предлагается (относительно новый) подход последовательного упрощения задачи аппроксимации условной вероятности, основанный на вероятностных деревьях [10, 12]. Рассмотрим матрицу эмпирических

[16] Это — априорная оценка автора, который относит себя, скорее, к сторонникам байесового подхода. Классический подход популярен среди физиков [1], байесовы вероятности — среди специалистов по обучению машин и анализу данных (`www.auai.org`).

данных D, строки которой соответствуют реализациям случайных переменных $\{\vec{X}; Y\}$. Эти данные содержат информацию об условной плотности вероятности $p\,(y \mid \vec{x})$. Нашей целью будет построение функциональной модели этой плотности.

Для простоты ограничимся случаем скалярной зависимой переменной [17]. Если последняя принимает непрерывный ряд значений на некотором отрезке, то задача аппроксимации плотности соответствует задаче (нелинейной) регрессии. При этом независимые переменные могут быть как непрерывными, так и дискретными (упорядоченными или неупорядоченными).

При построении модели условной плотности мы будем опираться на байесов подход вывода заключений из данных (*reasoning*). А именно, имеющуюся информацию о значениях входных переменных обученная машина переводит в апостериорную информацию о характере распределения выходной (зависимой) переменной.

Процесс вывода суждения может носить итерационный характер: машина последовательно использует информацию, содержащуюся во входах для построения сужающихся приближений к апостериорной плотности [18].

Вопрос заключается в эффективном способе синтеза (обучения) такой вероятностной машины на основе лишь имеющихся эмпирических данных.

Метод построения связей и выбора правил в узлах дерева

Остановимся на простой информационной трактовке обучения искомой вероятностной машины. Для этого на множестве значений зависимой переменной введем дискретизацию $y_{\min} < y_1 < \ldots < y_k < \ldots y_{\max}$ таким образом, что в каждый отрезок попадает одинаковое число наблюдений $n_k = p_k \cdot N_0$ из матрицы наблюдений D. Тогда *априорная* заселенность [19] всех интервалов одинакова, что соответствует максимальной энтропии:

$$S_0 = -N_0 \sum_k p_k \log p_k \,.$$

[17] Это равносильно обычному предположению о взаимной независимости зависимых переменных *при условии*, что значения независимых переменных заданы.

[18] В качестве априорного распределения у используется частотная гистограмма значений по совокупности данных D. Путем предобработки данных это распределение приводится к константе (т. н. неинформативный prior).

[19] Экспериментальные частоты в отсутствие другой информации являются наиболее вероятной оценкой *априорной* плотности вероятности.

Это (максимальное) значение энтропии отвечает полному отсутствию информации о возможном значении зависимой переменной. Здесь использовано предположение о независимости и одинаковом (стационарном) характере распределения всех отдельных наблюдений в матрице данных.

Построим теперь такой процесс вероятностного вывода, при котором дополнительная информация о входных (независимых) переменных приводит к уменьшению энтропии распределения выходной переменной.

Прямой путь состоит в генерации двоичного дерева, в каждом узле которого применяется простейшее решающее правило. Применение этого правила разделяет исходную совокупность данных на два множества. Идея заключается в том, чтобы суммарная энтропия распределений $S_1 = S_1^{'} + S_1^{''}$ в полученных множествах была меньше исходной. Другими словами, оптимальное решающее правило выбирается таким образом, чтобы его *информативность* была максимальной, и, соответственно, остаточная энтропия после его применения — минимальной [20].

Ограничимся в этой работе простым классом правил, в которых значение одной из зависимых переменных сравнивается с порогом. Правила такого типа широко применяются в классификаторах на основе бинарных деревьев [12].

Оптимальное (на данном уровне иерархии) правило выбирается в процессе последовательного решения M *одномерных* задач оптимизации (M — число независимых переменных X). Целевая функция — суммарная энтропия двух подмножеств, получаемых при применении правила, изменяемая переменная — значения порога для правила. Выбор останавливается на правиле, обеспечившем максимальное уменьшение энтропии.

Для дискретных переменных вместо решения задачи гладкой оптимизации выполняется перебор возможных значений классов с решающим правилом «свой класс—все остальные».

Иерархический процесс далее продолжается для подмножеств (дочерних ветвей и соответствующих им примеров) данного узла дерева. Процесс формально завершается по достижении нулевой энтропии для каждого узла самого нижнего уровня (значение зависимой переменной для всех примеров данного узла попадает в один интервал исходной дискретизации).

В итоге, каждому узлу полученного дерева приписывается:

[20] Заметим, что традиционно используются другие критерии выбора правил при построении деревьев, в частности, минимизируется дисперсия зависимой переменной, либо какие-то другие функционалы.

1. Эмпирическая оценка плотности условного распределения дискретизованной зависимой переменной (при условии отнесения примера к данному узлу).
2. Оценка выборочной энтропии распределения в этом узле.
3. Решающее правило, позволяющее выбрать дочернюю ветвь с дальнейшим уменьшением энтропии условного распределения.

Свойства вероятностного дерева

Обученная по предлагаемой методике машина предлагает в ответ на информацию о векторе независимых переменных целую серию последовательно уточняющихся приближений к оценке апостериорной плотности условного распределения зависимой переменной.

Обобщающая способность. В таком максимальном варианте дерево является, очевидно, *переобученным*, так как в нем полностью запомнен весь шум, содержавшийся в данных. На практике, иерархия предсказаний плотности может быть остановлена на более ранних уровнях (например, в узле должно содержаться не менее определенного числа примеров обучающей выборки, либо энтропия распределения должна быть выше порогового значения).

Для оценок можно воспользоваться методикой кросс-валидации на основе бутстрэп-выборок [25]. Эти простые эмпирические способы регуляризации могут быть заменены более последовательными методами минимизации длины описания модели и данных [2].

Сходимость метода. Нетрудно убедиться, что для любого конечного набора данных размера N, после отщепления одного примера суммарная энтропия полученных двух множеств (отщепленный пример и совокупность остальных примеров) строго меньше энтропии исходного набора. Действительно, энтропия отдельного примера равна 0, а энтропия оставшихся примеров лимитируется множителем $(N-1)/N$.

Тем самым, метод всегда сходится, по крайней мере за $(N-1)$ шагов. Практическая скорость убывания энтропии в типичных вычислительных экспериментах приведена на рис. 4.

При этом вклад различных входных переменных в снижение энтропии весьма неоднороден.

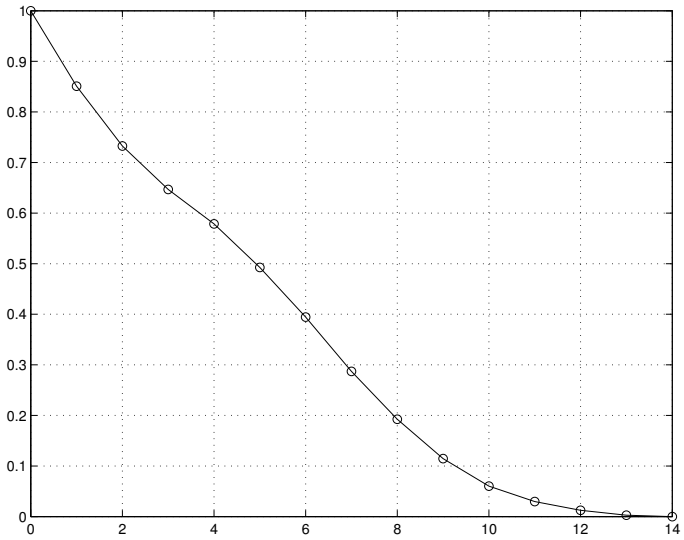

Рис. 4. Зависимость энтропии (в долях к исходному значению) от уровня иерархии дерева

Значимость независимых переменных (факторов). Важным «побочным» эффектом методики является возможность оценки относительной значимости факторов-входов. Все факторы (независимые переменные) можно упорядочить по степени уменьшения суммарной энтропии зависимой переменной, возникшей при использовании фактора в решающих правилах. Не использованные ни в одном из правил переменные в этом подходе признаются *незначимыми*.

Характерное распределение значимости факторов приведено на рис. 5.

Заметим, что предлагаемый метод приводит к весьма разреженному представлению модели данных, что может оказаться полезным при увеличении числа независимых переменных.

Вычислительная сложность метода. Применение метода требует решения множества задач *одномерной* оптимизации на отрезке. При этом затраты на вычисление оптимизируемой функции падают линейно по мере продвижения от корня дерева к листьям.

Задачи безусловной оптимизации функций на отрезке весьма подробно исследованы, и соответствующие методики успешно реализованы в пакетах прикладных программ [8].

Вычислительные затраты на одномерную оптимизацию несравненно меньше затрат на многомерный поиск, который требуется, например, при построении статистических моделей EM-алгоритмом, или обучении многослойных нейронных сетей [2].

Поиск решающих правил в ветвях дерева может быть непосредственно распараллелен, что повышает его эффективность при использовании современных ЭВМ.

Аналоги метода. Близким аналогом вероятностного дерева является *нейросетевая архитектура со встречным распространением* (counter-propagation, *R. Hecht-Nielsen*, [13]).

На кластеризующем [16] слое нейросеть со встречным распространением, являющаяся, по существу, одноуровневым деревом, происходит сквозной просмотр всех ветвей-кластеров. При этом, в отличие от предлагаемого подхода, для выбора ветви дерева используется метрика в пространстве входов вместо решающего правила. Таким образом, кроме повышенных вычислительных затрат, на практике могут возникать трудности с выбором подходящей метрики, что не всегда возможно в случае дискретных и непрерывных входов. Заметим также, что кусочно-постоянная аппроксимация в форме ячеек Вороного в нейросети со встречным распространением качественно отличается от глобального представления функции в виде иерархических сегментов постоянства функции при использовании дерева.

О применениях вероятностных деревьев

Оценка стоимости недвижимости в районе Бостона. В этой, ставшей классической [11,23], задаче предлагается построить прогностическую систему оценки стоимости недвижимости в районе Большого Бостона. База данных [11] содержит 506 наблюдений для 13 зависимых и одной независимой переменной (собственно цены).

Все переменные в базе данных представлены действительными числами, за исключением одной двоичной переменной (признак — находится ли строение в районе реки Чарльз Ривер)

Построение модели дерева заняло чуть больше минуты на персональной ЭВМ. Полное исчерпание энтропии в данных достигнуто при 439 узлах (число выведенных решающих правил чуть больше 200).

Относительная значимость факторов представлена на рис. 5.

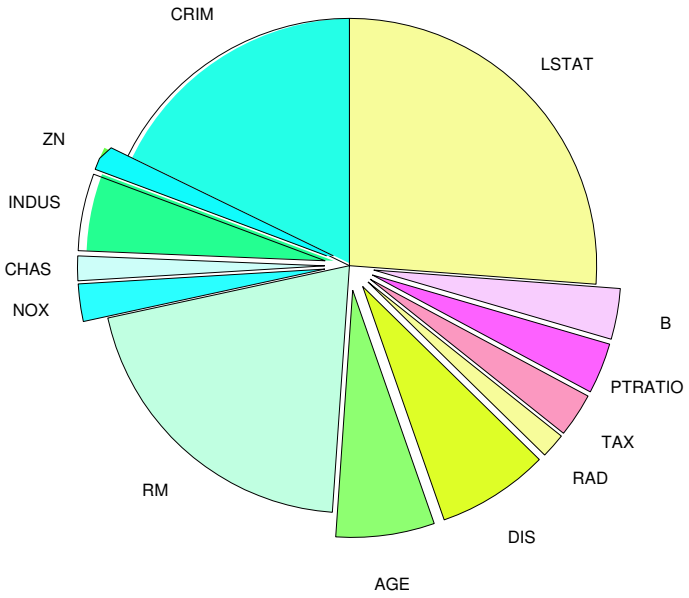

Рис. 5. Относительная значимость факторов в задаче Boston Housing

Почти $3/4$ всей неопределенности в прогнозе цены устраняется на основе информации лишь о трех «прозаических» факторах: ZN — процент земельной площади, выделенной для больших участков (свыше 25000 кв. футов), CRM — уровень преступности, RM — число комнат в доме.

Характер последовательного сужения плотности распределения при перемещении вниз по уровням иерархии дерева показан на рис. 6.

Прогноз свойств смеси химических компонент. Одним из успешных практических применений модели вероятностных деревьев явилось решение задачи о прогнозе свойств смеси химических красителей, применяемых в бытовой химии.

Особенность задачи состоит в том, что около полутора десятков химических компонентов вступают в сложные химические реакции, выход которых нелинейно зависит от концентрации веществ, при этом характер процессов является вероятностным. Последнее обстоятельство приводит к наличию шума и (статистических) противоречий в данных.

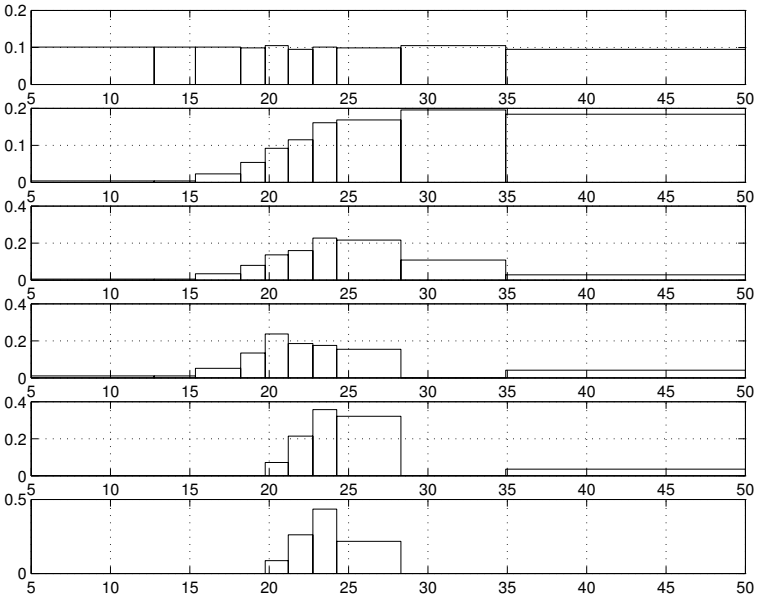

Рис. 6. Упорядочение плотности распределения цены при перемещении по иерархии дерева

Применение вероятностных деревьев позволило построить компактные прогностические модели свойств смеси.

Замечания о компьютерной реализации вероятностных деревьев. Проведенные вычислительные эксперименты показывают высокую эффективность и простоту применения обучающихся машин на основе вероятностных деревьев.

Основные применения полученных моделей состоят в анализе значимости входных факторов и возможности вероятностного прогноза значений выходных переменных при известных входах.

Приведение распределения выходов к константе не является обязательным, и выполнено в примерах здесь лишь для чистоты иллюстрации метода.

В предлагаемом подходе одинаково легко анализируются континуальные и дискретные входы. Малорелевантные входы автоматически игнорируются, что не приводит к усложнению методики.

Примеры применений байесовых сетей

Естественной областью использования байесовых сетей являются экспертные системы, которые нуждаются в средствах оперирования с вероятностями. Назовем лишь несколько областей с успешными примерами практических применений байесовых сетей. Некоторые коммерческие разработки и программное обеспечение обсуждаются в Приложении.

Медицина

Система PathFinder (Heckerman, 1990) разработана для диагностики заболеваний лимфатических узлов. PathFinder включает 60 различных вариантов диагноза и 130 переменных, значения которых могут наблюдаться при изучении клинических случаев. Система смогла приблизиться к уровню экспертов, и ее версия PathFinder-4 получила коммерческое распространение.

Множество других разработок (Child, MUNIN, Painulim, SWAN и др.) успешно применяются в различных медицинских приложениях [15].

Космические и военные применения

Система поддержки принятия решений Vista (*Eric Horvitz*) применяется в Центре управления полетами NASA (NASA Mission Control Center) в Хьюстоне. Система анализирует телеметрические данные и в реальном времени идентифицирует, какую информацию нужно выделить на диагностических дисплеях.

В исследовательской лаборатории МО Австралии системы, основанные на байесовых сетях, рассматриваются, как перспективные в тактических задачах исследования операций. В работе [9] описано применение пакета Netica (см. Приложение) к решению учебной задачи охраны территориальной зоны с моря “Operation Dardanelles”. Модель включает в себя различные тактические сценарии поведения сторон, данные о передвижении судов, данные разведнаблюдений и другие переменные. Последовательное

поступление информации о действиях противников позволяет синхронно прогнозировать вероятности различных действий в течение конфликта.

Компьютеры и системное программное обеспечение

В фирме Microsoft методики байесовых сетей применены для управления интерфейсными агентами-помощниками в системе Office (знакомая многим пользователям «скрепка»), в диагностике проблем работы принтеров и других справочных и wizard-подсистемах.

Обработка изображений и видео

Важные современные направления применений байесовых сетей связаны с восстановлением трехмерных сцен из двумерной динамической информации, а также синтеза статических изображений высокой четкости из видеосигнала (*Jordan*, 2002).

Финансы и экономика

В серии работ школы бизнеса Университета штата Канзас [22] описаны байесовы методики оценки риска и прогноза доходности портфелей финансовых инструментов. Основными достоинствами байесовых сетей в финансовых задачах является возможность совместного учета количественных и качественных рыночных показателей, динамическое поступление новой информации, а также явные зависимости между существенными факторами, влияющими на финансовые показатели.

Результаты моделирования представляются в форме гистограмм распределений вероятностей, что позволяет провести детальный анализ соотношений «риск–доходность». Весьма эффективными являются также широкие возможности по игровому моделированию.

Обсуждение

Байесовы вероятностные методы обучения машин являются существенным шагом вперед, в сравнении с популярными моделями «черных ящиков». Они дают понятное объяснение своих выводов, допускают логическую интерпретацию и модификацию структуры отношений между переменными

задачи, а также позволяют в явной форме учесть априорный опыт экспертов в соответствующей предметной области.

Благодаря удачному представлению в виде графов, байесовы сети весьма удобны в пользовательских приложениях.

Байесовы сети базируются на фундаментальных положениях и результатах теории вероятностей, разрабатываемых в течение нескольких сотен лет, что и лежит в основе их успеха в практической плоскости. Редукция совместного распределения вероятностей большого числа событий в виде произведения условных вероятностей, зависящих от малого числа переменных, позволяет избежать «комбинаторных взрывов» при моделировании.

Байесова методология, в действительности, шире, чем семейство способов оперирования с условными вероятностями в ориентированных графах. Она включает в себя также модели с симметричными связями (случайные поля и решетки), модели динамических процессов (марковские цепи), а также широкий класс моделей со скрытыми переменными, позволяющих решать вероятностные задачи классификации, распознавания образов и прогнозирования.

Благодарности

Автор благодарит своих коллег по работе *С. А. Шумского*, *Н. Н. Федорову*, *А. В. Квичанского* за плодотворные обсуждения. Отдельная благодарность *Ю. В. Тюменцеву* за высококачественную редакторскую работу, важные замечания и его энтузиазм в отношении организации лекций, без которого это издание было бы невозможным. Спасибо родному МИФИ — *Alma Mater* — за организацию Школы по нейроинформатике и гостеприимство.

Литература

1. *D'Agostini G.* Bayesian reasoning in high energy physics — principles and applications. – CERN Yellow Report 99-03, July 1999.
2. *Bishop C.M.* Neural networks for pattern recognition. – Oxford University Press, 1995.
3. *Вентцель Е. С.* Теория вероятностей. – М. Высшая Школа, 2001.
4. *Cheng J., Druzdel M. J.* Latin hypercube sampling in Bayesian networks // In: *Proc. of the Thirteenth International Florida Artificial Intelligence Research Symposium*

(FLAIRS-2000). – Orlando, Florida, AAAI Publishers, 2000. – pp. 287–292.
URL: http://www2.sis.pitt.edu/~jcheng/Latin.zip

5. *Coles S.* Bayesian inference. Lecture notes. – Department of Mathematics, University of Bristol, June 10, 1999.
URL: http://www.stats.bris.ac.uk/~masgc/teaching/bayes.ps
6. *Giarratano J., Riley G.* Expert systems: Principles and programming. – PWS Publishing, 1998.
7. *Гнеденко Б. В.* Курс теории вероятностей. – 7-е изд. – М.: Эдиториал УРСС, 2001.
8. *Дэннис Дж., Шнабель Р.* Численные методы безусловной оптимизации и решения нелинейных уравнений. – М.: Мир, 1988.
9. *Das B.* Representing uncertainties using Bayesian networks. – DSTO Electronics and Surveillance Research Laboratory, Department of Defence. – Tech. Report DSTO-TR-0918. – Salisbury, Australia, 1999.
URL: http://dsto.defence.gov.au/corporate/reports/DSTO-TR-0918.pdf
10. *Fukuda T., Morimoto Y., Morishita S., Tokuyama T.* Constructing efficient decision trees by using optimized numeric association rules // The VLDB Journal, 1996.
URL: http://citeseer.nj.nec.com/fukuda96constructing.html
11. *Harrison D., Rubinfeld D. L.* Hedonic prices and the demand for clean air // J. Environ. Economics & Management. – 1978. – vol. 5. – pp. 81–102.
12. *Hastie T., Tibshirani R., Friedman J.* The Elements of statistical learning. – Data Mining, Inference, and Prediction. Springer, 2001.
13. *Hecht-Nielsen R.* Neurocomputing. – Addison-Wesley, 1990
14. *Heckerman D.*. A tutorial on learning with Bayesian Networks. – Microsoft Tech. Rep. – MSR–TR–95–6, 1995.
15. *Jensen F. V.* Bayesian networks basics. – Tech. Rep. Aalborg University, Denmark, 1996.
URL: http://www.cs.auc.dk/research/DSS/papers/jensen:96b.ps.gz
16. *Kohonen T.* Self-organizing Maps. – Springer, 1995.
17. *Minka T.* Independence diagrams. – Tech. Rep. MIT, 1998.
URL: http://www-white.media.mit.edu/~tpminka/papers/diagrams.html
18. *Blake C. L., Merz C. J.* UCI Repository of Machine Learning Databases, 1998.
URL: http://www.ics.uci.edu/~mlearn/MLRepository.html
19. *Montgomery D. C.* Design and analysis of experiments. – 5th. Ed. – Wiley, 2001.
20. *Neal R. M.* Probabilistic inference using Markov chain Monte Carlo methods. – Technical Report CRG-TR-93-1, 25 Sep 1993, Dept. of Computer Science, University of Toronto.

21. *Pearl J.* Probabilistic reasoning in intelligent systems: Networks of plausible inference. – Morgan Kaufmann, 1988.

22. *Shenoy C., Shenoy P.* Bayesian network models of portfolio risk and return / Y. S. Abu-Mostafa, B. LeBaron, A. W. Lo (Eds.) Computational Finance, 85-104, MIT Press, 1999.

23. StatLib datasets archive. – CMU.
URL: http://lib.stat.cmu.edu/datasets/

24. *Соболь И.М.* Численные методы Монте-Карло. – М.: Наука, 1973.

25. *Терехов С. А.* Нейросетевые аппроксимации плотности в задачах информационного моделирования. Лекция для школы-семинара «Современные проблемы нейроинформатики», Москва, МИФИ, 25–27 января 2002 года.

26. *Терехов С. А.* Нейросетевые информационные модели сложных инженерных систем. Глава IV в кн.: *А. Н.Горбань, В. Л.Дунин-Барковский, А. Н. Кирдин, Е. М. Миркес, А. Ю. Новоходько, Д. А. Россиев, С. А. Терехов, М. Ю. Сенашова, В. Г. Царегородцев*. Нейроинформатика. – Новосибирск, Наука, 1998ю – с. 101–136.

27. *Терехов С. А.* Лекции по теории и приложениям искусственных нейронных сетей. – Снежинск, 1994–1998.
URL: http://alife.narod.ru/lectures/neural/Neu_index.htm

Приложение А. Обзор ресурсов Интернет по тематике байесовых сетей

Байесовы сети — интенсивно развивающаяся научная область, многие результаты которой уже успели найти коммерческое применение. В этом Приложении представлены лишь немногие популярные ресурсы Интернет, спектр которых, однако, отражает общую картину, возникшую в последние 10–15 лет.

AUAI — Ассоциация анализа неопределенности в искусственном интеллекте

URL: http://www.auai.org/

Ассоциация анализа неопределенности в искусственном интеллекте (Association for Uncertainty in Artificial Intelligence — AUAI) — некоммерческая организация, главной целью которой является проведение ежегодной *Конференции по неопределенности в искусственном интеллекте* (UAI).

Конференция UAI–2002 прошла в начале августа 2002 года в Университете Альберты (Эдмонтон, Канада). Конференции UAI проходят ежегодно, начиная с 1985 года, обычно в совместно с другими конференциями по смежным проблемам. Труды конференций издаются в виде книг, однако многие статьи доступны в сети.

NETICA

URL: http://www.norsys.com/index.html)

Norsys Software Corp. — частная компания, расположенная в Ванкувере (Канада). *Norsys* специализируется в разработке программного обеспечения для байесовых сетей. Программа Netica — основное достижение компании, разрабатывается с 1992 года и стала коммерчески доступной в 1995 году. В настоящее время Netica является одним из наиболее широко используемых инструментов для разработки байесовых сетей.

Версия программы с ограниченной функциональностью свободно доступна на сайте фирмы *Norsys*.

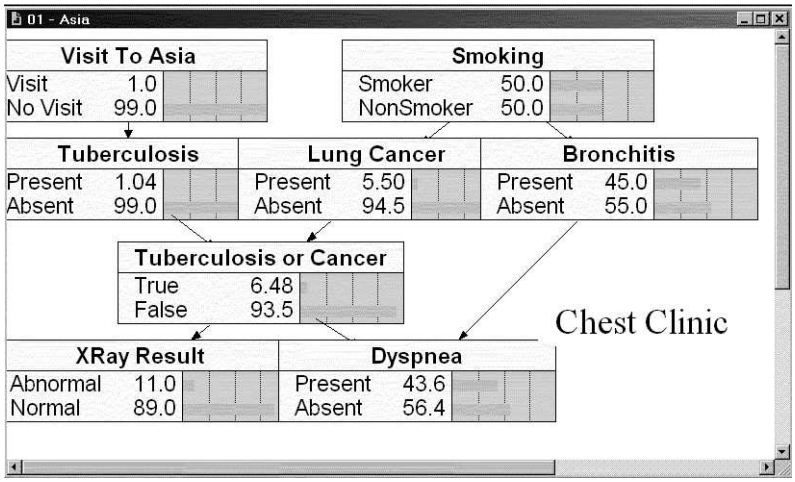

Рис. 7. Пример байесовой сети («Азия») в приложении Netica

Netica — мощная, удобная в работе программа для работы с графовыми вероятностными моделями. Она имеет интуитивный и приятный интерфейс пользователя для ввода топологии сети. Соотношения между переменными могут быть заданы,

как индивидуальные вероятности, в форме уравнений, или путем автоматического обучения из файлов данных (которые могут содержать пропуски).

Созданные сети могут быть использованы независимо, и как фрагменты более крупных моделей, формируя тем самым библиотеку модулей. При создании сетевых моделей доступен широкий спектр функций и инструментов.

Многие операции могут быть сделаны несколькими щелчками мыши, что делает систему Netica весьма удобной для поисковых исследований, и для обучения и для простого просмотра, и для обучения модели байесовой сети. Система Netica постоянно развивается и совершенствуется.

Knowledge Industries

URL: http://www.kic.com/

Knowledge Industries — ведущий поставщик программных инструментальных средств для разработки и внедрения комплексных диагностических систем. При проектировании сложных и дорогостоящих вариантов систем диагностик в компании используется байесовы сети собственной разработки.

Data Digest Corporation

URL: http://www.data-digest.com/home.html

Data Digest Corporation является одним из лидеров в применении методов байесовых сетей к анализу данных.

HUGIN Expert

URL: http://www.hugin.com/

Компания *Hugin Expert* была основана в 1989 году в Ольборге, (Дания). Их основной продукт Hugin начал создаваться во время работ по проекту *ESPRIT*, в котором системы, основанные на знаниях, использовались для проблемы диагностирования нервно-мышечных заболеваний. Затем началась коммерциализация результатов проекта и основного инструмента — программы Hugin. К настоящему моменту Hugin адаптирована во многих исследовательских центрах компании в 25 различных странах, она используется в ряде различных областей, связанных с анализом решений, поддержкой принятия решений, предсказанием, диагностикой, управлением рисками и оценками безопасности технологий.

BayesWare, Ltd

URL: http://www.bayesware.com/corporate/profile.html

Компания *BayesWare* основана в 1999 году. Она производит и поддерживает программное обеспечение, поставляет изготовленные на заказ решения, предоставляет программы обучения, и предлагает услуги консультирования корпоративным заказчикам и общественным учреждениям. Одна из успешных разработок компании, Bayesware Discoverer, основана на моделях байесовых сетей.

Персональные страницы специалистов по байесовым методам

URL: http://stat.rutgers.edu/∼madigan/bayes_people.html

На этой странице собраны адреса домашних страниц специалистов по байесовым сетям из лабораторий и фирм по всему миру.

Сергей Александрович ТЕРЕХОВ, кандидат физико-математических наук, заведующий лабораторией искусственных нейронных сетей Снежинского физико-технического института (СФТИ), заместитель генерального директора ООО «НейрОК». Область научных интересов — анализ данных при помощи искусственных нейронных сетей, генетические алгоритмы, марковские модели, байесовы сети, методы оптимизации, моделирование сложных систем. Автор 1 монографии и более 50 научных публикаций.

НАУЧНАЯ СЕССИЯ МИФИ–2003

НЕЙРОИНФОРМАТИКА–2003

V ВСЕРОССИЙСКАЯ
НАУЧНО-ТЕХНИЧЕСКАЯ
КОНФЕРЕНЦИЯ

ЛЕКЦИИ
ПО НЕЙРОИНФОРМАТИКЕ

Часть 1

Оригинал-макет подготовлен *Ю. В. Тюменцевым*
с использованием издательского пакета LaTeX 2_ε
и шрифтового набора *PSCyr*

Подписано в печать 02.12.2002 г. Формат 60 × 84 1/16
Печ. л. 11, 75. Тираж 200 экз. Заказ №

Московский инженерно-физический институт
(государственный университет)
Типография МИФИ
115409, Москва, Каширское шоссе, 31