



Нейрокомпьютеры: какие они?

Как начинался нейрокомпьютинг? Как выглядят современные нейрокомпьютеры? Какова их производительность и цена? Нейрокомпьютеры и нейро-эмуляторы. Как и где используют нейрокомпьютинг? Основные парадигмы нейрокомпьютинга. Анатомия нейросетей. Классификация нейро-архитектур.

 В науке мало пользы от моделей, которые рабски подчиняются всем нашим желаниям. Мы хотим иметь модели, которые дерзят нам, модели, которые имеют свой собственный ум. Мы хотим получить из наших моделей больше, чем мы в них вложили.


Т.Тоффоли, Н.Маролус, *Машины клеточных автоматов*

 Хватит. Пора наконец рассказать правду про моего друга ЭПИКАКа. Тем более, что он обошелся налогоплательщикам в 776 434 927 долларов 54 цента. Раз они выложили такие денежки, то имеют полное право узнать чистую правду.

К.Воннегут, *ЭПИКАК*

Краткая история нейрокомпьютинга

В прошлой главе появление нейрокомпьютеров представлено как закономерный этап развития вычислительной техники. В результате, у читателя может сложиться впечатление, что и сама идея нейрокомпьютинга - недавнее изобретение. Это, однако, не так. Пути Эволюции редко бывают прямыми. Идеи нейрокомпьютинга появились практически одновременно с зарождением последовательных ЭВМ.

 Ключевая работа Мак Каллока и Питтса по нейро-вычислениям (McCulloch and Pitts, 1943) появилась в 1943 году, на два года раньше знаменитой докладной записки фон Неймана о принципах организации вычислений в последовательных универсальных ЭВМ.

Однако, должны были пройти многие десятилетия, прежде чем радикальное удешевление аппаратуры позволило им заявить о себе в полный голос. Дело в том, что последовательная архитектура обладает весьма ценным преимуществом перед параллельной, решающим на ранних стадиях развития вычислительной техники. А именно, она позволяет получать полезные результаты уже при минимальном количестве аппаратуры. В следующей главе мы покажем, что обучение нейросетей требует больших вычислительных затрат (сложность обучения растет как третья степень размерности задачи). Поэтому нейрокомпьютинг предъявляет достаточно жесткие требования к вычислительной мощности аппаратуры. Только совсем недавно, когда рядовому пользователю РС¹ стала доступна производительность супер-ЭВМ 70-х,

¹ Мы будем пользоваться этим "народным" термином для персональных компьютеров (Personal Computer) вместо более наукообразной аббревиатуры ПЭВМ

нейросетевые методы решения прикладных задач стали приобретать популярность. Как мы увидим далее в этой главе, даже сейчас "настоящие" параллельные нейрокомпьютеры еще слишком дороги и не получили пока широкого распространения. Что уж говорить о конце 50-х, начале 60-х, когда появились первые образцы нейрокомпьютеров.

Первый экспериментальный нейрокомпьютер Snark был построен Марвином Минским в 1951 году. Однако, он не был приспособлен к решению практически интересных задач, и первый успех нейрокомпьютинга связывают с разработкой другого американца - Френка Розенблатта - *персептроном* (от английского perception - восприятие) (Rosenblatt, 1961). Персептрон был впервые смоделирован на универсальной ЭВМ IBM-704 в 1958 году, причем его обучение требовало около получаса машинного времени. Аппаратный вариант - Mark I Perceptron, был построен в 1960 году и предназначался для распознавания зрительных образов. Его рецепторное поле состояло из 400 пикселей (матрица фотоприемников 20x20), и он успешно справлялся с решением ряда задач - мог различать некоторые буквы. Однако по причинам, которые станут понятны по мере знакомства с теорией нейросетей, возможности первых персептронов были весьма ограничены. Позднее, в 1969 году Минский в соавторстве с Пейпертом дает математическое обоснование принципиальной, как им казалось, ограниченности персептронов (Minsky and Papert, 1969), что послужило началом охлаждения научных кругов к нейрокомпьютингу. Исследования в этом направлении были свернуты вплоть до 1983 года, когда они, наконец, получили финансирование от Агентства Перспективных Военных Исследований США, DARPA. Этот факт стал сигналом к началу нового нейросетевого бума.

Интерес широкой научной общественности к нейросетям пробудился в начале 80-х годов после теоретических работ физика Джона Хопфилда (Hopfield, 1982, 1984). Он и его многочисленные последователи обогатили теорию параллельных вычислений многими идеями из арсенала физики, такими как *коллективные взаимодействия* нейронов, *энергия* сети, *температура* обучения и т.д.

Однако, настоящий бум практических применений нейросетей начался после публикации Румельхартом с соавторами метода обучения многослойного персептрона, названного ими методом *обратного распространения ошибки* (*error backpropagation*) (Rumelhart et. al., 1986). Ограничения персептронов, о которых писали Минский и Пейперт, оказались преодолимыми, а возможности вычислительной техники - достаточными для решения широкого круга прикладных задач.

Далее в этой главе мы вкратце опишем современное (правда чрезвычайно быстро меняющееся) состояние нейрокомпьютинга: нейросетевые продукты (как специализированное hardware, так и более доступное software), их сегодняшние применения, а также основные принципы нейровычислений.

Современные нейрокомпьютеры

Элементная база нейрокомпьютеров

Практически все действующие нейрокомпьютеры используют традиционную элементную базу: *микрозлектронные СБИС*. Сотни миллиардов долларов, уже вложенные в развитие этой технологии, дают ей решающее преимущество перед другими альтернативами, такими, как оптические вычисления.

Современная электроника опирается, в основном, на цифровую обработку сигналов, устойчивую к помехам и технологическим отклонениям в параметрах базовых элементов. Цифровая схемотехника предоставляет нейро-конструкторам наиболее богатый инструментарий. Поэтому

неудивительно, что наибольшее распространение получили именно цифровые нейροкомпьютеры. Это по существу - специализированные матричные ускорители, использующие матричный, послойный характер обработки сигналов в нейросетях. Широко используются стандартные процессоры обработки сигналов (DSP - Digital Signal Processors), оптимизированные под такие операции.

Примером современного DSP-процессора, приспособленного для ускорения нейро-вычислений является продукт Texas Instruments TMS320C80 производительностью 2 млрд. операций в секунду. Этот кристалл включает пять процессоров и реализует сразу две технологии - DSP и RISC (4 32-разрядных сигнальных процессора с фиксированной точкой и управляющий процессор с плавающей арифметикой).

Однако, сама природа нейросетевой обработки информации - *аналоговая*, и дополнительного выигрыша в скорости вычислений (по некоторым оценкам $\sim 10^3$ - 10^4) и плотности вычислительных элементов можно добиться, используя специализированную аналоговую элементную базу (Mead, 1989). Наиболее перспективны, по-видимому аналоговые микросхемы с локальными связями между элементами (т.н. клеточные нейросети, CNN - Cellular Neural Networks), например силиконовая ретина фирмы Synaptics. С другой стороны, разработка аналоговых чипов с использованием нетрадиционных схемотехнических решений требует дополнительных и немалых затрат. В настоящее время эти работы на Западе развернуты широким фронтом, например, в рамках проекта SCX-1 (Silicon Cortex - кремниевая кора). Этот проект отличает принципиальная ориентация на массовых производителей аппаратуры, обеспечиваемая совместимостью разрабатываемых нейроплат со стандартами шины VME. Вот как оценивает перспективы этих разработок один из пионеров российского нейрокомпьютинга Феликс Владимирович Широков:

"Системы промышленной автоматизации, построенные на VME, обретут нейроморфный мозг, способность видеть и слышать, ощущать электрические и магнитные поля, воспринимать ультразвуки и радиацию. Они смогут анализировать обстановку и принимать решения. Это будет прививкой разума системам промышленной автоматизации." (Широков, 1998.)

Преимущества обоих подходов пытаются совместить гибридные микросхемы, имеющие цифровой интерфейс с остальной аппаратурой, но исполняющие наиболее массовые операции аналоговым способом.

Приведенные ниже таблицы (Таблица 1 и Таблица 2) дают некоторое представление о сильных и слабых сторонах различных элементных баз и достигнутых результатах.

Таблица 1. Сравнение типов элементной базы

<i>Тип элементной базы</i>	<i>Преимущества</i>	<i>Недостатки</i>
<i>Аналоговая оптическая</i>	Допускает массовые межсоединения	Нет замкнутой технологии оптических вычислений
<i>Аналоговая электрическая</i>	Концептуальная простота схемотехники, выигрыш в емкости схем и скорости вычислений	Жесткие технологические требования, чувствительность к дефектам и внешним воздействиям, малая точность вычислений, трудность реализации массовых соединений
<i>Цифровая электрическая</i>	Развитая замкнутая технология, точность вычислений, устойчивость к технологическим вариациям	Сложность схемных решений, многотактовое выполнение базовых операций, трудность реализации массовых соединений
<i>Гибридная (аналого-цифровая схемотехника, оптоэлектроника)</i>	Аналоговое ускорение базовых операций при цифровом интерфейсе с внешними устройствами, возможности оптической коммутации	Требует дополнительных технологических разработок

Таблица 2. Сравнительные характеристики некоторых нейросхем

Название нейросхемы (фирма-производитель)	Тип элементной базы	Емкость (кол-во нейронов / кол-во синапсов)	Производительность (кол-во умножений с суммированием в сек)
Silicon Retina (Synaptics)	Аналоговая	48 x 48	?
ETANN (Intel)	Аналоговая	64 / 10^4	$2 \cdot 10^9$
N64000 (Inova)	Цифровая	64 / 10^5	$9 \cdot 10^8$
MA-16 (Siemens)	Цифровая	16 / 256	$4 \cdot 10^8$
RN-200 (Ricoh)	Гибридная	16 / 256	$3 \cdot 10^9$
NeuroClassifier (Mesa Research Institute)	Гибридная	7 / 426	$2 \cdot 10^{10}$

Как устроены нейрокомпьютеры

Преимущества нейрокомпьютинга состоит в возможности организовать массовые параллельные вычисления. Поэтому базовые процессорные элементы обычно соединяют в вычислительные комплексы: как можно больше - на одном чипе, а что не поместилось - в мультипроцессорные платы.

Эти платы затем либо вставляют в персональные компьютеры и рабочие станции в качестве нейро-ускорителей, либо собирают в полномасштабные нейрокомпьютеры. В последнем случае избегают задержек в относительно медленных системных шинах РС, правда ценой удорожания аппаратуры.

Таблица 3. Сравнительные характеристики некоторых нейрокомпьютеров

Название нейросхемы, фирма-производитель	Тип	Количество процессорных элементов	Производительность (кол-во умножений с суммированием в сек)	Ориентировочная цена, \$
CNAPS/PC (Adaptive Solutions)	PC-ускоритель	2 CNAPS-1016 процессора (128 нейронов)	$2.5 \cdot 10^9$	$\sim 10^4$
CNAPS (Adaptive Solutions)	Нейро-компьютер	8 CNAPS-1016 процессоров (512 нейронов)	10^{10}	$\sim 10^5$
SYNAPSE-1 (Siemens)	Нейро-компьютер	8 MA-16 процессоров (512 нейронов)	$3 \cdot 10^9$	$\sim 10^5$

Сравнение вычислительных возможностей искусственных и природных нейросетей

Из приведенных в предыдущих таблицах данных следует, что возможности современных нейрокомпьютеров пока довольно скромны, особенно в сравнении с биологическими нейросистемами (см. Рисунок 1). Они едва-едва достигли уровня мухи и еще не дотягивают до таракана.

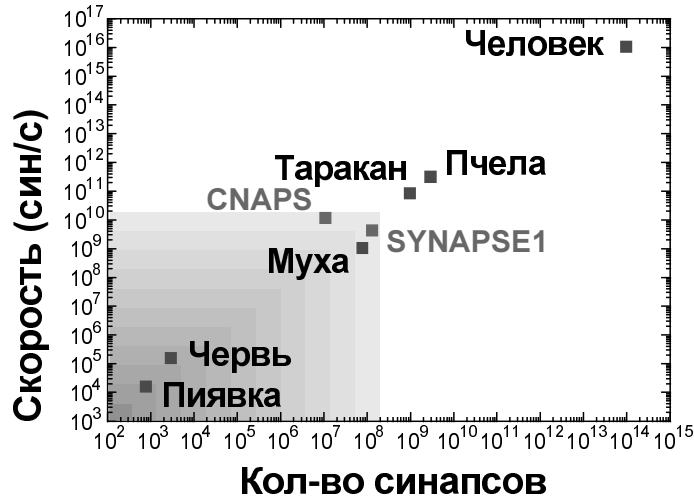


Рисунок 1 Сравнение современных нейрокомпьютеров с естественными нейросистемами

Вспомним, однако, искусство маневра, характерное для полета мухи, и возможности современных нейрокомпьютеров уже не покажутся такими уж скромными.

Сравнение стоимости обычных и нейро- вычислений

Производительность современных персональных компьютеров составляет примерно 10^7 операций с плавающей точкой в секунду (при тактовой частоте системной шины 66 МГц, положив в среднем 6 тактов на одну операцию). Итак, при стоимости всего на порядок больше обычных РС, нейроускоритель в несколько сот раз превосходит их в быстродействии. Таким образом, удельная стоимость современных нейровычислений примерно на порядок ниже, чем у традиционных компьютеров. Это всего лишь следствие специализации матричных процессоров (DSP), имеющих ту же элементную базу, что и универсальные микропроцессоры.

Однако, выигрыш на один порядок в стоимости вычислений редко когда способен стать решающим аргументом для использования специализированной аппаратуры, сопряженной с дополнительными затратами, в том числе на обучение персонала. Поэтому реально нейрокомпьютеры используются в специализированных системах, когда требуется обучать и постоянно переобучать сотни нейросетей, объединенных в единые информационные комплексы, или в системах реального времени, где скорость обработки данных критична.

Например, при обработке экспериментов на современных ускорителях элементарных частиц скорость поступления событий достигает десятков МГц, тогда как конечная информация записывается на ленту со скоростью десятки Гц. Т.е. требуется отбор одного события из миллиона - причем в реальном масштабе времени. Для этого используют несколько каскадов фильтрации событий. Здесь-то вычислительные возможности параллельных нейрокомпьютеров оказываются весьма кстати. Так, в эксперименте Н1 на ускорителе HERA всю аппаратную часть фильтрации событий планируется реализовать на основе нейроплат CNAPS (данные 1994 г.).

Большинство же прикладных систем нейросетевой обработки данных использует *эмуляцию* нейросетей на обычных компьютерах, в частности на РС. Такие программы называются *нейро-эмуляторами*.

Нейро-эмуляторы

Доступность и возросшие вычислительные возможности современных компьютеров привели к широкому распространению программ, использующих принципы нейросетевой обработки данных, о которых мы поговорим подробнее в следующих разделах, но исполняемых на последовательных компьютерах. Этот подход не использует преимуществ присущего нейровычислениям параллелизма, ориентируясь исключительно на способность нейросетей решать неформализуемые задачи².

Преимущества нейро-эмуляторов

Преимущества таких "виртуальных" нейрокомпьютеров для относительно небольших задач очевидны:

- Во-первых, не надо тратить на новую аппаратуру, если можно загрузить уже имеющиеся компьютеры общего назначения.
- Во-вторых, пользователь не должен осваивать особенности программирования на спец-процессорах и способы их сопряжения с базовым компьютером.
- Наконец, универсальные ЭВМ не накладывают никаких ограничений на структуру сетей и способы их обучения, тогда как спец-процессоры зачастую имеют ограниченный набор "защитых" в них функций активации и достигают пиковой производительности лишь на определенном круге задач.

В общем, если речь идет не о распознавании изображений в реальном времени или других приложениях такого рода, а, скажем, об обработке и анализе обычных баз данных, не имеет особого смысла связываться с нейро-ускорителями. Скорее всего, производительности хорошей РС окажется вполне достаточно. Поскольку большинство финансовых применений относится пока именно к этому классу задач, мы будем ориентировать нашего читателя на использование нейро-эмуляторов. Несколько условно нейро-*software* можно разделить на готовые нейро-пакеты общего назначения, более дорогие системы разработки нейроприложений, обладающие большими возможностями, но требующие и больших знаний, и, наконец, готовые комплексные решения с элементами нейросетевой обработки информации, обычно скрытыми от глаз пользователя.

Готовые нейро-пакеты

Это законченные независимые программные продукты, предназначенные для широкого класса задач, в основном - для предсказаний и статистической обработки данных. Большинство из имеющихся на рынке нейропакетов имеет дружественный интерфейс пользователя, не требующий знакомства с языками программирования.

Множество нейро-эмуляторов начального уровня можно найти в Internet как shareware или freeware. Это, обычно, многослойные перцептроны с одним или несколькими правилами обучения. Исключение составляет вполне профессиональный Штутгартский симулятор с большим набором возможностей, работающий, правда, только на UNIX-машинах.

² Некоторые нейро-пакеты, впрочем, предусматривают возможность использования специальных нейро-ускорителей.

Коммерческие пакеты отличаются от свободно распространяемых большим набором средств импорта и предобработки данных, дополнительными возможностями по анализу значимости входов и оптимизации структуры сети. Стоимость коммерческих эмуляторов - масштаба \$1000. Как правило, такие пакеты (BrainMaker Professional, NeuroForecaster, Лора-IQ300) имеют собственный встроенный блок предобработки данных, хотя иногда для этой цели удобнее использовать стандартные электронные таблицы. Так, нейро-продукты группы нейрокомпьютинга ФИАН встраиваются непосредственно в Microsoft Excel в качестве специализированных функций обработки данных. При этом всю предобработку данных и визуализацию результатов можно проводить стандартными средствами Excel, который, кроме того, имеет богатый и расширяемый набор конверторов для импорта и экспорта данных.

Такие пакеты нацелены на решение информационных задач в диалоговом режиме - при непосредственном участии пользователя. Они не применимы в условиях потоковой обработки данных. Кроме того, они не приспособлены для разработки сложных систем обработки данных, состоящих из многих блоков, содержащих, скажем, сотни нейросетей, адаптивно настраивающихся и дообучающихся на вновь поступающих данных. Разработка таких "серьезных" систем требует специального инструментария.

Инструменты разработки нейроприложений

Главное, что отличает этот класс программного обеспечения - способность генерировать "отчуждаемые" нейросетевые продукты, т.е. генерировать программный код, использующий обученные нейросети для обработки данных. Такой код может быть встроен в качестве подсистемы в любые сколь угодно сложные информационные комплексы.³

Примерами подобных систем, способных генерировать исходные тексты программ являются NeuralWorks Professional II Plus (стоимостью от \$3000) фирмы NeuralWare и отечественный Neural Bench (нейро-верстак). Последний интересен, кроме прочего, тем, что может генерировать коды на многих языках, включая Java. Такие Java-апплеты могут использоваться для организации различного рода сервисов в глобальных и локальных сетях. Удобным инструментом разработки сложных нейросистем является MATLAB с прилагающимся к нему нейросетевым инструментарием, органично вписавшимся в матричную идеологию этой системы. MATLAB предоставляет удобную среду для синтеза нейросетевых методик с прочими методами обработки данных (wavelet-анализ, статистика, финансовый анализ и т.д.). Разработанные в системе MATLAB приложения могут быть затем перетранслированы в C++.

Подобные средства разработки используются фирмами, в частности, для создания основанных на нейросетевой обработке данных готовых решений в различных областях.

Готовые решения на основе нейросетей

Это - конечный результат. Здесь нейросети спрятаны от пользователя в недрах готовых автоматизированных комплексов, предназначенных для решения конкретных производственных задач. Например, уже упоминавшийся продукт Falcon встраивается в банковскую автоматизированную систему обслуживания платежей по пластиковым карточкам. В другом случае это будет автоматизированная система управления заводом или реактором. Конечного пользователя, как правило, не интересует способ достижения результата, ему важно лишь качество продукта. Поскольку многие такие готовые решения обладают уникальными возможностями (пока специалисты по нейрокомпьютингу еще в дефиците) и обеспечивают

³ Известно сравнение нейросетей с изюмом, который хорош не столько сам по себе, сколько как добавка, например, в булочки.

реальные конкурентные преимущества, их цена может быть довольно высока - порядка $\$10^6$ - гораздо выше, чем стоимость нейро-hardware.

Испанская компания SEMP занимается повышением эффективности обслуживания кредитных карт VISA, эмитируемых испанскими банками. Количество подобных транзакций - от 500,000 до 1,000,000 в день. Нейросетевая система, разработанная для нее учеными из Мадридского Института Инженерии Знаний (Instituto de Ingenieria del Conocimiento), уменьшила вероятность несанкционированного использования карт на 30-40% для основных каналов мошенничества.

Нейросетевой консалтинг

Описание рынка нейро-продуктов будет не полным без упоминания о нейро-консалтинге. Вместо того, чтобы продавать готовые программы либо инструменты для их разработки, можно торговать и услугами. Например, до изобретения радио большим спросом пользовались барометры, как инструменты предсказания погоды. Теперь же мы просто узнаем погоду по радио или TV, а не предсказываем ее кустарными методами. Некоторые задачи, например такие, как предсказание рыночных временных рядов, являются настолько сложными, что доступны лишь настоящим профессионалам. Не каждая компания может позволить себе издержки, ассоциируемые с передовыми научными разработками (например, постоянное участие в международных конференциях). Поэтому приобретают популярность фирмы, единственной продукцией которых являются *предсказания* рынков. При большом числе клиентов цена таких предсказаний может быть весьма умеренной.

Примером здесь может служить Prediction Company, основанная в 1991 году физиками Дойном Фармером и Норманом Паккардом - специалистами в области динамического хаоса. Первый до этого руководил группой исследования сложных систем в ядерной лаборатории Лос Аламоса, а второй работал в Институте Перспективных Исследований в Принстоне (где когда-то трудился Эйнштейн). Продукция компании пользуется большим успехом среди Швейцарских банков, скупающих прогнозы "на корню" для игры на фондовых и валютных рынках.

Таблица 4. Секторы рынка нейросетевых программных продуктов

<i>Сегмент рынка нейро-продуктов</i>	<i>Преимущества продукции</i>	<i>Недостатки продукции</i>
Нейро-пакеты общего назначения	Не требуют самостоятельного программирования, легко осваиваются, инструмент быстрого и дешевого решения прикладных задач	Не способны к расширению, не способны генерировать отчуждаемые приложения, не могут использоваться для разработки сложных систем или их подсистем
Системы разработки нейроприложений	Могут использоваться для создания сложных систем обработки данных в реальном времени (или их подсистем)	Требуют навыков программирования, более глубокого знания нейросетей
Готовые решения на основе нейросетей	Не предполагают знакомства пользователя с нейросетями, предоставляют комплексное решение проблемы	Как правило - дорогое удовольствие
Нейро-консалтинг	Не предполагает участия пользователя в получении прогнозов, потенциальная дешевизна услуг	Нет возможности дополнить предсказания своим know how. Доступность конфиденциальной информации

Рынок нейропродукции

Объем рынка нейропродукции, структуру которого мы попытались выше очертить, растет стремительными темпами: по разным оценкам - от 30% до 50% в год, перевалив недавно за миллиард долларов (Рисунок 2).



Рисунок 2. Рост объемов продаж нейропродукции

Такой же рост наблюдался в начале 80-х годов на формирующемся в то время рынке персональных компьютеров. Миллиардный оборот тогда явился сигналом для вступления на этот рынок гиганта компьютерной индустрии - IBM. Все мы прекрасно помним как в результате

возникшей конкуренции за деньги конечного пользователя преобразился весь компьютерный мир...

Нейрокомпьютеры и их программные эмуляторы, естественно, интересны не сами по себе, а как инструмент решения практических задач. Только в этом случае нейропродукция будет обладать потребительской стоимостью и иметь соответствующий объем рынка. Где же и как используется нейросетевая продукция сегодня?

Приложения нейрокомпьютинга

Какие задачи решают нейросети

Нейросети наиболее приспособлены к решению широкого круга задач, так или иначе связанных с обработкой образов. Вот список типичных постановок задач для нейросетей:

- Аппроксимация функций по набору точек (регрессия)
- Классификация данных по заданному набору классов
- Кластеризация данных с выявлением заранее неизвестных классов-прототипов
- Сжатие информации
- Восстановление утраченных данных
- Ассоциативная память
- Оптимизация, оптимальное управление

Этот список можно было бы продолжить и дальше. Заметим, однако, что между всеми этими внешне различными постановками задач существует глубокое родство. За ними просматривается некий единый прототип, позволяющий при известной доле воображения сводить их друг к другу.

Возьмем, например, задачу *аппроксимации* функции по набору точек. Это типичный пример некорректной задачи, т.е. задачи не имеющей единственного решения. Чтобы добиться единственности, такие задачи надо *регуляризовать* - дополнить требованием минимизации некоторого регуляризирующего функционала. Минимизация такого функционала и является целью обучения нейросети. Задачи *оптимизации* также сводятся к минимизации целевых функций при заданном наборе ограничений. С другой стороны, *классификация* - это ни что иное, как аппроксимация функции с дискретными значениями (идентификаторами классов), хотя ее можно рассматривать и как частный случай заполнения пропусков в базах данных, в данном случае - в колонке идентификаторов класса. Задача *восстановления* утраченных данных, в свою очередь - это *ассоциативная память*, восстанавливающая прообраз по его части. Такими прообразами в задаче *кластеризации* выступают центры кластеров. Наконец, если информацию удастся восстановить по какой-нибудь ее части, значит мы добились *сжатия* этой информации, и т.д.

Многие представители разных наук, занимающихся перечисленными выше задачами и уже накопившими изрядный опыт их решения, видят в нейросетях лишь перепев уже известных им мотивов. Каждый полагает, что перевод его методов на новый язык нейросетевых схем ничего принципиально нового не дает. Статистики говорят, что нейросети - это всего лишь частный

способ статистической обработки данных, специалисты по оптимизации - что методы обучения нейросетей давно известны в их области, теория аппроксимации функций рассматривает нейросети наряду с другими методами многомерной аппроксимации. Нам же представляется, что именно синтез различных методов и идей в едином нейросетевом подходе и является неоценимым достоинством нейрокомпьютинга. Нейрокомпьютинг предоставляет единую методологию решения очень широкого круга практически интересных задач. Это, как правило, ускоряет и удешевляет разработку приложений. Причем, что обычно забывают за неразвитостью соответствующего hardware, но что, видимо, в конце концов сыграет решающую роль, нейросетевые алгоритмы решения всех перечисленных выше задач заведомо параллельны. Следовательно, все что может быть решено - может быть при желании решено гораздо быстрее и дешевле.

Где применяются нейросети

Наверное, в каждой предметной области при ближайшем рассмотрении можно найти постановки нейросетевых задач. Список областей, где решение такого рода задач имеет практическое значение уже сейчас, приведенный ниже, ограничен лишь опасением утомить читателя. По этой же причине мы ограничились единственным примером конкретного применения в каждой из этих областей.

- **Экономика и бизнес:** предсказание рынков, автоматический дилинг, оценка риска невозврата кредитов, предсказание банкротств, оценка стоимости недвижимости, выявление пере- и недооцененных компаний, автоматическое рейтингование, оптимизация портфелей, оптимизация товарных и денежных потоков, автоматическое считывание чеков и форм, безопасность транзакций по пластиковым карточкам.

☞ Программное обеспечение компании RETEK, дочерней фирмы HNC Software, - лидер среди крупных ритейлеров с оборотом свыше \$1 млрд. Ее последний продукт января 1998 года Retek Predictive Enterprise Solution включает развитые средства нейросетевого анализа больших потоков данных, характерных для крупной розничной торговли. Он также содержит прогнозный блок, чтобы можно было заранее просчитать последствия тех или иных решений. (<http://www.retek.com>)

- **Медицина:** обработка медицинских изображений, мониторинг состояния пациентов, диагностика, факторный анализ эффективности лечения, очистка показаний приборов от шумов.

☞ Группа *НейроКомп* из Красноярска (под руководством Александра Николаевича Горбаня) совместно с Красноярским межобластным офтальмологическим центром им. Макарова разработали систему ранней диагностики меланомы сосудистой оболочки глаза. Этот вид рака составляют почти 90% всех внутриглазных опухолей и легко диагностируется лишь на поздней стадии. Метод основан на косвенном измерении содержания меланина в ресницах. Полученные данные спектрофотометрии, а также общие характеристики обследуемого (пол, возраст и др.) подаются на входные синапсы 43-нейронного классификатора. Нейросеть решает, имеется ли у пациента опухоль, и если да, то определяет ее стадию, выдавая, кроме этого, процентную вероятность своей уверенности (<http://www.chat.ru/~neurocom/>).

- **Авионика:** обучаемые автопилоты, распознавание сигналов радаров, адаптивное пилотирование сильно поврежденного самолета.

☞ Компания McDonnell Douglas Electronic Systems разработала автоматический переключатель режимов полета в реальном масштабе времени в зависимости от вида повреждения самолета. Данные от 20 сенсорных датчиков и сигналов от пилота

используются нейросетью для выработки около 100 аэродинамических параметров полета. Сильной стороной является возможность сети адаптироваться к непредсказуемым аэродинамическим режимам, таким как потеря части крыла и т.д. (SIGNAL Magazin, февраль 1991)

- **Связь:** сжатие видео-информации, быстрое кодирование-декодирование, оптимизация сотовых сетей и схем маршрутизации пакетов.

✉ Нейросети уже продемонстрировали коэффициент сжатия 120:1 для черно-белого видео. Цветное видео допускает примерно вдвое большую степень сжатия 240:1 за счет специальной схемы кодирования цветов. (<http://www.ee.duke.edu/~cec/JPL/paper.html>)

- **Интернет:** ассоциативный поиск информации, электронные секретари и агенты пользователя в сети, фильтрация информации в push-системах, коллаборативная фильтрация, рубрикация новостных лент, адресная реклама, адресный маркетинг для электронной торговли.

✉ Фирма Autonomy отделилась от родительской фирмы Neurodynamics в июне 1996 года с уставным капиталом \$45 млн и идеей продвижения на рынок Internet электронных нейросетевых *агентов*. Согласно ее пресс-релизу, первоначальные вложения окупились уже через год. Компания производит семейство продуктов AGENTWARE, создающих и использующих профили интересов пользователей в виде персональных автономных нейро-агентов. Такие компактные нейро-агенты (<1 Кб) могут представлять пользователя в любом из продуктов компании. Например, агенты могут служить в качестве нейро-секретарей, фильтруя поступающую по информационным каналам информацию. Они также могут постоянно находиться на сервере провайдера, или посылаться для поиска в удаленных базах данных, осуществляя отбор данных на месте. В будущем, когда эта технология получит достаточное распространение, она позволит снизить нагрузку на трафик Сети. (<http://www.agentware.com>)

- **Автоматизация производства:** оптимизация режимов производственного процесса, комплексная диагностика качества продукции (ультразвук, оптика, гамма-излучение, ...), мониторинг и визуализация многомерной диспетчерской информации, предупреждение аварийных ситуаций, робототехника.

✉ Ford Motors Company внедрила у себя нейросистему для диагностики двигателей после неудачных попыток построить экспертную систему, т.к. хотя опытный механик и может диагностировать неисправности он не в состоянии описать алгоритм такого распознавания. На вход нейро-системы подаются данные от 31 датчика. Нейросеть обучалась различным видам неисправностей по 868 примерам. "После полного цикла обучения качество диагностирования неисправностей сетью достигло уровня наших лучших экспертов, и значительно превосходило их в скорости." (Marko K, et. al. , Ford Motors Company, Automotive Control Systems Diagnostics, IJCNN 1989)

- **Политические технологии:** анализ и обобщение социологических опросов, предсказание динамики рейтингов, выявление значимых факторов, объективная кластеризация электората, визуализация социальной динамики населения.

✉ Уже упоминавшаяся ранее группа *НейроКомп* из Красноярска довольно уверенно предсказывает результаты президентских выборов в США на основании анкеты из 12 вопросов. Причем, анализ обученной нейросети позволил выявить пять ключевых вопросов, ответы на которых формируют два главных фактора, определяющих успех президентской кампании. Этот пример будет рассмотрен более подробно в главе, посвященной извлечению знаний с помощью нейросетей.

- **Безопасность и охранные системы:** системы идентификации личности, распознавание голоса, лиц в толпе, распознавание автомобильных номеров, анализ аэро-космических снимков, мониторинг информационных потоков, обнаружение подделок.

✉ Многие банки используют нейросети для обнаружения подделок чеков. Корпорация Nestor (Providence, Rhode Island) установила подобную систему в Mellon Bank, что по оценкам должно сэкономить последнему \$500,000 в год. Нейросеть обнаруживает в 20 раз больше подделок, чем установленная до нее экспертная система. (*Tearing up the Rules*, Banking Technology, ноябрь 1993)

- **Ввод и обработка информации:** Обработка рукописных чеков, распознавание подписей, отпечатков пальцев и голоса. Ввод в компьютер финансовых и налоговых документов.

✉ Разработанные итальянской фирмой RES Informatica нейросетевые пакеты серии FlexRead, используются для распознавания и автоматического ввода рукописных платежных документов и налоговых деклараций. В первом случае они применяются для распознавания не только количества товаров и их стоимости, но также и формата документа. В случае налоговых деклараций распознаются фискальные коды и суммы налогов.

- **Геологоразведка:** анализ сейсмических данных, ассоциативные методики поиска полезных ископаемых, оценка ресурсов месторождений.

✉ Нейросети используются фирмой Амосо для выделения характерных пиков в показаниях сейсмических датчиков. Надежность распознавания пиков - 95% по каждой сейсмо-линии. По сравнению с ручной обработкой скорость анализа данных увеличилась в 8 раз. (J.Veezhinathan & D.Wadner, Amoco, First Break Picking, IJCNN, 1990)

Обилие приведенных выше применений нейросетей - не рекламный трюк. Просто нейросети - это не что иное, как новый инструмент анализа данных. И лучше других им может воспользоваться именно специалист в своей предметной области. Основные трудности на пути еще более широкого распространения нейротехнологий - в неумении широкого круга профессионалов формулировать свои проблемы в терминах, допускающих простое нейросетевое решение. Данный курс призван как раз помочь усвоить типовые постановки задач для нейросетей. Для этого, прежде всего, нужно четко представлять себе основные особенности нейросетевой обработки информации - *парадигмы* нейрокомпьютинга.

Парадигмы нейрокомпьютинга

Что такое парадигмы

Парадигмы нейрокомпьютинга - это родовые черты, объединяющие принципы работы и обучения всех нейрокомпьютеров. Главное, что их объединяет - нацеленность на обработку образов. Эта их особенность, аналогичная способу функционирования мозга, уже обсуждалась ранее во вводной главе. Теперь же мы сформулируем эти парадигмы в концентрированном виде безотносительно к биологическим прототипам, как способы обработки данных. Эти общие сведения послужат фундаментом для более подробного разбора отдельных нейро-архитектур в последующих главах.

Коннекционизм

Отличительной чертой нейросетей является *глобальность связей*. Базовые элементы искусственных нейросетей - *формальные нейроны* - изначально нацелены на работу с широкополосной информацией. Каждый нейрон нейросети, как правило, связан со всеми нейронами предыдущего слоя обработки данных (см. Рисунок 3, иллюстрирующий наиболее широко распространенную в современных приложениях архитектуру многослойного персептрона). В этом основное отличие формальных нейронов от базовых элементов последовательных ЭВМ - *логических вентилях*, имеющих лишь два входа. В итоге, универсальные процессоры имеют сложную архитектуру, основанную на иерархии модулей, каждый из которых выполняет строго определенную функцию. Напротив, архитектура нейросетей проста и универсальна. Специализация связей возникает на этапе их обучения под влиянием конкретных данных.

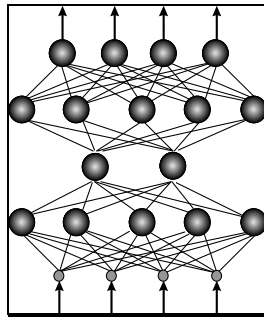


Рисунок 3. Глобальность связей в искусственных нейросетях

Типичный формальный нейрон производит простейшую операцию - взвешивает значения своих входов со своими же локально хранимыми весами и производит над их суммой нелинейное преобразование:

$$y = f(u), \quad u = w_0 + \sum_i w_i x_i$$

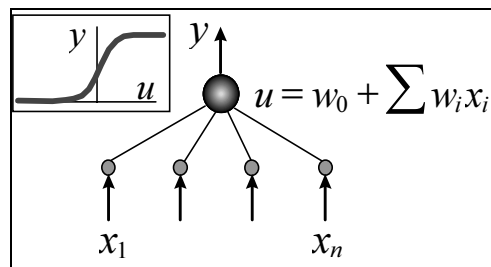


Рисунок 4. Нейрон производит нелинейную операцию над линейной комбинацией входов

Нелинейность выходной *функции активации* $f(\cdot)$ принципиальна. Если бы нейроны были линейными элементами, то любая последовательность нейронов также производила бы линейное преобразование, и вся нейросеть была бы эквивалентна одному нейрону (или одному слою нейронов - в случае нескольких выходов). Нелинейность разрушает линейную суперпозицию и приводит к тому, что возможности нейросети существенно выше возможностей отдельных нейронов.

Локальность и параллелизм вычислений

Массовый параллелизм нейро-вычислений, необходимый для эффективной обработки образов, обеспечивается *локальностью обработки информации* в нейросетях. Каждый нейрон реагирует лишь на локальную информацию, поступающую к нему в данный момент от связанных с ним таких же нейронов, без апелляции к общему плану вычислений, обычной для универсальных ЭВМ. Таким образом, нейросетевые алгоритмы локальны, и нейроны способны функционировать параллельно.

Программирование: обучение, основанное на данных

Отсутствие глобального плана вычислений в нейросетях предполагает и особый характер их программирования. Оно также носит локальный характер: каждый нейрон изменяет свои "подгоночные параметры" - *синаптические веса* - в соответствии с поступающей к нему локальной информацией об эффективности работы всей сети как целого. Режим распространения такой информации по сети и соответствующей ей адаптации нейронов носит характер *обучения*. Такой способ программирования позволяет эффективно учесть специфику требуемого от сети способа обработки данных, ибо алгоритм не задается заранее, а порождается самими данными - *примерами*, на которых сеть обучается. Именно таким образом в процессе самообучения биологические нейросети выработали столь эффективные алгоритмы обработки сенсорной информации.

Характерной особенностью нейросетей является их способность к *обобщению*, позволяющая обучать сеть на ничтожной доле всех возможных ситуаций, с которыми ей, может быть, придется столкнуться в процессе функционирования. В этом их разительное отличие от обычных ЭВМ, программа которых должна заранее предусматривать их поведение во всех возможных ситуациях. Эта же их способность позволяет кардинально удешевить процесс разработки приложений.

Универсальность обучающих алгоритмов

Привлекательной чертой нейрокомпьютинга является единый принцип обучения нейросетей - минимизация эмпирической ошибки. Функция ошибки, оценивающая данную конфигурацию сети, задается извне - в зависимости от того, какую цель преследует обучение. Но далее сеть начинает постепенно модифицировать свою конфигурацию - состояние всех своих синаптических весов - таким образом, чтобы минимизировать эту ошибку. В итоге, в процессе обучения сеть все лучше справляется с возложенной на нее задачей.

Не вдаваясь в математические тонкости, образно этот процесс можно представить себе как поиск минимума функции ошибки $E(\mathbf{w})$, зависящей от набора всех синаптических весов сети \mathbf{w} (см. Рисунок 5).

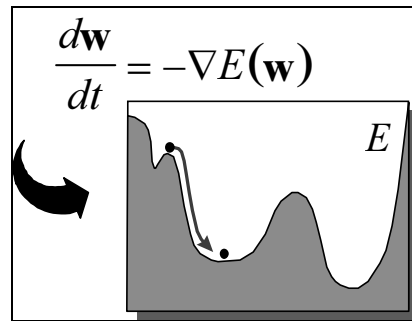


Рисунок 5. Обучение сети как задача оптимизации

Базовой идеей всех алгоритмов обучения является учет локального градиента в пространстве конфигураций для выбора траектории быстрого спуска по функции ошибки. Функция ошибки, однако, может иметь множество локальных минимумов, представляющих суб-оптимальные решения. Поэтому градиентные методы обычно дополняются элементами стохастической оптимизации, чтобы предотвратить застревание конфигурации сети в таких локальных минимумах. Идеальный метод обучения должен найти глобальный оптимум конфигурации сети⁴.

В дальнейшем нам встретится множество конкретных методов обучения сетей с разными конфигурациями межнейронных связей. Чтобы не потеряться за деревьями леса, полезно заранее ознакомиться с базовыми нейро-архитектурами. В следующем разделе мы приведем такого рода классификацию, основанную на способах *кодирования* информации в сетях (обучения) и *декодирования* (обработки) информации нейросетями.

Классификация базовых нейроархитектур

Типы обучения нейросети

Ошибка сети зависит, как уже говорилось, от конфигурации сети - совокупности всех ее синаптических весов. Но эта зависимость не прямая, а опосредованная. Ведь непосредственные значения весов скрыты от внешнего наблюдателя. Для него сеть - своего рода черный ящик, и оценивать ее работу он может лишь основываясь на ее поведении, т.е. на том, каковы значения выходов сети при данных входах. Иными словами, в общем виде функция ошибки имеет вид:

$$E(\mathbf{w}) = E\{\mathbf{x}^\alpha, \mathbf{y}^\alpha, \mathbf{y}(\mathbf{x}^\alpha, \mathbf{w})\}.$$

Здесь $\{\mathbf{x}^\alpha, \mathbf{y}^\alpha\}$ - набор *примеров* (т.е. пар входов-выходов), на которых обучается нейросеть, а $\{\mathbf{y}(\mathbf{x}^\alpha, \mathbf{w})\}$ - реальные значения выходов нейросети, зависящие от конкретных значений ее синаптических весов. Такой способ обучения, когда действительный выход нейросети сравнивают с эталонным, называют *обучением с учителем*.

Иногда выходная информация известна не полностью. Например, вместо эталонных ответов известно лишь хуже или лучше данная конфигурация сети справляется с задачей (вспомним

⁴ Напомним определение идеала как принципиально недостижимой цели, к которой, тем не менее, следует стремиться.

детскую игру “холоднее-горячее” или лабораторную мышь в лабиринте с лакомствами и электрошоком). Этот тип обучения называют *обучением с подкреплением* (*reinforcement learning*).

Вообще говоря, возможен и такой режим обучения, когда желаемые значения выходов вообще неизвестны, и сеть обучается только на наборе входных данных $\{\mathbf{x}^\alpha\}$:

$$E(\mathbf{w}) = E\{\mathbf{x}^\alpha, \mathbf{y}(\mathbf{x}^\alpha, \mathbf{w})\}.$$

Такой режим обучения сети называют *обучением без учителя*. В этом случае сети предлагается самой найти скрытые закономерности в массиве данных. Так, избыточность данных допускает сжатие информации, и сеть можно научить находить наиболее компактное представление таких данных, т.е. произвести оптимальное кодирование данного вида входной информации.

Таблица 5. Сравнение режимов обучения нейросетей

Вид обучения:	С "учителем"	С "подкреплением"	Без "учителя"
Что подается в качестве обучающих примеров	Набор пар входо-выходов $\{\mathbf{x}^\alpha, \mathbf{y}^\alpha\}$	Оценка выходов сети $\{\mathbf{x}^\alpha, \mathbf{y}(\mathbf{x}^\alpha)\}$	Только набор входных значений $\{\mathbf{x}^\alpha\}$
Что требуется от сети	Найти функцию, обобщающую примеры, в случае дискретных \mathbf{y}^α - классифицировать входы. В целом - научиться реагировать схожим образом в схожих ситуациях.	Научиться заданной "правильной" линии поведения.	Найти закономерности в массиве данных, отыскать порождающую данные функцию распределения, найти более компактное описание данных.

С практической точки зрения, "помеченные" данные $\{\mathbf{x}^\alpha, \mathbf{y}^\alpha\}$ зачастую дороги и не столь многочисленны, как "непомеченные" $\{\mathbf{x}^\alpha\}$, например, в случае, когда "учителем" является человек - эксперт. В силу этого обстоятельства на таких данных можно обучить лишь относительно простые и компактные нейросети. Напротив, нейросети, обучаемые без учителя часто используют для переработки больших массивов "сырых" данных - в качестве преобразовывающих фильтров. Указанное различие, однако, исчезает, когда данные естественным образом распадаются на входы-выходы, например при предсказании временных рядов, где следующее значение ряда является выходом, а предыдущие несколько значений - соответствующими входами обучаемой нейросети.

Архитектура связей

На способ обработки информации решающим образом сказывается наличие или отсутствие в сети петель обратных связей. Если обратные связи между нейронами отсутствуют (т.е. сеть имеет структуру последовательных слоев, где каждый нейрон получает информацию только с

предыдущих слоев), обработка информации в сети однонаправленна. Входной сигнал обрабатывается последовательностью слоев и ответ гарантированно получается через число тактов равное числу слоев.

Наличие же обратных связей может сделать динамику нейросети (называемой в этом случае *рекуррентной*) непредсказуемой. В принципе, сеть может "заиклиться" и не выдать ответа никогда. Причем, согласно Тьюрингу, не существует алгоритма, позволяющего для произвольной сети определить придут ли когда-либо ее элементы в состояние равновесия (т.н. *проблема останова*).

Вообще говоря, то, что нейроны в рекуррентных сетях по многу раз принимают участие в обработке информации позволяет таким сетям производить более разнообразную и глубокую обработку информации. Но в этом случае следует принимать специальные меры к тому, чтобы сеть не заикливалась (например, использовать *симметричные* связи, как в сети Хопфилда, или принудительно ограничивать число итераций, как это делается в эмуляторе *MultiNeuron* группы *НейроКомп*).

Таблица 6. Сравнение архитектур связей нейросетей

<i>Сравнение сетей:</i>	<i>Без обратных связей (многослойные)</i>	<i>С обратными связями</i>
<i>Преимущества</i>	Простота реализации. Гарантированное получение ответа после прохождения данных по слоям.	Минимизация размеров сети - нейроны многократно участвуют в обработке данных. Меньший объем сети облегчает процесс обучения.
<i>Недостатки</i>	Требуется большее число нейронов для алгоритмов одного и того же уровня сложности. Следствие - большая сложность обучения.	Требуется специальные условия, гарантирующие сходимость вычислений.

Классификация по типу связей и типу обучения (Encoding-Decoding)

Ограничившись лишь двумя описанными выше факторами, разделяющими сети по типу обучения (программирования) и функционирования, получим следующую полезную классификацию базовых нейро-архитектур, впервые предложенную, по-видимому, Бартом Коско (Таблица 7).

Таблица 7. Классификация нейросетей

<i>Тип обучения (Coding) →</i> <i>Тип связей (Decoding) ↓</i>	С "учителем"	Без "учителя"
Без обратных связей	Многослойные персептроны (аппроксимация функций, классификация)	Соревновательные сети, карты Кохонена (сжатие данных, выделение признаков)
С обратными связями	Рекуррентные аппроксиматоры (предсказание временных рядов, обучение в режиме on-line)	Сеть Хопфилда (ассоциативная память, кластеризация данных, оптимизация)

В этой таблице различные архитектуры сетей, которые встретятся нам далее в этой книге, распределены по ячейкам в зависимости от способа обработки ими информации и способа их обучения. В скобках указаны основные типы задач, обычно решаемых данным классом нейросетей.

В следующих главах, после более близкого знакомства с перечисленными выше основными нейро-архитектурами, мы рассмотрим конкретные примеры задач из области финансов и бизнеса, интересные с практической точки зрения.

Литература

А.Н.Горбань. *Обучение нейронных сетей*. М: СП Параграф, 1990.

А.Н.Горбань, Д.А.Росси́ев. *Нейронные сети на персональном компьютере*. Новосибирск. Наука, Сибирская издательская фирма РАН, 1996.

Ф.В. Широков (1998). "Нейросети на шине VME". Краткая история нейроинформатики. TANA Ltd.,

Anderson, J. A. and Rosenfeld, E. *Neurocomputing: Foundations of Research*. The MIT Press, 1988.

Anderson, J. A., Pellionisz, A. and Rosenfeld, E. *Neurocomputing 2: Directions for Research*. The MIT Press, 1990.

Beltratti A., Margarita S., Terna P. *Neural Networks for Economic and Financial Modeling*. ITCP, 1995.

Bishop C.M. *Neural Networks and Pattern Recognition*. Oxford Press, 1995.

Haykin, S. *Neural Networks, a Comprehensive Foundation*. Macmillan, 1994.

Hecht-Nielsen, R. *Neurocomputing*. Addison-Wesley, 1990.

Hopfield, J.J. (1982). "Neural networks and physical systems with emergent collective computational abilities", *Proc. Natl. Acad. Sci.*, **79**, 2554-2558.

Hopfield, J.J. (1984). "Neurons with graded response have collective computational properties like those of two-state neurons", *Proc. Natl. Acad. Sci.*, **81**, 3088-3092.

McCulloch, W.S., Pitts, W. (1943). "A logical calculus of the ideas immanent in nervous activity", *Bulletin of Math. Bio.*, **5**, 115-133.

Maren, A., Harston, C. and Pap, R. *Handbook of Neural Computing Applications*. Academic Press, 1990

Mead, C (1989). *Analog VLSI and neural systems*. Addison Wesley.

Minsky, M. and Papert, S. (1969). *Perceptrons*. MIT Press, Cambridge MA.

M. Muller, B. and Reinhardt, J. *Neural Networks, An Introduction*. Springer-Verlag. 1990. Rosenblatt, F. (1969). *Principles of Neurodynamics*. Spartan Books, Washington DC, 1961.

Murray A.F. *Applications of Neural Nets*. Kluwer Academic Publishers, 1995.

Pao, Y. H. *Adaptive Pattern Recognition and Neural Networks*. Addison-Wesley, 1989.

Rumelhart, D.E., Hinton, G.E., and Williams, R.J. (1986). "Learning internal representations by error propagation", in: McClelland, J. L. and Rumelhart, D. E. (Eds.). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Volume 1, 318-362. MIT Press, Cambridge MA.