«ФИЗИКА ЭЛЕМЕНТАРНЫХ ЧАСТИЦ И АТОМНОГО ЯДРА»
1993, ТОМ 24, ВЫП.6

УДК 681.322

ПРИМЕНЕНИЕ НЕЙРОННЫХ СЕТЕЙ В ЭКСПЕРИМЕНТАЛЬНОЙ ФИЗИКЕ

И.В.Кисель, В.Н.Нескоромный, Г.А.Ососков

Объединенный институт ядерных исследований, Дубна

Изложены теоретические основы различных моделей искусственных нейронных сетей (ИНС) и их применения к актуальным задачам ассоциативной памяти, оптимизации и распознавания. Дан обзор многочисленных приложений ИНС в экспериментальной физике как для аппаратной реализации триггерных средств быстрого отбора событий, так и для последующего программного решения задачи распознавания данных траекторных измерений.

The theoretical foundations of numerous models of artificial neural networks (ANN) and their applications to the actual problems of associative memory, optimization and pattern recognition are given. This review contains also numerous using of ANN in the experimental physics both as the hardware realization of fast triggering systems for event selection and for the following software implementation of the trajectory data recognition.

1.ВВЕДЕНИЕ

Возникновение и бурное развитие в последние 10—15 лет различных теорий искусственных нейронных сетей (ИНС) явилось проявлением и попыткой преодоления драматического разрыва между огромным фактическим материалом, относящимся к биологическим механизмам работы мозга, накопленным в нейрофизиологии в конце XIX — начале XX веков, и неадекватностью имевшегося математического аппарата и вычислительных средств его технической реализации. Главное достоинство и преимущество способностей мозга выполнять логические, распознающие и вычислительные функции — их принципиальная параллельность, нелинейность и нелокальность — не согласовывались с довлеющим принципом последовательных вычислений, с ориентацией математического аппарата на локальность, линейность и стационарность описаний.

Вряд ли можно с полной определенностью утверждать, что этот разрыв заполнен или хотя бы успешно заполняется сейчас. В настоящем обзоре мы, собственно, и не претендуем на какое-то описание уже накопленных наукой теорий работы мозга. Тем не менее некоторые математические модели функционирования нейронов в их коллективном взаи-

модействии, возникшие в процессе этого накопления, оказались чрезвычайно эффективным инструментом решения самых актуальных проблем, если не в биологии, то в самой математике и ее многочисленных приложениях.

В число этих проблем входит ряд задач, решение которых осложняется именно нелинейностью, нелокальностью, дискретностью и часто нестационарностью постановки. Сюда, в частности, относятся задачи распознавания образов, конструирования ассоциативной памяти (т.е. запоминающих устройств для быстрого обмена не по адресу, а по значению фрагмента запоминаемой или извлекаемой величины) и оптимизации (т.е. поиска максимума функционала при наличии ограничений на его параметры).

Таким образом, термин «нейрон» в данной работе не следует понимать буквально, и, несмотря на биологическое происхождение, мы специально подчеркиваем в нашей терминологии искусственность описываемых нейронных сетей. По существу теория ИНС является частью общей теории динамических систем, в которой особое внимание уделяется исследованию сложного коллективного поведения совокупности очень большого числа сравнительно простых логических объектов, называемых нейронами.

Существенно, что, хотя для решения указанного класса задач применяются до сих пор обычные цифровые последовательные ЭВМ, теория нейронных сетей стимулировала возникновение как совершенно новых алгоритмов их решения, так и соответствующих специализированных параллельных компьютеров, приспособленных для наилучшей технической реализации этих алгоритмов, в том числе аналоговых быстродействующих нейрокомпьютеров на базе новейших достижений оптики.

Материал настоящей статьи организован следующим образом. В разд.2 описываются биологические принципы, лежащие в основе большинства моделей, использующих алгоритмы и методы нейронных сетей. Строится простейшая нейронная сеть, в процессе символической «эволюции» которой минимизируется некоторый функционал, называемый обычно энергией нейронной сети. Его экстремумы отвечают искомой конфигурации нейронной сети, совпадающей либо с распознаваемым образом, либо, в зависимости от решаемой задачи, с оптимальным удовлетворением определенных априорных требований. Аналогия бинарных нейронов с хорошо разработанной теоретической моделью магнетика Изинга приводит к глауберовой динамике для решения задачи нахождения глобального минимума энергетической функции и непопадания в ее локальные минимумы. При этом используется теория среднего поля с ненулевой температурой, что дает возможность туннелирования в глобальный минимум спиновой конфигурации.

В третьем разделе приведены постановки и дается обзор решений ряда актуальных проблем, использующих нейронные сети: задачи ассоциативной памяти, задачи оптимизации и две задачи распознавания — с обучением и без. Последние две задачи разбираются более детально, так как результат их решения особенно широко применяется в физике высоких энергий.

В разд.4 описаны различные способы реализации нейронных сетей: виртуальный (в виде алгоритма на компьютере), электрический и оптический (в виде физической схемы). В зависимости от класса решаемых задач описываются также различные типы архитектуры нейронных сетей, содержащие множество связанных друг с другом функциональных уровней и проявляющие некоторые зачатки искусственного интеллекта.

В пятом разделе рассматриваются возможные варианты применения нейронных сетей на примерах проблем, возникающих на трех основных этапах обработки экспериментальных данных в физике высоких энергий (on-line детектирование, off-line обработка для получения физических параметров и последующая интерпретация с проверкой гипотез).

Особое внимание уделено постановке и методам решения задач распознавания треков, событий и проверки статистических гипотез. На примерах работ, ведущихся в ОИЯИ, дается детальный разбор вычислительных трудностей, возникающих как при начальном конструировании нейронной сети, так и при поиске глобального минимума ее энергетического функционала. Приводятся различные методики ускорения сходимости нейронных сетей.

2. БИОЛОГИЧЕСКИЕ ПРЕДПОСЫЛКИ И ПРОСТЕЙШИЕ МОДЕЛИ

Нейробиологические истоки нейронных сетей. Первые нейронные клетки мозжечка были открыты в 1836 г. нейрофизиологом Я.Пуркинье. Через 70 лет Нобелевская премия по физиологии была присуждена К.Гольджи и С.Рамон-и-Кахалю за исследования структуры нейронов и нервной системы человека и позвоночных. Само слово «нейрон» было введено в научный обиход В.Вальдейром (1836—1921). Понятие «синапс» как преобразователь сигналов, поступающих в нейрон, ввел в 1897 г. Ч.С.Шеррингтон*.

Так появилась *нейронная доктрина*, охватывающая как единый биологический принцип нервные процессы всех организмов от самых

*Подробнее с историей нейробиологии можно ознакомиться в статье Д.Хьюбела [1].

простых до имеющих центральную нервную систему, включая и человека. Эта доктрина рассматривает нервную систему как структуру, состоящую из множества параллельно и связанно работающих элементов — нейронов. Отдельный нейрон — это обрабатывающая единица, клетка, состоящая из «сомы» — тела нейрона, к которому через отростки — «дендриты» подходят «аксоны» — многочисленные линии передач от других нейронов. Стыковочная часть нейрона — «синапс» осуществляет преобразование входной информации в сигналы, воспринимаемые нейроном и переводящие его в одно из двух устойчивых состояний — возбуждения или торможения.

Теории биохимических молекулярных процессов, происходящих в нервной системе, биологических мембран, осуществляющих передачу сигналов, были развиты много позже.

Главное принципиальное положение нейронной доктрины состоит в том, что нейроны рассматриваются как неотъемлемые элементы высокопараллельной нервной системы, процессы обработки информации в которой несводимы к явлениям, происходящим в отдельных нейронах.

Именно поэтому, в частности, возникла потребность в мультидисциплинарном подходе к изучению и описанию таких процессов. Теории моделей мозга могли развиваться только на стыке многих наук: биологии, физики, химии, математики и микроэлектроники. Наше дальнейшее изложение в рамках выбранного направления также можно рассматривать как иллюстрацию этого положения. От нейрофизиологических принципов и описаний мы переходим к простым моделям, оказывающимся аналогом неупорядоченных магнитных систем, так называемых спиновых стекол. Их теория является частью статистической физики, которая подсказала также и другие подходы к решению важнейших проблем ИНС, таких как применение теории среднего поля и имитационного отжига. Клеточные автоматы, примыкающие к ИНС, также оказались эффективным средством решения проблемы выбора начальных состояний ИНС. Еще больше междисциплинарных рассмотрений возникает при изучении проблем разработки нейрокомпьютеров.

Завершая перечисление биологических предпосылок ИНС, укажем на еще два важных исследования, в значительной степени опередивших развитие теории нейронных сетей. Первая — это работа Мак-Каллока и Питтса [2], в которой система биологических нейронов моделируется набором бинарных (т.е. имеющих только два состояния) объектов, связанных друг с другом и имеющих некоторый пороговый уровень возбуждения, при достижении которого состояние нейрона изменяется. Итак, каждый i-й нейрон имеет два состояния, характеризующихся тем, что выход нейрона V_i принимает одно из значений: V_i^0 или V_i^1, которые часто равны

просто 0 или 1. На практике выходом реального нейрона является частотно-модулированный сигнал, так что разным состояниям отвечает разная частота сигнала. Мы же в дальнейшем будем считать, что на выходе нейрона образуется некоторый потенциал V_i, поскольку реальный сигнал в дальнейшем все равно подвергается частотной демодуляции. Выходной сигнал нейрона через аксон подается на синапс и через синаптическую связь (которая отличается для разных пар нейронов) подается через входные отростки нейрона (дендриты) на вход j-го нейрона. Образуется глобальная петля обратной связи, в связи с чем система нейронов обладает нетривиальным нелинейным поведением.

Входной сигнал каждого нейрона состоит из двух составляющих: внешнего по отношению к нейронной сети сигнала I_i и взвешенного с синаптическими весами T_{ij} сигнала, поступающего от остальных нейронов. Общий входной сигнал, таким образом, имеет вид

$$H_i = \sum_{j \neq i}^{N} T_{ij} V_j + I_i. \tag{1}$$

Эволюция нейронной сети происходит по следующим правилам. Каждый нейрон время от времени оценивает свой входной сигнал по отношению к пороговому уровню U_i, после чего либо оставляет свой выходной сигнал без изменений, либо изменяет его в соответствии со стохастическим алгоритмом:

$$V_i \to \begin{cases} V_i^0, \text{ если } \sum\limits_{j \neq i} V_{ij} V_j + I_i < U_i \\ V_i^1, \text{ если } \sum\limits_{J \neq i} T_{ij} V_j + I_i > U_i. \end{cases} \tag{2}$$

Таким образом, алгоритм в модели Мак-Каллока — Питтса является асинхронным. Можно указать по меньшей мере два существенных отличия реальной нейронной сети от модельной. Во-первых, реальный нейрон должен обладать непрерывно меняющимся выходным сигналом в зависимости от входного сигнала. Во-вторых, распространение любого сигнала по сети из-за емкостей происходит с запаздыванием, так что распространение сигнала должно описываться не дискретным алгоритмом, а дифференциальным уравнением с источником шума из-за детектирования. Поэтому более принятой формой описания эволюции сети является нелинейная функция не ступенчатого, а «сигмоидного» типа:

$$V_i = g(H_i), \tag{3}$$

где $g(t)$ изображена на рис.1. Схематическое изображение такого нейрона дано на рис.2.

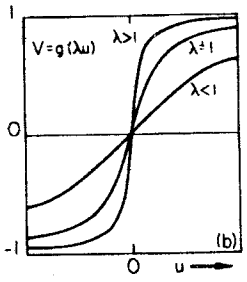
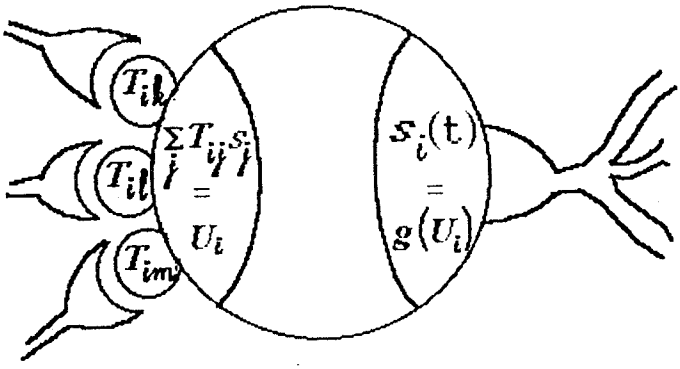


Рис.1. Общий вид сигмоидной функции в зависимости от параметра λ

Рис.2. Схематическое изображение нейрона



Вторая работа, на которую мы хотим обратить внимание, нейропсихолога Д.О.Хебба [3] заложила основы синаптической теории памяти и обучения, т.е. понимания свойств нейронных сетей, которые и обуславливают многообразие их возможностей, в том числе и возникновение интеллекта. Хебб сформулировал свое правило обучения следующим образом: «Если аксон клетки *A* достаточно близок к клетке *B*, чтобы возбуждать ее, и если он повторно или настойчиво принимает участие в ее срабатывании, то в одной из этих клеток или в обеих происходит некоторый процесс метаболического изменения, так что эффективность клетки *A* как одной из клеток, вызывающих срабатывание клетки *B*, увеличивается».

Правило Хебба, меняя синаптические связи, обеспечивает тренировку сети для работы в качестве ассоциатора (классификатора) образов. Многократное предъявление на вход стимулирующего образа должно побуждать сеть генерировать другой образ (признак), ассоциируемый с первым. И наоборот, предъявление сети фрагмента образа в соответствии с правилом Хебба вызывает генерацию полного образа. По сути, это — основная идея адресации по содержанию и ассоциативной памяти [4].

Математические модели ИНС. Поскольку Хебб, будучи нейропсихологом, не пользовался математикой и его книга не содержит ни одной формулы, математическое выражение этого правила было предложено Саттоном [5]:

$$T_{ij}^{(m+1)} = T_{ij}^{(m)} + \eta V_j^{(m)} H_i^{(m)}, \tag{4}$$

где m — дискретное время обучения, т.е. номер тренировочной итерации, η — положительная константа, определяющая скорость обучения, а V_j — один из входных сигналов, дающих на выходе H_i.

В книге [6] правило Хебба (4) рассматривается как специальный случай более общего правила обучения, в котором синаптический вес изменяется пропорционально усилительному сигналу $r_{ij}^{(m)}$:

$$T_{ij}^{(m+1)} = T_{ij}^{(m)} + \eta r_{ij}^{(m)}. \tag{5}$$

Для правила Хебба усилительный сигнал имеет вид $r_{ij}^{(m)} \equiv V_j^{(m)} H_i^{(m)}$. Другую форму усилительного сигнала предложили Видроу и Хофф [7] для так называемого обучения с учителем (супервизором):

$$r_{ij}^{(m)} = [z_i^{(m)} - H_i^{(m)}] V_j^{(m)}, \tag{6}$$

где $z_i^{(m)}$ — специальный обучающий эталонный сигнал, $H_i^{(m)} = \sum_{j=1}^{n} T_{ij}^{(m)} V_j^{(m)}$. Обычно обозначают

$$\Delta T_{ij}^{(m)} = T_{ij}^{(m+1)} - T_{ij}^{(m)} = \eta r_{ij}^{(m)}, \tag{7}$$

и правило обучения называют дельта-правилом. Оно обеспечивает сходимость весов при обучении.

Концепция обучаемых нейронных сетей, называемых персептронами, была введена Ф.Розенблаттом [8,9]. Персептрон, как показано на рис.3, является прямоточной ИНС из одного или нескольких слоев нейронов, которые соединены связями между слоями, а также с входами и выходами. Слои между входным и выходным слоями называются скрытыми. Сигналы, поступающие на вход каждого из нейронов, преобразуются нелинейно согласно (1), а также (2) либо (3). Согласно (1) и (3) персептрон разделяет все N-мерное пространство входных переменных $\{x_i\}$ на K классов посредством гиперплоскостей, определяемых уравнениями

$$\sum_{j \neq i} T_{ij} x_j - I_i = 0, \quad i = 1,...,K.$$

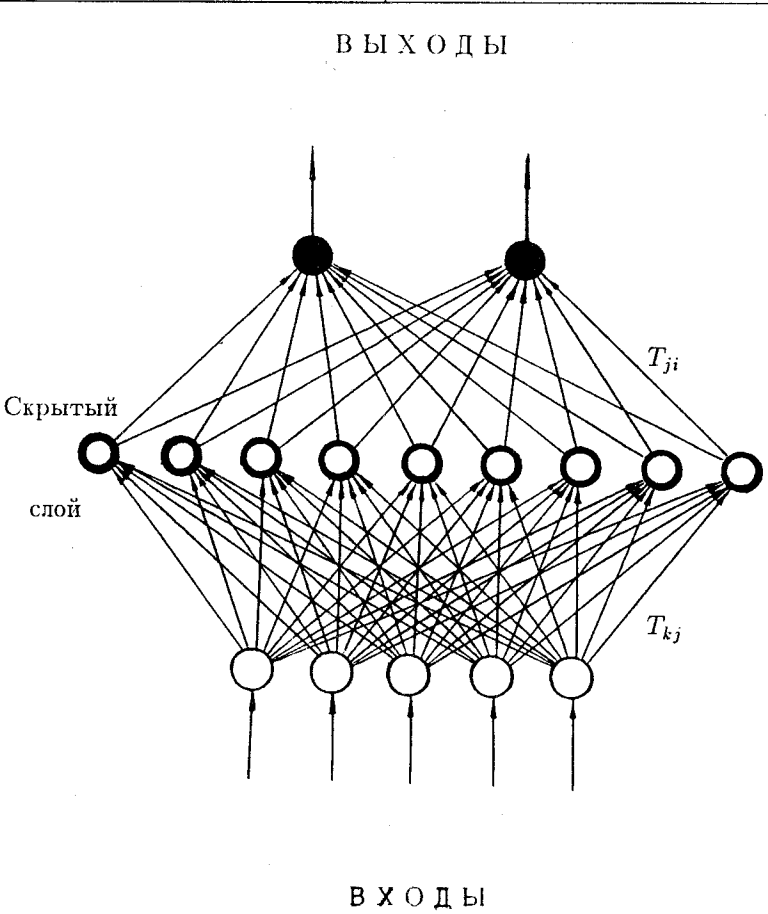


Рис.3. Схема трехслойного персептрона

Может существовать не более 2^K таких классов. Веса связей и пороги изменяются при обучении, а затем фиксируются для использования персептрона как классификатора.

При обучении по Δ-правилу на выборке из M тренировочных циклов обычно используется алгоритм *обратного распространения ошибок*, который является обобщением метода наименьших квадратов применительно к многослойным персептронам, т.к. при этом минимизируется по всем значениям весов функционал из суммы квадратов невязок между

выходным сигналом персептрона $\{y_i^{(m)}\}$ в каждом из циклов и эталонным образцом $\{z_i^{(m)}\}$, предъявляемым в этом цикле:

$$E = \sum_{m=1}^{M} \sum_{i} (y_i^{(m)} - z_i^{(m)})^2 \to \min. \tag{8}$$

Например, для трехслойного персептрона имеем N входных сигналов $\mathbf{X}_N = (x_1,...,x_N)$, скрытый слой n нейронов $\mathbf{H}_n = (h_1,...,h_n)$ и l выходных сигналов. Обозначая T_{ij} веса, связывающие входные сигналы со скрытыми нейронами, и T_{jk} — связи последних с выходными сигналами персептрона, получим согласно (1), (3)

$$h_j = g(H_j), \quad H_j = \sum_k T_{jk}\, x_k;$$

$$y_i = g(Y_i), \quad Y_i = \sum_j T_{ij}\, h_j. \tag{9}$$

Дифференцируя (8) по T_{ij} и T_{jk} и приравнивая нулю производные, получим систему линейных уравнений, из которой находим:

$$\Delta T_{ij}^{(m+1)} = -\eta (y_i^{(m)} - z_i^{(m)})\, g'(y_i) h_j, \tag{10}$$

$$\Delta T_{ik}^{(m+1)} = -\eta \sum_i T_{ij}^{(m)} g'(y_i^{(m)}) g'(h_j^{(m)}) x_k, \tag{11}$$

что фактически соответствует Δ-правилу.

Для простого случая $M = 1$, когда происходит классификация входов только на два класса, Розенблатт показал [9], что если входные векторы, принадлежащие разным классам по их близости эталонам, разделены в пространстве входов некоторой гиперплоскостью (гипотеза компактности), то указанный алгоритм обучения сходится. К сожалению, такие простые трехслойные персептроны не позволяют строить более сложные разделяющие поверхности в пространстве входов. Введение дополнительных скрытых слоев позволяет формировать более сложные выпуклые разделяющие поверхности, но в то же время значительно усложняет обучение.

Модель Хопфилда. Поистине «взрывной» характер приобрели публикации по нейронным сетям после работ Хопфилда [10—12], в которых указанные вопросы были решены на основе построения функционала «энергии» для нейронной сети, являющегося для динамики исходной нейронной сети не чем иным, как функцией Ляпунова [13—16]. Доказательство существования стационарной стабильной точки для алгоритма (2)

было проведено для произвольной симметричной матрицы синаптических весов T_{ij}. Действительно, оказалось достаточным построить функцию

$$E = -\frac{1}{2}\sum_{i \neq j}^{N}\sum_{j=1}^{N} T_{ij}V_iV_j - \sum_{i=1}^{N} I_iV_i + \sum_{i=1}^{N} U_iV_i. \qquad (12)$$

Изменение E при изменении состояния i-го нейрона на ΔV_i, очевидно, равно

$$\Delta E = -\left(\sum_{j \neq i} T_{ij}V_j + I_i - U_i\right)\Delta V_i.$$

Но положительной скобке отвечает положительное ΔV_i, и наоборот, следовательно, $\Delta E \leq 0$ в процессе эволюции и, поскольку E ограничена, существует предельное стационарное состояние.

В реальной нейронной сети матрица T_{ij} вовсе не обязана быть симметричной в силу топологии связей между нейронами. Действительно, i-й нейрон получает сигнал от j-го через один синапс, а j-й от i-го — через другой. Динамика нейронной сети с несимметричной матрицей синаптических весов исследовалась в работах [17,18]. При этом в процессе эволюции могли меняться и сами T_{ij}. В результате, на основе [19] было получено, что условие симметричности во многих случаях несущественно, хотя при его нарушении затруднительно интерпретировать E как энергию нейронной сети. Кроме того, при несимметричных весах в системе возможно возникновение предельных циклов и осцилляций, подробно исследованных в [20—22].

Интересным оказывается вопрос о предельной точке нейронной сети, состоящей из нейронов с непрерывной функцией отклика (сигмоидной функцией), изображенной на рис.1 в зависимости от параметра, управляющего степенью отличия этой непрерывной функции от ступенчатой [11]. Для такой нейронной сети тоже возможно построение функции E, монотонно убывающей на решениях дифференциального уравнения, определяющего непрерывную динамику нейронов. Доказано, что при стремлении сигмоидной функции к ступенчатой стационарная точка системы непрерывных нейронов с заданной матрицей T_{ij} стремится к стационарной точке дискретных нейронов с той же матрицей связей. Это связано с тем, что для непрерывных V_i первое слагаемое в формуле (12) является определяющим для минимума (при отсутствии I_i), а оно линейно для каждого V_i, что дает положение минимума в углах гиперкуба

$-1 \le V_i \le 1$. Для очень пологой сигмоидной функции, напротив, единственной стабильной точкой является тривиальное решение $V_i = 0$.

Рассмотрим теперь, как в модели Хопфилда «запоминается» конкретный образ. Иными словами, пусть задана некоторая фиксированная конфигурация нейронной сети. Требуется так подобрать синаптические веса T_{ij}, чтобы для любой начальной конфигурации, расположенной достаточно близко к требуемому образу, в результате динамики (2) в пределе получалась требуемая конфигурация нейронов. Здесь отметим, что динамика (2) модели с функцией (12) является не чем иным, как глауберовой динамикой изинговского магнетика с дальнодействием при нулевой температуре. Для этого необходимо заменить нейроны на классические спины $S_i = \pm 1$, тогда энергия будет иметь следующий вид (здесь и далее порог возбуждения принят равным нулю):

$$E = -\frac{1}{2}\sum_{i,j=1}^{N} T_{ij} S_i S_j. \tag{13}$$

Здесь суммирование идет по всем спинам, при этом принято, что $T_{ii} = 0$.

Проблема запоминания конкретного образа была решена в работе [23]. Рассмотрим ее на примере одного образа. Пусть требуемый образ представляет собой последовательность $\{\xi_i\}$ ($\xi_i = \pm 1$). Очевидно, что квадратичная форма

$$E = -\frac{1}{N}\sum_{i,j=1}^{N} (\xi_i S_i)(\xi_j S_j)$$

будет иметь минимум именнно при $S_i = \xi_i$ (или при $S_i = -\xi_i$, что тоже решает задачу). Тогда для матрицы синаптических весов получим выражение

$$T_{ij} = \frac{1}{N}\xi_i \xi_j,$$

что решает задачу запоминания одного образа. Множитель $1/N$ введен для того, чтобы при больших N энергия имела необходимую для термодинамики пропорциональность числу спинов. Для нескольких образов $\{\xi_i^p\}$ ($i = 1,...,N$, $p = 1,...,M$) при не очень большом M решением будет следующий магнетик Маттиса:

$$T_{ij} = \frac{1}{N}\sum_{p=1}^{M} \xi_i^p \xi_j^p. \tag{14}$$

Интересно, что будучи построенным чисто формально, на основе только математических соображений, выражение (14) имеет также глубокую

биологическую основу. Если при запоминании образа i-й нейрон посылает импульс на j-й нейрон, то соответствующая этой паре нейронов синаптическая связь усиливается, и сигналы между этими нейронами проходят с большим весом, что приводит к вышеупомянутому правилу Хебба [3].

Известно, что в термодинамическом пределе $N \to \infty$ энергетическая функция магнетика с взаимодействием (14) будет иметь ровно $2M$ локальных минимумов при условии некоррелированности образов, т.е. когда

$$\frac{1}{N}\sum_{i=1}^{N} \xi_i^p \xi_i^q = O(N^{-1/2}) \quad \text{при } p \neq q.$$

Оказывается, что в термодинамическом пределе возможно также увеличение числа образов, записанных в нейронную сеть, так что число образов будет пропорционально N:

$$M = \alpha N. \tag{15}$$

Исследование максимально возможного коэффициента α, при котором еще происходит восстановление большей части образов, было осуществлено Хопфилдом [10]. Получено, что при $0{,}14 < \alpha < 0{,}16$ нейронная сеть быстро насыщается, при этом возникают ложные минимумы от перекрытия разных образов, а часть минимумов исчезает. Численный эксперимент был проведен с нейронной сетью, содержащей $N = 100$ нейронов. При записи в нейронную сеть $M = 0{,}15N$ образов восстановление нужного образа происходило с вероятностью 0,85. В 10% случаев система приходила в минимум, не отвечающий ни одному образу. При дальнейшем увеличении числа образов до $\alpha = 1$ модель нейронной сети оказывается идентичной модели спинового стекла [24]. На рис.4 приведен пример восстановления образа, представляющего букву A, записанного в нейронную сеть из $N = 20{\times}20$ спинов наряду с другими случайными образами общим числом $M = 30$. Начальное состояние нейронной сети выбиралось

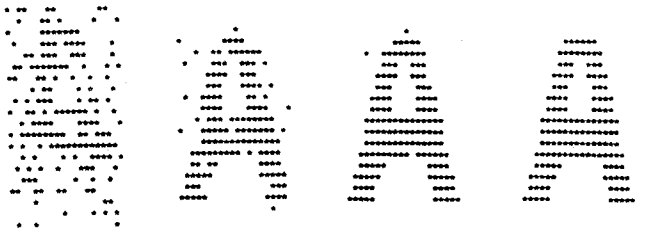


Рис.4. Восстановление образа, представляющего букву A, записанного в нейронную сеть из $N = 20{\times}20$ спинов наряду с другими $M = 30$ случайными образами [24]

в виде той же буквы A, в которой с вероятностью $p = 0{,}3$ спины были изменены на противоположные. Нейронной сети оказалось достаточно четырех итераций, чтобы практически без искажений «вспомнить» исходный образ.

Указанное свойство нейронной сети восстанавливать предъявленный слегка измененный образ (либо с наложенным шумом, либо содержащий лишь важную часть исходного образа) называется *ассоциативной памятью*. Это позволяет на основе нейронной сети создавать базы данных, поиск в которых осуществляется не по каталогу, а в процессе вычислений на основе предъявленного фрагмента требуемого образа. Все вышесказанное было справедливо для некоррелированных образов. Если же образы имеют общие характерные черты (например, при записи в нейронную сеть еще нескольких букв алфавита), то минимумы на энергетической поверхности, отвечающие похожим буквам, сливаются, и нейронная сеть выдает неверный результат. На рис.5 показано, как в ту же нейронную сеть записали всего 5 букв (левая колонка), и что вышло после работы нейронной сети (правая колонка). Видно, что нейронная сеть восстановила только совпадающие фрагменты изображений.

Выход из такой ситуации оказался довольно неожиданным. Для этого вспомним, что информация в нейронной сети хранится не в ее элементах (нейронах), а в связях между ними (т.е. в матрице синаптических весов). Общее число различных элементов в симметричной матрице равно $N(N-1)/2$. Иными словами, даже если $M = 0{,}15N$, в нейронной сети содержится много избыточной информации об образах, что повышает надежность нейронной сети даже при исключении большого числа связей между нейронами. В случае коррелированных образов оказалось достаточным исключить «испорченные» связи. Последние определяются следующим образом. По формуле (14) вычисляются все синаптические веса T_{ij}. Если хотя бы для одного образа знак соответствующего слагаемого в сумме не совпадает со знаком всей суммы, то эта связь между нейронами исключается: $T_{ij} = 0$. На рис.6 продемонстрирован результат работы такой модернизированной нейронной сети, содержащей изображения пяти букв латинского алфавита. Если до исключения нежелательных элементов из матрицы синаптических весов нейронная сеть отказывалась распознавать даже первоначальные «чистые» изображения букв, то после процедуры исключения, когда число оставшихся связей составило лишь 16% от из первоначального числа, нейронная сеть неплохо восстанавливает даже зашумленные с уровнем шума 20% (левая колонка) изображения букв (правая колонка рис.6).

Проблема увеличения емкости нейронной сети очень важна. Результат Хопфилда, естественно, справедлив только для простейшей одно-

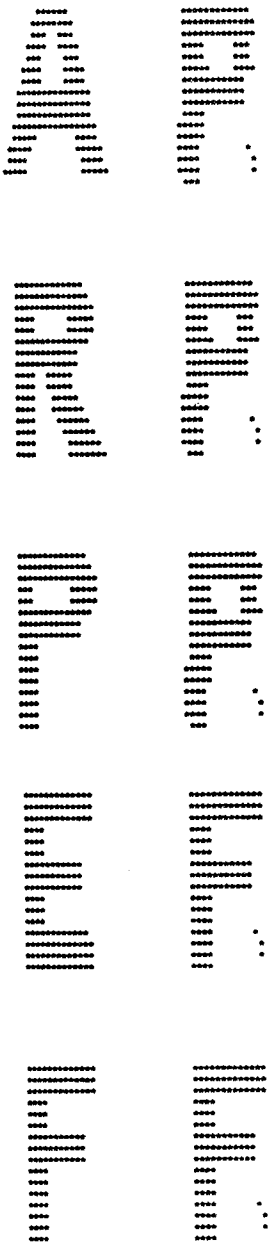


Рис.5. Распознавание коррелированных образов [24]

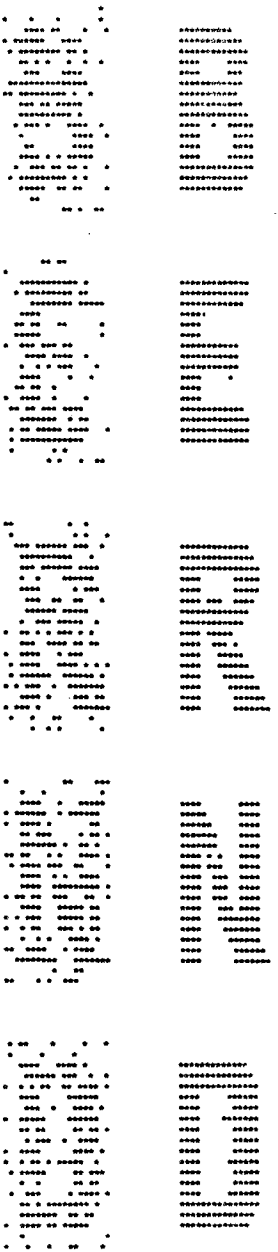


Рис.6. Распознавание коррелированных образов с учетом исключения испорченных связей [24]

слойной нейронной сети. Существенное увеличение емкости нейронной сети за счет введения скрытого слоя продемонстрировано в работе [25]. В работе [26], как и в случае с запоминанием букв, улучшение характеристик нейронной сети также достигнуто за счет ее упрощения. Увеличение емкости нейронной сети произошло при сужении диапазона возможных значений T_{ij}. Если у Хопфилда $T_{ij} \in [-M,M]$, то здесь веса принимают только три возможных значения: −1, 0 и +1. Это дает также дополнительные преимущества по экономии места при реализации нейронной сети в виде электронного чипа.

Даже в стандартной нейронной сети возможно увеличение числа хранимых образов. Действительно, пусть нам необходимо запомнить, к примеру, 1 млн. некоррелированных образов. Тогда простейшая нейронная сеть должна содержать $N \approx 6{,}7$ млн. нейронов. В таком случае каждый образ может содержать более 6 Мбит информации. Но зачастую в образе информации много меньше, т.е. он содержит только X бит информации, большая часть из которых к тому же нули, а единиц в образе только $K \sim \sqrt{X}$. Тогда для записи M таких образов требуется нейронная сеть лишь из $N = 2\sqrt{MX}$ элементов. По сравнению с организацией обычного каталога происходит существенный выигрыш во времени контекстно-адресного поиска в нейронной сети: требуется всего $KN \approx 2X\sqrt{M}$ операций вместо $\sim M$ для обычного перебора изображений [27].

Заметим, что монотонное убывание функции энергии нейронной сети первоначально было получено для последовательной динамики нейронной сети, т.е. когда в каждый момент времени только один нейрон (по очереди либо выбранный случайным образом) оценивает свое состояние и затем изменяет его. Очень важный результат был получен в работе [28], где утверждается, что различия последовательной динамики и параллельной (когда все нейроны эволюционируют одновременно), с точки зрения получения стационарного состояния, несущественны. Это открывает дорогу к широкому использованию параллельных компьютерных систем для моделирования нейронных сетей.

В этом разделе основное внимание было уделено проблеме ассоциативной памяти и задаче распознавания образов. Как было показано, в таких задачах сначала происходит «тренировка» связей, т.е. обучение ИНС, при котором модифицируется матрица синаптических весов. Эта операция потребляет достаточно много машинных ресурсов, поэтому итоговая матрица синаптических весов может составлять предмет авторского права наряду со стандартными базами данных. После этого алгоритмы нейронных сетей позволяют достаточно экономно найти ближайший к начальной конфигурации локальный минимум энергетической функции.

Существует, однако, обширный круг задач оптимизации, в которых необходимо найти не локальный, а глобальный минимум для специальным образом выбранного функционала, имеющего вид энергетической функции (12). В этих задачах соответственно требуемым условиям сначала необходимо выбрать вид матрицы T_{ij}.

3. ЗАДАЧИ ОПТИМИЗАЦИИ

Теория среднего поля в модели Хопфилда. Итак, рассмотрим задачу нахождения глобального минимума системы классических спинов, энергия которых определяется выражением

$$E = -\frac{1}{2}\sum_{i,j=1} T_{ij} S_i S_j .$$

Для нахождения глобального минимума методами нейронных сетей необходимо, чтобы система могла перейти от одного минимума к другому, более низкому. Это возможно, если в модель ввести «температуру». После этого появляются две возможности. Во-первых, подобно методу Монте-Карло, можно допустить в системе наличие флуктуаций энергии с тем, чтобы она выходила за пределы локального минимума. Во-вторых, можно разрешить системе иметь «подбарьерные» конфигурации с тем, чтобы система «туннелировала» в соседний, более низкий минимум. Выбор между этими альтернативами, над- и подбарьерными переходами зависит от соотношения высоты барьера и его ширины. Для нейронной сети с сильно нерегулярными константами T_{ij} энергетическая поверхность похожа на энергетическую поверхность модели спинового стекла, а для последней известно, что в модели существуют очень высокие, но достаточно узкие потенциальные барьеры между локальными минимумами [29].

Таким образом, мы приходим к выводу о предпочтительности туннелирования сквозь барьеры. Математически это можно реализовать, используя теорию среднего поля. Суть ее состоит в следующей выкладке:

$$0 \approx -\frac{1}{2}\sum_{i,j=1}^{N} T_{ij}(S_i - \langle S_i\rangle_T)(S_j - \langle S_j\rangle_T) =$$

$$= -\frac{1}{2}\sum_{i,j=1}^{N} T_{ij} S_i S_j + \sum_{i,j=1}^{N} T_{ij} S_i \langle S_j\rangle_T - \frac{1}{2}\sum_{i,j=1}^{N} T_{ij}\langle S_i\rangle_T \langle S_j\rangle_T ,$$

т.е. энергию приближенно можно записать как линейное выражение плюс константа:

$$E = -\frac{1}{2}\sum_{i,j=1}^{N} T_{ij} S_i S_j \approx -\sum_{i,j=1}^{N} T_{ij} S_i \langle S_j \rangle_T + \text{const}, \qquad (16)$$

где через $\langle S_i \rangle_T$ обозначены температурные средние значения спинов S_i.

Существует обширная литература, касающаяся математически строгого обоснования корректности указанной процедуры. Мы не будем ее подробно касаться, особенно в применении к модели Хопфилда, укажем лишь на классическую работу [30]. Как правило, теория среднего поля становится точной в термодинамическом пределе $N \to \infty$ для моделей с исчезающе слабым взаимодействием бесконечно большого радиуса. Поэтому, чтобы корректно применить теорию среднего поля в конкретном случае, необходимо проследить, чтобы число взаимодействующих спинов было очень большим, а интенсивность взаимодействия каждой пары спинов становилась малой. Отметим, что наличие множителя $1/N$ в матрице T_{ij} (см. выражение (14)) и двойное суммирование по всем спинам позволяют с большой достоверностью воспользоваться результатами теории среднего поля для больших, но конечных значений N.

Обозначим среднее поле, создаваемое остальными спинами в i-м узле:

$$U_i = \langle H_i \rangle_T = \sum_{j=1}^{N} T_{ij} \langle S_j \rangle_T . \qquad (17)$$

Тогда, в соответствии с простейшей задачей о среднем значении спина во внешнем поле, получим выражение вида

$$V_i = \langle S_i \rangle_T = \tanh\left(\sum_{j=1}^{N} T_{ij} V_j / T\right). \qquad (18)$$

Уравнение (18) в теории среднего поля называется уравнением самосогласования. В нем неизвестные значения V_i выражаются через все остальные V_j. Для пространственно однородного случая, когда V не зависит от номера узла j, имеем одно уравнение, для общего случая нейронной сети — систему уравнений. Решить эту систему точно практически невозможно, однако нейронная сеть может ее решить сама (!) последовательными итерациями.

Как и в предыдущем классе задач ассоциативной памяти, нейронная сеть на каждом шаге эволюции вычисляет локальное поле (17) для каждого нейрона (т.е. вычисляет аргумент гиперболического тангенса в выражении (18)), после чего определяет новые средние значения спина V_i. Отличие динамики среднего поля с температурой от динамики

(2) состоит в том, что функция отклика является непрерывной сигмоидной функцией, а не ступенчатой. В силу этого спин ± 1 заменяется своим средним значением V_i, принимающим любое значение в интервале $(-1,1)$. Для указанной динамики существует критическая «температура» T_c, выше которой возможно только тривиальное конечное состояние $V_i = 0$ при неограниченном возрастании времени [11]. Таким образом, чтобы найти глобальный минимум функционала (16) необходимо взять начальную температуру T достаточно большую, но ниже критической, а после попадания в окрестность глобального минимума постепенно ее уменьшать до нуля с целью улучшения точности решения задачи.

Ниже, в разд.5, модель Хопфилда используется как основной инструмент в задачах распознавания треков заряженных частиц благодаря тому, что априорная информация о магнитном поле и геометрии детектора позволяет заранее определить матрицу синаптических весов.

Процедура «имитационного отжига». В процедуре минимизации энергетической функции ИНС возникает реальная опасность попадания в один из локальных минимумов вместо глобального. Возможный способ избежать этого лучше всего пояснить на механической аналогии с тяжелым шариком, перекатывающимся по поверхности энергетической функции (см. рис.7). При умеренном встряхивании системы шарик с большой вероятностью перепрыгнет на более глубокий уровень и останется там. Но если встряхнуть сильно, то он с равной вероятностью может попасть на любой из уровней. Поэтому наилучшей стратегией будет сначала встряхнуть сильно, а потом встряхивать все более умеренно. В этом можно усмотреть сходство с процедурой отжига металлов, когда низкая энергия связей в металле достигается начальным нагреванием до очень высокой температуры, вызывающей фазовый переход через точку разрушения кристаллических связей, с последующим медленным охлаждением. Поэтому соответствующая процедура достижения глобального минимума энергетической ИНС получила название имитационного отжига.

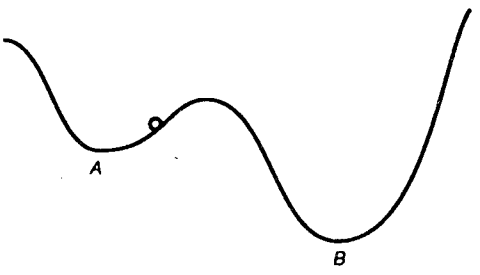


Рис.7. К механизму имитационного отжига

Вот как выглядит алгоритм имитационного отжига для сети Хопфилда [31]. Обозначим температуру T, время работы сети при каждом значении температуры t, число входных узлов N.

1. Начальное состояние $T_0 = 1{,}5$, $t_0 = N$.

2. Медленный нагрев: $T \Leftarrow T/0{,}8$, $t = N$ до тех пор, пока удельная флуктуация энергии $(\langle E^2\rangle - \langle E\rangle^2)/T$ не станет меньше 0,05.

3. Постепенное охлаждение: $T \Leftarrow T \cdot 0{,}95$, $t = N$, пока число изменивших свое состояние нейронов не превысит 50%.

4. Замедленное охлаждение: $T \Leftarrow T \cdot 0{,}95$, $t = 16N$ до установления стабильного состояния сети.

Метод деформируемых образцов. Метод деформируемых образцов в теории ИНС состоит в том, чтобы нейронная сеть для заданного набора эталонных образцов, определяемых рядом параметров, вычислила бы оптимальные значения этих параметров. Продемонстрируем работу этой модели на одной из типичных задач определения параметров наперед заданного числа треков, имеющих форму пространственной спирали [32].

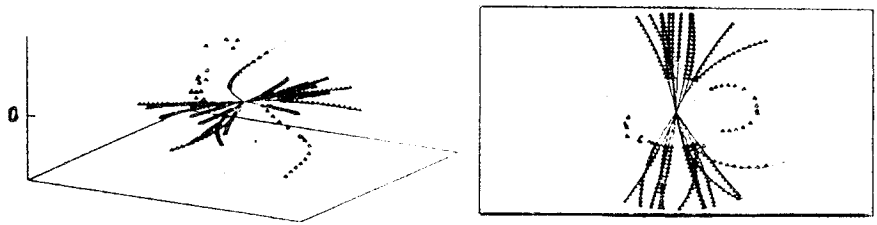
Предположим, что мы имеем дело с множеством M спиральных деформируемых образцов, заданных углом θ эмиссии трека в плоскости XY, кривизной κ и параметром γ, определяющим наклон трека к оси Z (вдоль магнитного поля): $(\theta_a, \kappa_a, \gamma_a)$, $a = 1,\ldots,M$. Требуется аппроксимировать ими экспериментальные точки (x_i, y_i, z_i). Мера качества аппроксимации

$$E[S_{ia}; \theta_a, \kappa_a, \gamma_a] = \sum_{i,a} S_{ia} M_{ia} + \lambda \sum_i \left\{ \sum_a S_{ia} - 1 \right\}^2, \qquad (19)$$

где двоичный нейрон S_{ia} определяет принадлежность i-й экспериментальной точки a-му треку: $S_{ia} = 1$ — в случае принадлежности и $S_{ia} = 0$ — в противном случае. Мы хотим минимизировать $E[S_{ia}; \theta_a, \kappa_a, \gamma_a]$ по отношению к S_{ia}, θ_a, κ_a и γ_a при условии, что каждая точка может или лежать только на одном треке, или не лежать ни на одном треке (быть шумовой).

Первый член уравнения (19) описывает *взаимодействие точки с треком*, а M_{ia} характеризует силу этого взаимодействия. Обычно M_{ia} берется равным квадрату расстояния от точки до трека и в этом случае уравнение является обобщением метода наименьших квадратов в терминах нейронной сети.

Второй член в уравнении (19) определяет *критическое расстояние*, до которого точка действует на трек: если $M_{ia} > \lambda$, то энергетически выгодно взять $S_{ia} = 0$. В случае M_{ia}, определяемого как квадрат расстояния от трека до нейрона, критическое расстояние равно $\sqrt{\lambda}$. Такой подход соответствует робастности восстановления параметров трека [33].



а *б*

Рис.8. Результат работы метода деформируемых образцов: *а* — трехмерное изображение, *б* — проекция на плоскость *xy*

Задача поиска глобального минимума энергетической функции (19) решается аналогично стандартной нейронной сети. Для последовательности уменьшающихся температур метод градиентного спуска дает:

$$\Delta \alpha_a = -\eta \sum_i \hat{S}_{ia} \frac{\partial M_{ia}}{\partial \alpha_a} \tag{20}$$

с $\alpha_a = (\theta_a, \kappa_a, \gamma_a)$. В этом уравнении выделен множитель

$$\hat{S}_{ia} = \frac{e^{-M_{ia}/T}}{e^{-\lambda/T} + \sum_b e^{-M_{ib}/T}}. \tag{21}$$

Метод деформируемых образцов был протестирован на моделированных данных для CERN DELPHI TPC детектора. Результат показан на рис.8. Алгоритм работал достаточно хорошо даже в случае пересекающихся или близко проходящих треков. Такая надежность метода определяется в основном множителем (21).

По сути, метод деформируемых образцов является интерпретацией так называемого метода «гибкой руки» или эластичной сети [34,35], которые весьма перспективны для распознавания треков в условиях высокой множественности и зашумленности (сравнение с обычными методами дано в [35]).

Самообучающиеся ИНС. К существенным недостаткам ИНС типа классифицирующего персептрона с обучением по выборке относятся прежде всего необходимость знать заранее число классифицируемых признаков, а также умение выбрать наиболее представительные и малокоррелированные из них. Поэтому наряду с ИНС, обучаемыми по выборке, значительный интерес вызывают самообучающиеся ИНС (СИНС)

[36]. СИНС не требует предварительного знания искомых признаков, а выделяет их сама в соответствии со структурой анализируемых данных, используя алгоритм кластеризации.

Для нейронов h_j из скрытого слоя выбирается «победитель» h_m, например, по максимуму среднего поля

$$h_m = \max_j(h_j), \tag{22}$$

который в простейшей СИНС с организацией «победитель-забирает-все» становится более «чувствительным» к предъявленным данным x_k путем пересчета весовых коэффициентов T_{jk}. В СИНС с «соревновательным» обучением среднее поле H_j скрытого нейрона h_j вычисляется не так, как в (17), а как скалярное произведение весового вектора $\mathbf{T}_j = (T_{j1}, T_{j2}, \ldots T_{jn})$ и входного вектора $\mathbf{X}_N = (x_1, x_2, \ldots, x_N)$

$$H_j = |\mathbf{T}_j||\mathbf{X}_N| \cos\theta_j. \tag{23}$$

Таким образом, нахождение победителя h_m в (22) ведет к определению $\mathbf{T}_j$, ближайшего к входному вектору $\mathbf{X}_N$.

Для этого, как и в (8), минимизируется среднеквадратичная ошибка их разности

$$E = \frac{1}{2} \sum_{\mathbf{X} \in M} (\mathbf{X} - \mathbf{T}_j)^2 \to \min_k,$$

где M — подмножество входных узлов (кластер), дающее h_m как максимум. Подмножество M может меняться в процессе самообучения.

Динамика обновления весов получается при использовании градиентного спуска

$$\Delta T_{jk} = -\eta \frac{\partial E}{\partial T_{jk}} = \eta \sum_{\mathbf{X} \in M} (x_k - T_{jk}),$$

где η — по-прежнему параметр скорости обучения. На практике, однако, весовой вектор $\mathbf{T}_j$ обновляется при каждом сравнении с образцом $\mathbf{X} \in M$ в соответствии с Δ-правилом

$$\Delta T_{jk} = \eta(x_k - T_{jk})_{\mathbf{X} \in M}$$

с перенормировкой

$$T_{jk} \to \frac{T_{jk}}{\sqrt{\sum_{k'} T_{jk'}^2}}.$$

Определение подмножества M может осуществляться по близости к победителю, так что Δ-правило изменяется так:

$$\Delta \mathbf{T}_j = \eta \, \Lambda \, (j, m)(\mathbf{X} - \mathbf{T}_j),$$

где $\Lambda \, (j, m)$ — функция соседства, предназначенная для подавления (исключения) боковых связей. Обычно $\Lambda \, (j, m)$ имеет форму «мексиканской шляпы» или, как это принято в распространенном пакете JETNET 2.0 [36], является характеристической функцией множества $\| \mathbf{h}_j - \mathbf{h}_m \| \le \lambda$ узлов $\mathbf{h}_j$, отстоящих от«"победителя» $\mathbf{h}_m$ на λ:

$$\Lambda \, (j, m) = \begin{cases} 1, & \text{если } \| \mathbf{h}_j - \mathbf{h}_m \| \le \lambda; \\ 0, & \text{для остальных } j. \end{cases}$$

Теоретические основы СИНС изложены в [15]. Результаты применения СИНС для распознавания струйных кластеров в адронных экспериментах, где требовалось разделить данные с различными типами кварков, приведены в работе [36].

4. РЕАЛИЗАЦИЯ НЕЙРОННЫХ СЕТЕЙ

С простейшим способом реализации нейронной сети мы уже фактически познакомились. Это виртуальная нейронная сеть в виде программы для компьютера. Правда, желательно, чтобы это был достаточно производительный компьютер параллельного типа. Отметим здесь характерные черты нейронной сети, позволяющие и другие возможности ее реализации:

— нейрон представляет собой очень простое логическое устройство;

— система состоит из очень большого числа одинаковых нейронов, причем результат работы нейронной сети малочувствителен к характеристикам конкретного нейрона;

— каждый нейрон связан с очень большим числом других нейронов;

— веса связей различны и, в зависимости от решаемой задачи, могут изменяться;

— в системе возможна существенная параллельность обработки информации [37].

Существует целый ряд фирм, специализирующихся в этом направлении, которыми созданы пакеты успешно работающих на компьютерах IBM PC программ по обработке результатов экспериментов. Одна из этих программ, JETNET 2.0 [36], специально предназначена для обработки данных в физике высоких энергий и, в частности, для поиска струй при множественном рождении частиц. Оригинальный текст программы,

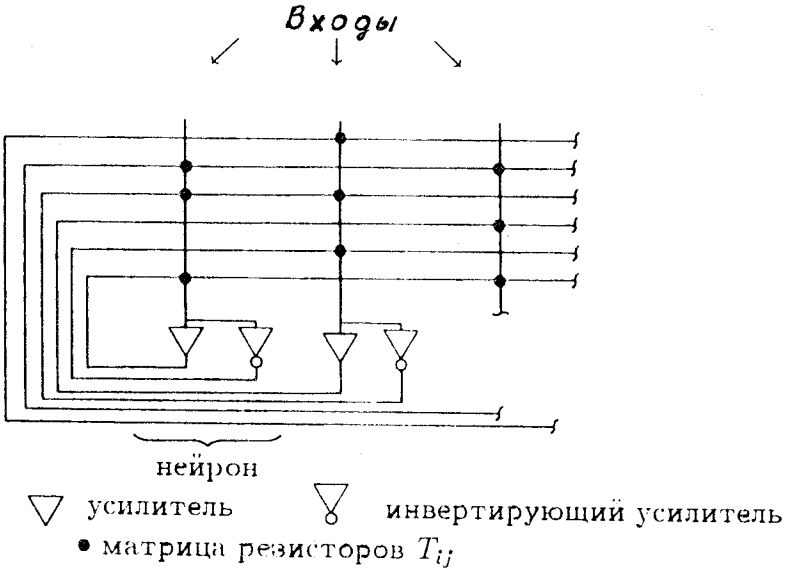


Рис.9. Реализация ИНС в виде электрической схемы [11]

написанный на фортране, распространяется через библиотеку программ журнала «Computer Physics Communications».

С развитием технологии оказалось возможным создать специализированный аналоговый компьютер, в котором функции нейрона исполняет простейший процессор, либо имеющий дискретный выход, либо представляющий собой небольшой усилитель с сигмоидной выходной характеристикой. Функции матрицы синаптических весов играет матрица омических сопротивлений, расположенных в местах пересечения выходов и входов этих усилителей, причем возможно реализовать планарную схему, когда все элементы соединены со всеми (рис.9). Дифференциальные уравнения для напряжения приведены в работе [11], где проанализировано влияние непрерывной выходной характеристики нейронов на получающееся предельное состояние нейронной сети. Подобная действующая реализация нейронной сети описана в работе [38].

В своем интервью в газете «Поиск» (№ 13, 1993) директор Научного центра нейрокомпьютеров (НЦН РАН) А.Галушкин определяет нейрокомпьютеры как вычислительную систему, алгоритмы решения задач в которой представлены в логическом базисе ИНС. Помимо реализации нейромашины на обычной исследовательской ЭВМ уже существуют раз-

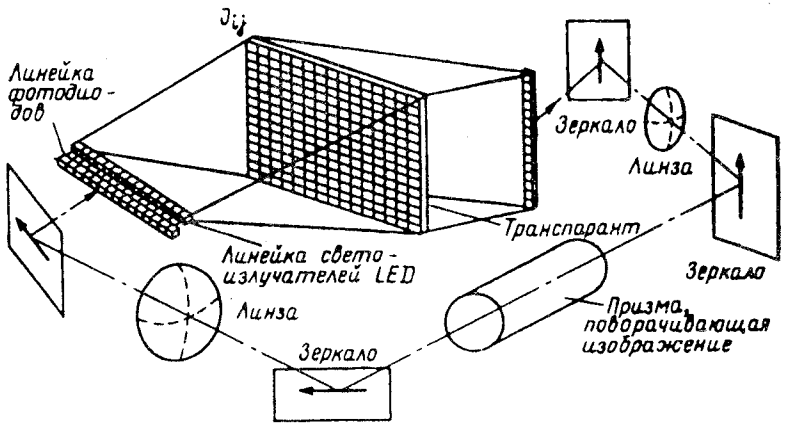


Рис.10. Оптическая реализация нейронной сети

ные варианты параллельных нейрокомпьютеров вплоть до мощной супер-ЭВМ с реализацией множества нейронов на одном кристалле. Цель создания этих машин — расчеты оптимизационных задач на 2—3 порядка быстрее, чем в существующих супер-ЭВМ при одинаковых стоимости и энергопотреблении. Уже идет разговор о производительности в терафлопах — 10^{12} операций с плавающей запятой в секунду.

Интересно, что первые нейрокомпьютеры в России начали создаваться в то же время, что и в США, — в конце 60-х — начале 70-х годов. В них реализовывался трехслойный персептрон с весами, предварительно вычисленными при обучении на обычных ЭВМ. Они использовались для задач, не требовавших больших скоростей изменения параметров, таких как задачи гидролокации и медицинской диагностики. Последующий «застой» в развитии нейрокомпьютеров в России был вызван общим отставанием в технологии, преодолеваемым только к настоящему времени.

Самой перспективной реализацией нейронной сети, открывающей путь к архитектуре ЭВМ шестого поколения, является оптический нейрокомпьютер. В нем возможно достижение массового параллелизма и быстродействия на 2—3 порядка выше существующих супер-ЭВМ при прочих равных условиях. Большим преимуществом является то, что световые лучи при распространении не оказывают влияния друг на друга и могут многократно перекрещиваться. Матрица синаптических связей, будучи реализованной в виде объемной голограммы размером 1 $см^3$, может содержать более триллиона связей для записанных в нее оптических изображений [39]. Оптическая реализация модели Хопфилда описана в рабо-

те [40], а оптическая реализация модели ассоциативной памяти — в [41] (см. рис.10). При этом ничто не мешает все более миниатюризировать элементы нейронной сети, что позволяет с успехом использовать достижения нанотехнологий.

Впечатляюще выглядит краткая сводка результатов применения на тевatроне FNAL [44] специально сконструированных многослойных персептронов, реализованных на одном кристалле, как для быстрого (несколько микросекунд) on-line распознавания мюонных треков с точностью, всего вдвое уступающей сложным off-line методам, так и для распознавания струй в $p\bar{p}$-столкновениях, а также триггеров, реализованных аппаратно на коммерческой ИНС (INTEL 8170NX Electrical Trainable Analog Neural Network, Intel Corp. Santa Clare, California).

С развитием технологии, когда транспьютерные системы перешли рубеж микроминиатюризации в 1,5 мкм, возникает возможность создания и «нейрочипов». К сожалению, нам эта технология еще недоступна в силу жесткого запрета КОКОМ. Тем не менее попытки разработок нейрочипов на менее тонкой отечественной технологии могут привести к значительному росту результатов по производительности. Отечественное производство нейрокомпьютеров на стандартной микропроцессорной базе фактически не уступает американскому. В частности, как утверждает А.Галушкин в указанной статье, отечественный нейрокомпьютер «Геркулес» не отстает по производительности от известного нейрокомпьютера AWZA.

Наибольшей трудностью в настоящее время являются проблемы перевода многих актуальных задач, связанных с распознаванием образов при обработке сигналов и изображений, с управлением нелинейными динамическими системами, возникающими в автоматизации механических, физических и химических процессов, в алгоритмы, пригодные для ввода их в нейрокомпьютеры. В этой связи возник новый раздел вычислительной математики — «нейроматематика».

5. ПРИЛОЖЕНИЯ ИНС В ФИЗИКЕ ЭЛЕМЕНТАРНЫХ ЧАСТИЦ

Три этапа анализа экспериментальных данных. Среди все расширяющегося числа приложений ИНС значительное место занимают задачи анализа экспериментальных данных в физике элементарных частиц. Их можно условно разбить на три группы в соответствии с тремя основными этапами обработки данных (on-line детектирование, off-line обработка для получения физических параметров и последующая интерпретация с проверкой гипотез). Саму эту обработку можно рассматривать как жесткий отбор, направленный на кардинальное сокращение избыточ-

ности в экспериментальной информации, так как в современных высокостатистических экспериментах полезное событие может содержаться среди 10^{10}—10^{12} фоновых. Жесткость методов не должна при этом допускать потери этого одного единственного полезного события. Примеры экспериментальных установок и формулировки возникающих математических проблем обработки на всех трех этапах можно найти в обзорах [50,51].

Первый этап такого отбора осуществляется уже в процессе самого эксперимента с помощью так называемой системы триггеров, т.е. средств сверхбыстрой электроники, для выделения возможной полезной информации из гигантского потока экспериментальных данных (свыше 10^7 событий в секунду). Это дает резкое (на 4—5 порядков) повышение вероятности нахождения интересных событий и соответствующее сокращение объема регистрируемых данных, подлежащих обработке на следующем, втором этапе. Приложения ИНС на первом этапе анализа экспериментальной информации проиллюстрированы на примерах в наиболее свежих публикациях [42,52,53], где приведены описания реализации триггеров второго уровня в различных калориметрах и системах быстрого трекинга.

Использованию ИНС в триггере второго уровня для обнаружения струй и центров ливней в данных калориметров были посвящены работы [45,46], причем в первой из них использовалась коммерческая ИНС. В обоих случаях применение трехслойной сети с одним выходом при соответствующем обучение на нескольких сотнях событий показало ее полную пригодность для триггера 2-го уровня. Дополнительного исследования требует вопрос о числе нейронов в скрытом слое и длине обучающей выборки. Исследования по применению ИНС типа трехслойных персептронов для проектирования калориметрического триггера по отбору событий с B-мезонами проводились в ОИЯИ [52,53].

На втором этапе данные фильтруются от фона, распознаются траектории заряженных частиц и диссипация энергии нейтральных частиц, события реконструируются в пространстве для определения искомых физических параметров.

Исследования по применению ИНС и клеточных автоматов на втором этапе обработки для фильтрации трековой информации и восстановления траекторий заряженных частиц будут рассмотрены более подробно в дальнейшем на примере экспериментальных данных, зарегистрированных на цилиндрическом спектрометре АРЕС (ОИЯИ). Детально представлены методы ускорения вычислений и преодоления трудностей, возникающих при росте множественности событий и зашумленности данных.

Данные, полученные на втором этапе, используются на заключительном, третьем этапе для выбора наиболее правдоподобной физической гипотезы с помощью статистических критериев. На третьем этапе ИНС применяется в задачах классификации кварков, идентификации распадов короткоживущих частиц и оценки двух эмпирических распределений.

Задача определения на треке излома под малым углом решалась с помощью трехслойного персептрона [43], на вход которого вводились отклонения в двух плоскостях измеренных трехмерных координат от подогнанной геликоиды и кривизна последней. Выходной нейрон принимал значение ± 1 в зависимости от наличия или отсутствия излома. В случае излома те же данные подавались на вход другой ИНС, выдававшей радиальную координату точки излома. Тренировка сети и тестирование ее робастности по отношению к пропускам отсчетов при измерении проводились на данных, моделировавших эксперимент ALEPH. Сравнение результатов с наиболее мощным статистическим методом, использующим фильтр Калмана, показало, что при сравнимой точности и надежности метод ИНС работал в несколько десятков раз быстрее.

В работе [47] многослойный персептрон был применен после обучения на 27 тыс. событиях, соответствующих классам с b-, c-, и легкими кварками для последующей классификации 50 тыс. событий по этим трем классам. Значительное внимание было уделено сокращению числа входных переменных (до 15) путем применения методов дискриминантного анализа, что обеспечило более быстрое обучение и последующую работу сети.

Аналогичный подход был предложен в работе [48] для классификации распадов Z^0-бозонов по данным эксперимента DELPHI и в [49] для идентификации возможных каналов τ-распадов. В работе [48] приведены подробные объяснения по выбору числа нейронов в скрытом слое и изменению обучающих параметров, а также введены такие параметры качества работы сети после обучения, как сигнальная эффективность ε_s (число событий, правильно отнесенных к определенному классу, по отношению к общему числу событий в этом классе) и частота p (число событий, правильно отнесенных к определенному классу, по отношению к общему числу событий, классифицированных как входящие в этот класс).

Клеточный автомат. Клеточный автомат можно рассматривать как упрощенный локальный вариант нейронной сети. Обладая простотой и наглядностью, он позволяет понять основные особенности параллельного алгоритма.

Клеточные автоматы [54] являются динамическими системами, эволюция которых развертывается в дискретных, обычно плоских, пространствах, состоящих из клеток (ячеек). Каждая клетка может принимать несколько значений, в простейшем случае однобитовой клетки — 0 и 1.

Законы эволюции локальны, т.е. динамика системы задается неизменным набором правил, например таблицей, по которой осуществляется вычисление нового состояния клеток в зависимости от состояния окружающих ее ближайших соседей. Существенно, что эта смена состояний происходит одновременно и параллельно, а время идет дискретно.

Таким образом, клеточные автоматы позволяют моделировать многие естественнонаучные явления, а также создавать общие модели параллельных вычислительных процессов (подобно машине Тьюринга, позволяющей моделировать самые общие последовательные вычисления).

Особенно широкую популярность клеточные автоматы приобрели в 70-е годы благодаря публикациям М.Гарднера в журнале «Scientific American» (позже переведенным и на русский язык), посвященным игре Дж.Конвея «Жизнь» [55]. Правила простого клеточного автомата в этой игре имитируют развитие колонии стилизованных организмов — клеток, рождающихся и умирающих (т.е. принимающих значения 1 и 0), в зависимости от числа соседей, окружающих каждую из этих клеток. Мерой течения времени служит смена поколений колонии, которая происходит по следующим правилам:

1. Соседями клетки считаются все живые клетки, находящиеся в восьми ячейках, расположенных рядом с данной по горизонтали, вертикали и диагонали.

2. Если у некоторой клетки меньше двух соседей, она погибает от одиночества. Если клетка имеет больше трех соседей, она погибает от тесноты.

3. Если рядом с пустой ячейкой окажется ровно три живые соседние клетки, то в этой ячейке рождается живая клетка.

4. Гибель и рождение происходят в момент смены поколений. Таким образом, гибнущая клетка может способствовать рождению новой, но рождающаяся клетка не может воскресить гибнущую, и гибель одной клетки, уменьшив локальную плотность населения, не может предотвратить гибель другой.

Такая колония может все время расти, непрерывно меняя свое расположение, форму и число клеток. Однако чаще колония становится в конце концов стационарной или циклически повторяет один и тот же конечный набор состояний. Длина цикла называется периодом колонии. Пример развития колонии показан на рис.11. Номер поколения увеличивается направо. Верхний ряд представляет осциллирующий триплет, так называемый блинкер. Конфигурация, становящаяся стабильной на третьем шаге, изображена в среднем ряду. В нижнем ряду представлена более сложная конфигурация, сначала растущая до седьмого шага, а затем распадающаяся на четыре блинкера.

На основе этого примера можно сформулировать общие правила построения клеточных автоматов:

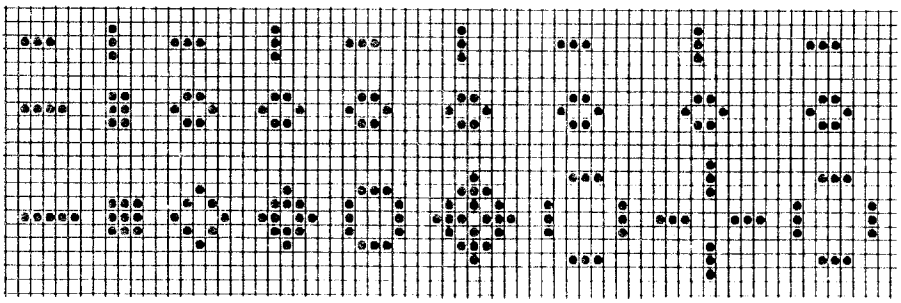


Рис.11. Примеры развития колоний в игре «Жизнь»

1. Состояние клеток дискретно (обычно 0 и 1, хотя могут быть автоматы и с большим числом состояний).

2. Соседями являются ограниченное число клеток, часто это ближайшие клетки.

3. Правила, задающие динамику развития клеточного автомата, обычно имеют простую функциональную форму и зависят от решаемой проблемы.

4. Клеточный автомат является тактируемой системой, т.е. смена состояний клеток происходит одновременно.

На основании правил построения клеточного автомата сформулируем специфические особенности автомата для фильтрации треков в пропорциональных камерах.

Во-первых, определим живую клетку как кластер, т.е. непрерывную группу сработавших проволочек, а мертвой клеткой назовем пустую ячейку, не содержащую отсчета. Для поддержания жизнеспособности разрывных (из-за несрабатывания камер) треков необходимо также введение клеток-фантомов, которые соответствовали бы кластерам в случае правильного срабатывания камер. Таким образом, в нашем случае клетка имеет 4 состояния.

Во-вторых, для задания правила определения соседей рассмотрим характерные особенности дискретного детектора типа многопроволочной пропорциональной камеры [56].

В общем случае многопроволочная пропорциональная камера (рис.12) состоит из плоской сетки эквидистантно расположенных анодных проволочек, которая лежит между двумя параллельными плоскостями заземленных катодов. Когда заряженная частица проходит сквозь сетку, в газе, заполняющем камеру, возникает электронная лавина, приводящая к импульсу на ближайшей проволочке. Эквипотенциальные

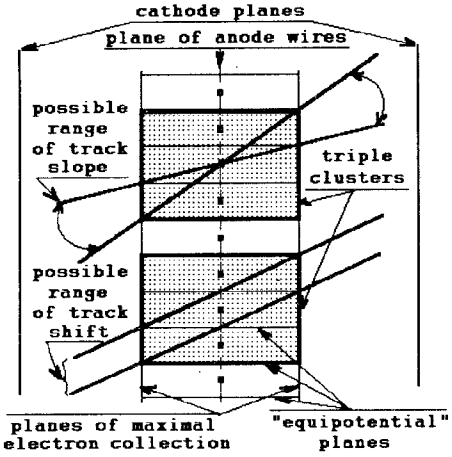


Рис.12. Схематическое представление работы пропорциональной камеры

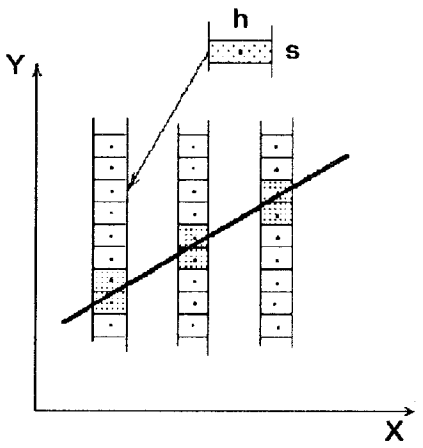


Рис.13. Модель детектирующей установки

линии в пропорциональной камере позволяют провести между проволочками воображаемые эквидистантные плоскости, перпендикулярные плоскости проволочек. Кроме того, вследствие присутствия электроотрицательных добавок в газе камер, функция распределения сбора электронов является ступенчатой, т.е. по обе стороны от плоскости проволочек можно провести еще две воображаемые плоскости, такие, что все электроны, образовавшиеся внутри, будут собраны на проволочке. Но электроны, дрейфующие извне области, ограниченной этими двумя плоскостями, будут поглощены газом.

Обе эти особенности позволяют нам рассматривать камеру как цепочку воображаемых прямоугольников, каждый из которых окружает сигнальную проволочку (рис.13). Две параллельные стороны каждого прямоугольника определяются «эквипотенциальными» плоскостями, две другие формируются плоскостями максимального сбора электронов. Такая простая модель является достаточно хорошим приближением для описания эффектов функционирования пропорциональной камеры.

Когда заряженная частица пересекает прямоугольник, проволочка внутри него срабатывает. Если трек пересекает несколько соседних прямоугольников в камере, все они срабатывают, образуя кластер. В последнем случае трек пересек левую сторону нижнего прямоугольника и правую сторону верхнего (или наоборот, в зависимости от направления прохождения трека). Наиболее важной чертой этой модели дискретного детектора является возможность приближенного решения обратной за-

дачи восстановления трека, а именно: зная кластерную структуру, можно сделать заключение об области возможных углов прохождения трека через кластер. Кроме того, когда наклон трека фиксирован, кластерная структура определяет возможную область точек пересечения трека с плоскостью сигнальных проволочек.

Итак, зададим правило определения соседей, опираясь на характерные особенности трека в дискретном детекторе: соседями могут быть такие клетки, лежащие на смежных камерах, через которые можно провести физически разумный (допустимый) трек. Определим также область возможных соседей как область, которую заметают на смежных камерах такие допустимые треки.

В-третьих, определим правила эволюции автомата, чтобы отсеять шум и оставить треки. В первую очередь, мы должны восстановить несработавшие кластеры в камерах. Предположим, что если на данной камере в месте пересечения областей возможных соседей смежных камер соседей нет, то произошел сбой в работе камеры или электроники. В этом случае необходимо родить клетку-фантом, являющуюся соседом клеток, указывающих на нее. Затем необходимо уничтожить шумовые точки, т.е. точки, у которых или слишком мало соседей, или слишком много. Для случая трехтрековых событий будем уничтожать клетки, имеющие меньше двух или больше четырех соседей, Чтобы не допустить вымирание треков с концов, введем условные 0 и $(N + 1)$ камеры, заполненные соседями, поддерживающими все клетки крайних камер.

В-четвертых, разнесем рождение и смерть клеток на разные такты. На первом такте будем осуществлять рождение клеток, а на втором — вымирание. И так на каждом шаге. Эта мера обеспечивает выживание разрывных треков.

Чтобы уловить выход автомата на стабильное состояние или циклическую повторяемость, для каждого поколения считается контрольная сумма CRC и в случае совпадения контрольных сумм итерации прекращаются.

В остальных деталях трековый фильтр совпадает со стандартным клеточным автоматом.

Клеточный автомат был опробован на трехтрековых данных, полученных в эксперименте по поиску распада $\mu^+ \to e^+e^+e^-$ [57] и исследованию распада $\pi^+ \to e^+\nu_e e^+e^-$ [58]. В этом эксперименте было включено 10 камер, т.е. на один трек в среднем приходится 10 экспериментальных точек (кластеров).

Пример работы созданного клеточного автомата показан на рис.14. На этом рисунке крестиками обозначены кластеры, отброшенные автоматом как шумовые. В верхней части камер 2 и 3 (начиная от центра)

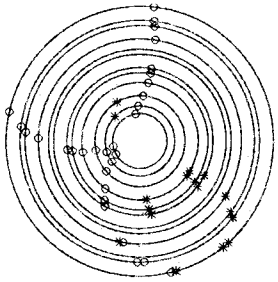


Рис.14. Пример работы клеточного автомата

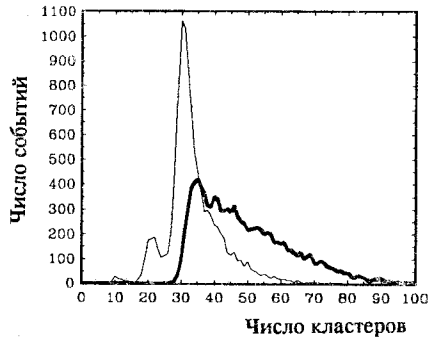


Рис.15. Распределение числа событий по количеству кластеров в событии до работы клеточного автомата и после (более тонкая линия)

видны два отброшенных шумовых кластера. Трек в нижней части рисунка возвратился в детектор после столкновения с внешней стенкой, и эта часть была также отброшена автоматом. В камерах 9,11 и 12 близко к треку есть кластеры от дельта-электронов. Кластерная длина не показана на рисунке, но два из этих кластеров были отброшены именно на основе анализа кластерной длины. Однако дельта-электронный кластер в камере 11 оказался той же длины, что и кластер, лежащий на треке. Поэтому автомат оставил оба эти кластера как подходящие. Отметим отсутствие срабатывания камеры 10 на этом треке. В этой камере клеточный автомат произвел клетку-фантом, что спасло этот разрывный трек от уничтожения. Эта клетка-фантом не показана на рисунке, так как использовалась временно только на стадии эволюционного процесса клеточного автомата. Камеры 7 и 8 не использовались во время эксперимента.

На рис.15 приведено распределение числа событий по количеству кластеров в событии (более тонкая линия — распределение после работы автомата). На первоначальном распределении видно, что в результате работы программ предварительного отбора (on-line обработка) почти нет событий с числом кластеров меньше 30, а также виден длинный шумовой хвост, простирающийся до 100 кластеров на событие. На распределении, полученном в результате работы автомата, четко обозначены пики в области однотрековых и двухтрековых событий, оставшихся в результате ложной интерпретации на предыдущем этапе обработки, и трехтрековых событий (10, 20 и 30 кластеров соответственно). Видно, что клеточный автомат хорошо отсеивает шум (в среднем 65—70%), после чего мы можем проводить естественное определение числа треков по числу кластеров.

Достоинствами автомата являются его простота и быстрота работы. Программа, реализующая клеточный автомат на персональном компьютере, обрабатывает примерно 25 событий в секунду. Постановка в компьютер специализированной платы, способной выполнять *параллельную* работу, увеличивает скорость обработки до 1500 событий в секунду и выше. Такая скорость позволяет использовать клеточный автомат в линию с экспериментом (on-line режим).

Но наиболее важным для дальнейшего результатом работы клеточного автомата является группировка клеток по принципу возможной принадлежности треку. Очевидно, что локальность работы клеточного автомата (учет только ближайших соседей) не дает ему возможности разделить близкие или пересекающиеся треки. Эта задача должна быть выполнена на следующем этапе обработки.

Эффективность работы клеточного автомата оценивалась просмотром по нескольким выборкам и составила 98%.

Применение ИНС для обнаружения треков заряженных частиц. *Модель сегментов.* Первую попытку использовать нейронные сети для распознавания треков сделали Петерсон [59] и Денби [60]. Их подход (так называемый метод сегментов) вкратце можно описать следующим образом.

Имеется множество N экспериментальных точек на плоскости. Требуется провести непрерывные гладкие кривые (треки) через эти N точек. Предполагается, что треки не имеют изломов и разветвлений.

Вводятся бинарные нейроны s_{ij}, определяющие, соединяются точки i и j ($s_{ij} = 1$) или нет ($s_{ij} = 0$), т.е. принадлежит данный направленный ($s_{ij} \neq s_{ji}$) сегмент треку или нет. Энергетическая функция конструируется в виде:

$$E = E_{\text{cost}} + E_{\text{constr}}. \tag{24}$$

Первый член (стоимостный) выбирается так, чтобы он поощрял короткие смежные сегменты с малым углом между ними (см. рис.16):

$$E_{\text{cost}} = -\frac{1}{2}\sum_{ijkl}\delta_{jk}\frac{\cos^m\theta_{ijl}}{r_{ij}\,r_{jl}}, \tag{25}$$

где m — нечетный целочисленный показатель степени.

Второй член (штрафной) состоит из двух частей:

$$E_{\text{constr}} = \frac{\alpha}{2}\left[\sum_{ik} s_{ik}s_{kl} + \sum s_{ij}s_{jl}\right] + \frac{\beta}{2}\left[\sum_{ij} s_{ij} - N\right]^2. \tag{26}$$

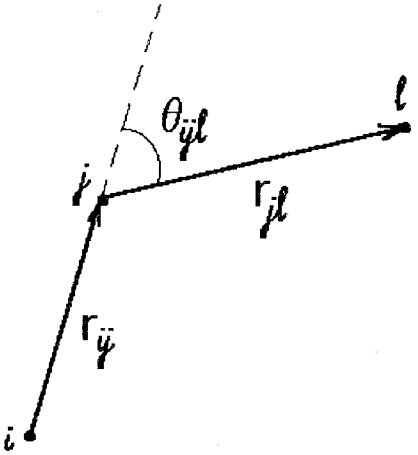


Рис.16. Модель сегментов

Первая часть учитывает запрет разветвлений, а вторая — устанавливает баланс между числом активных нейронов и числом экспериментальных точек. Параметры α и β являются множителями Лагранжа*.

Применение метода градиентного спуска к энергетической функции (24) приводит к уравнению эволюции для нейрона s_{ij} в виде ступенчатой функции:

$$s_{ij} = \frac{1}{2}\left(1 + \operatorname{sign}\left(-\frac{\Delta E}{\Delta s_{ij}}\right)\right). \quad (27)$$

Однако такая процедура скорее всего приведет в какой-нибудь локальный минимум энергетической функции, что не является удовлетворительным решением задачи. Стандартным путем преодоления этой проблемы является введение статистического шума в систему при помощи теории среднего поля, что приводит к замене ступенчатой функции (27) на функцию сигмоидного вида:

$$v_{ij} = \frac{1}{2}\left(1 + \tanh\left(-\frac{\partial E}{\partial v_{ij}}\frac{1}{T}\right)\right). \quad (28)$$

Здесь по-прежнему $v_{ij} = \langle s_{ij} \rangle_T$ — усредненные по температурному ансамблю значения дискретных нейронов, являющиеся уже непрерывными нейронами с областью изменения [0,1]. Уравнения (28) решаются итерационно до достижения стабильного состояния. Параметры α, β, T являются подгоночными.

В модели сегментов N экспериментальных точек приводят к $N(N-1)$ нейронам и уравнениям. В принципе требуется N^3 операций для выполнения каждой итерации. Однако, вследствие локальной природы большинства задач поиска треков, это число может быть существенно уменьшено. Очень маловероятно, что две экспериментальные точки, находящиеся далеко друг от друга, будут соединены непосредственно. Поэтому можно ввести радиус обрезания R_{cut}, который характеризует область взаимодействия нейронов. Это приводит к уменьшению

*Обычно решения оказываются устойчивыми по отношению к этим параметрам [61].

Рис.17. Время сходимости модели сегментов в зависимости от числа активных нейронов

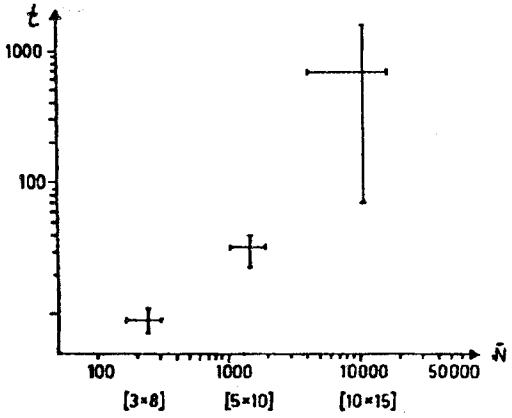
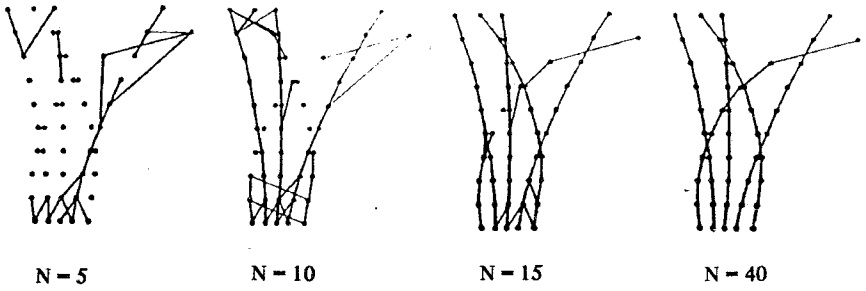


Рис.18. Эволюция нейронной сети в модели сегментов [62]. Изображены только нейроны, у которых $v_{ij} > 0{,}5$, N — число итераций



числа активных нейронов в 2—3 раза. При среднем числе активных партнеров $\widetilde{N}$ в области взаимодействия R_{cut} требуется $O(N\widetilde{N}^2)$ вычислений. Результаты не очень чувствительны к R_{cut}.

В конце эволюции активными считаются нейроны с $v_{ij} > 0{,}5$. Затем конечное состояние подвергается «чистке», т.к. сеть может оставить некоторое число разветвлений.

Метод сегментов хорошо распознает события с небольшим количеством треков при отсутствии шумовых экспериментальных точек. На рис.18 показана типичная эволюция состояния нейронной сети в процессе итерационного решения на различных стадиях для 5-трекового события.

Было проведено изучение работы метода сегментов при увеличении множественности $N_{\text{track}} \times N_{\text{signals/track}} = 3{\times}8$, $5{\times}10$ и $10{\times}15$. Время сходимости оказалось линейно зависимым от числа активных нейронов $\widetilde{N}$ (см. рис.17).

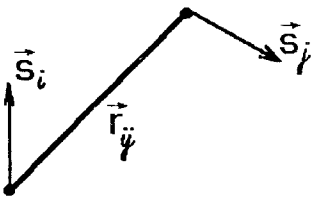


Рис.19. Роторная модель

Замечена чувствительность такой сети к наличию шумовых отсчетов. Для преодоления этого недостатка было предложено увеличивать вес тех нейронов-сегментов, вдоль которых есть дополнительные «подтверждающие» экспериментальные точки [35].

Роторная модель. Дальнейшим развитием модели сегментов можно считать роторную модель [59,62] нейронной сети.

В этой модели (рис.19) нейроны представляются в виде роторов (единичных векторов) s_i. Таким образом, нейроны характеризуются координатой (они располагаются в экспериментальных точках) и направлением (оно отражает касательную к треку в данной точке). Динамической переменной является угол. Эти роторы взаимодействуют друг с другом (по принципу некоторой близости) и с вектором r_{ij}, соединяющим их. Для выстраивания роторов в треки предлагается следующая энергетическая функция:

$$E = -\frac{1}{2}\sum_{ij}\frac{s_i s_j}{|r_{ij}|^m} - \frac{\alpha}{2}\sum_{ij}\frac{(s_i r_{ij})^2}{|r_{ij}|^m}, \qquad (29)$$

где первый член выстраивает роторы параллельно друг другу, а второй член поворачивает их вдоль трека. Множитель $|r_{ij}|^{-m}$ устанавливает локальность взаимодействия. Баланс этих двух влияний устанавливается множителем Лагранжа α.

Применение теории среднего поля дает следующие уравнения динамики:

$$v_i = \frac{-\frac{\partial E}{\partial v_i} I_1\left(\left|-\frac{\partial E}{\partial v_i}\right|/T\right)}{\left|-\frac{\partial E}{\partial v_i}\right| I_0\left(\left|-\frac{\partial E}{\partial v_i}\right|/T\right)}, \qquad (30)$$

где $v_i = \langle s_i \rangle_T$ — переменные среднего поля, I_1, I_0 — функции Бесселя. Заметим, что v_i уже не являются единичными векторами: область их изменения $0 \le |v_i| \le 1$. Длина нейрона v_i выражает вероятность его принадлежности треку. Это связано с тем, что в теории среднего поля с ненулевой температурой возникает зависимость длины вектора среднего спина v_i от величины локального поля $-\frac{\partial E}{\partial v_i}$ в i-м узле, которая

максимальна для точек, расположенных на треке, и сильно спадает при удалении от него.

Основным достоинством роторной модели является сведение количества динамических переменных до $2N$.

Модифицированная роторная модель. Модифицированная роторная модель нейронной сети первоначально была разработана для восстановления треков в пропорциональных камерах [63,64].

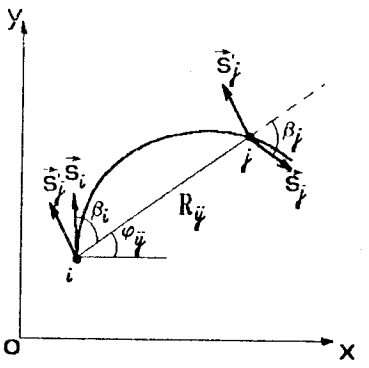


Рис.20. Модифицированная роторная модель

Естественным следствием приведенной выше модели функционирования пропорциональной камеры (см. рис.12) является роторная модель нейронной сети, в которой нейрон характеризуется значением, координатой и наклоном. Предлагается следующая модификация роторной модели. В качестве исходной предпосылки примем, что сигналы от частицы достаточно хорошо ложатся на некоторую окружность. А нейронная сеть в процессе работы должна расположить векторы $v_i = \langle s_i \rangle_T$ по касательной к этой окружности и сильно уменьшить v_j для тех j, которые не лежат на какой-либо истинной траектории частицы.

Вначале напишем двухчастичную энергию взаимодействия. Пусть через две точки i, j, проходит окружность (см. рис.20), тогда касательные векторы в них связаны соотношением

$$\beta_i = -\beta_j, \tag{31}$$

где β_i, β_j — углы между ними и хордой R_{ij}. Таким образом, если мы возьмем энергетическую функцию вида

$$E_{ij} = -s_i s'_j, \tag{32}$$

где s'_j — вектор s_j, отраженный относительно хорды R_{ij}, то она будет иметь минимум для точек, лежащих на одном треке. Отметим, что использование скалярного произведения делает поля от разных нейронов аддитивными.

Вектор s'_j проще всего вычислить из следующих геометрических соображений: для того чтобы отразить s_j относительно R_{ij}, достаточно повернуть R_{ij} на угол $-\varphi_{ij}$, чтобы он совпал с осью OX, затем произвести отра-

жение относительно OX, а потом повернуть обратно на угол $+\varphi_{ij}$. В матричной форме это преобразование запишется следующим образом:

$$T_{ij} = \begin{pmatrix} \cos(2\varphi_{ij}) & \sin(2\varphi_{ij}) \\ \sin(2\varphi_{ij}) & -\cos(2\varphi_{ij}) \end{pmatrix}. \tag{33}$$

Окончательно энергетическая функция принимает вид, значительно более простой по сравнению с (26) и (29):

$$E = -\frac{1}{2}\sum_{i,j} \mathrm{s}_i T_{ij} \mathrm{s}_j. \tag{34}$$

Для дальнейшего рассмотрения удобно ввести поле h_i^j, созданное в точке нейрона i нейроном j. Очевидно, поле, созданное всеми нейронами в этой точке, равняется сумме полей от каждого нейрона:

$$\mathrm{H}_i = \sum_j \mathrm{h}_i^j = \sum_j T_{ij} \mathrm{s}_j. \tag{35}$$

В новых обозначениях наша задача формулируется так:

$$E = -\frac{1}{2}\sum \mathrm{H}_i \mathrm{s}_i \to \min. \tag{36}$$

Перейдем к формулировке уравнений динамики. В первую очередь опишем эволюцию угла наклона нейрона. Она оказывается такой же, как и в роторной модели:

$$\mathrm{v}_i^{(m+1)} = \frac{\mathrm{H}_i^m}{|\mathrm{H}_i^m|} \frac{I_1(|\mathrm{H}_i^m|/T)}{I_0(|\mathrm{H}_i^m|/T)}. \tag{37}$$

Причем при определении поля здесь учитываются только допустимые с точки зрения дискретной структуры связи между нейронами.

Кроме изменения угла наклона нейрона, дискретная структура детектора позволяет нам двигать положение нейрона в пределах ширины ячейки относительно центра кластера (см. рис.12). В нашем случае изменение положения нейрона можно производить при нулевой температуре, т.е. скачком перемещать нейрон в точку максимального поля или в крайнюю допустимую точку, если максимум поля лежит вне допустимого интервала положений нейрона.

Программа, реализующая модифицированную роторную модель нейронной сети, была протестирована на реальных трехтрековых событиях, полученных на спектрометре АРЕС во время экспериментов по поиску запрещенного распада $\mu^+ \to e^+e^+e^-$ [57] и исследованию редкого распада $\pi^+ \to e^+\nu_e e^+e^-$ [58]. В этом эксперименте использовались только 10 камер. Это значит, что каждый трек в среднем содержит 10 экспериментальных точек (кластеров).

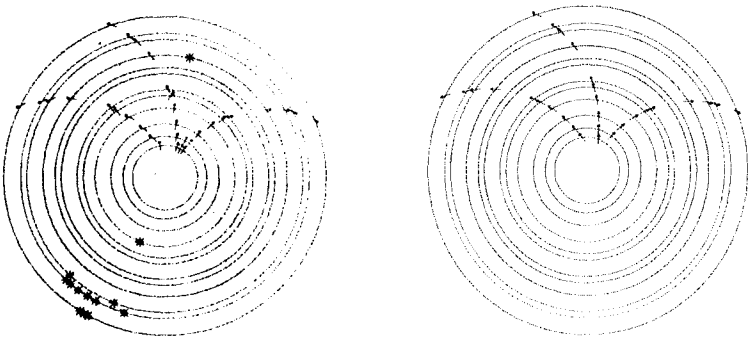


Рис.21. Пример работы нейронной сети: начальное и конечное состояния

На рис.21 приведены начальное состояние нейронной сети и результат ее работы. Отметим, что построение удачного начального состояния было возможно благодаря применению на предварительном этапе обработки клеточного автомата, который отсеял шумовые точки (на рисунке помечены крестиками) и произвел группировку экспериментальных точек по принципу их возможностей принадлежности разным трекам. Результат работы нейронной сети показывает также ее хорошую работу в случае близко лежащих треков.

В среднем нейронная сеть решает задачу за 4—5 итераций. Такая высокая степень сходимости стала возможной благодаря предварительному применению клеточного автомата, как уже отмечалось выше, и учету дискретных особенностей детектора. Эффективность работы нейронной сети оценивалась просмотром по нескольким выборкам и составляет 98%.

Анализ работы нейронной сети показывает, что компьютерное время поиска глобального минимума энергетической функции зависит от двух главных факторов:

— способности алгоритма ИНС избежать «сваливания» в один из локальных минимумов;

— числа степеней свободы ИНС, которое для ИНС с N входными сигналами в стандартных алгоритмах поиска пропорционально N^2.

Среди методов, позволяющих преодолеть первый фактор, наиболее известным является процедура имитационного отжига, описанная выше и потребляющая значительное количество машинного времени. В этой связи в [65] был использован иной подход: предлагалось так задать начальную конфигурацию роторов ИНС, чтобы оказаться в окрестности глобального минимума. Один из способов уже упоминался — применение клеточного автомата как для сокращения числа входных данных путем

их фильтрации, так и для частичного объединения отдельных отсчетов в треки по их близости. В то же время вышеприведенная роторная модель была слишком ориентирована на специфику пропорциональных камер и несомненно нуждалась в доработке для учета эффектов, связанных с продвижением в область энергии первичных пучков порядка нескольких ТэВ, связанную с большой множественностью и пересечением близких треков под малыми углами. Кроме того, имелись указания на повышенную чувствительность таких моделей ИНС к шуму [35].

В этой связи было необходимо исследовать следующие проблемы извлечения трековой информации с помощью ИНС:

— создание оптимальной начальной конфигурации ИНС с помощью алгоритма, достаточно общего для применений как при наличии, так и в отсутствие магнитного поля;

— робастность по отношению к сильно зашумленным входным данным (отношение сигнал/шум вплоть до 100%);

— стабильность к растущей множественности событий.

Подобный и, по-видимому, более общий подход был рассмотрен в [32], где авторы предложили комбинацию локального преобразования Хафа и метода деформируемых образцов. Однако авторы работы [65] предложили свои нововведения почти на всех этапах ИНС-алгоритма.

1. На стадии начального формирования роторов предложено определять их исходные направления в каждой экспериментальной точке по пику специальной угловой гистограммы, а длину ротора — по относительному числу точек, попавших в этот пик.

В принципе, это могло дать хорошую возможность создания деформируемого образца (кандидата в треки) и оценку их общего количества, но авторы предоставили самой ИНС развить эту начальную информацию для определения наиболее существенных связей.

2. В качестве своеобразного фактора отталкивания посторонних и шумовых нейронов предложено домножать веса T_{ij} на специальные робастные множители, взятые из теории робастных M-статистик [66] типа

$$w(t) = \begin{cases} \left(1 - \left(\dfrac{t}{C_T}\right)^2\right)^2 & |t| \le C_T; \\ 0 & |t| > C_T, \end{cases}$$

где $t = \widehat{v_i v_j'}$, $\mathbf{v}_j' = T_{ij}\,\mathbf{v}_j$ (см. (33), (35), (37)), $C_T = 2°$.

3. Процедура имитационного отжига заменена на ИНС-динамику (36), (37) с оптимально подобранной температурой $T = 1{,}5$.

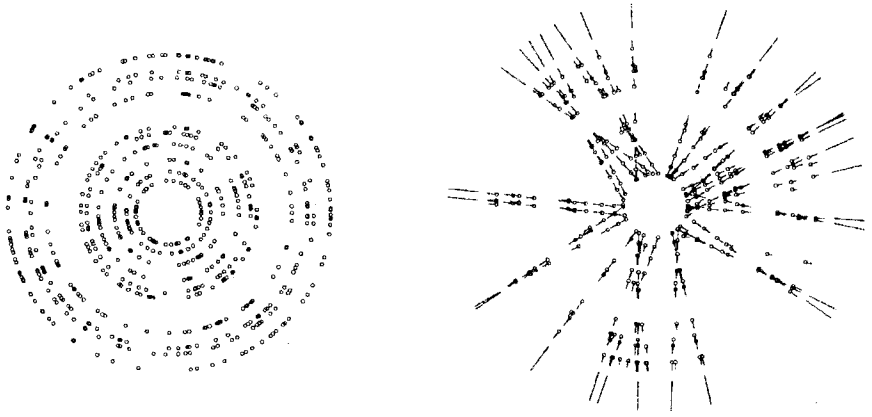


Рис.22. Фильтрация нейронной сетью 30-трекового события со 100% шумом

Этот подход применим как к прямым, так и к дуговым трекам, и был детально тестирован в [65] на использованной еще в [32] модели с прямыми треками, радиально выходящими из мишени. Модель была усложнена тем, что каждый трек выходил из своей «вершины» в мишени, что дало пересечение треков под очень малыми углами. Для большей близости к реальным событиям авторы [65] придерживались параметров цилиндрического спектрометра APEC. Множественность менялась от 10 до 50 треков (при большей множественности треки начинали сливаться из-за сплошной засветки камер). В данные после дискретизации вводились шумовые точки в количестве, равном числу «полезных» отсчетов (100% отношение «сигнал/шум»).

Таким образом, нейронная сеть применялась в качестве фильтра шумовых отсчетов. Пример такого 30-трекового события показан слева на рис.22. В этом событии имеется 300 точек, соответствующих трекам, и 300 шумовых точек (по 10 точек на трек). Результат применения нейронной сети показан справа. Здесь нейроны, отброшенные как шумовые, не показаны, а активные нейроны представлены векторами с длиной, пропорциональной их активационному уровню. Внешние линии указывают моделированные прямые треки. В данном случае 251 шумовая точка была отброшена сетью, а оставшиеся 49 шумовых точек были присоединены к трековым группам (в среднем 1,6 точек на трек).

Как показала статистика из 1000 модельных событий, среднее число итераций возрастало от 5 до 20 с ростом множественности от 10 до 50 (хотя разброс этого числа был достаточно велик в зависимости от случайных сгущений треков).

Надежность такой ИНС как фильтра оказалось 100%, т.е. предложенный критерий определения глобального минимума

$$\max \mid v_i^{(m+1)} - v_i^{(m)} \mid < 0{,}05,$$

где m — номер очередной итерации, позволяет включить все модельные треки в множество треков-кандидатов, исключив подавляющее большинство шумовых точек (до 85% для множественности 30). Небольшое число фиктивных треков, образованных из фрагментов близко лежащих треков с добавлением шумовых точек, легко устранялось при последующем спрямлении полученных треков, наряду и с 1—2 посторонними точками, «прилипшими» к реальным трекам.

6. ЗАКЛЮЧЕНИЕ

В работе над обзором авторы наряду с немногочисленными обзорами, имевшимися в отечественной литературе [37,67,68], использовали главным образом зарубежные источники, среди которых можно отметить известные вводные обзоры первого тома журнала «Neural Networks» Т.Кохонена, С.Гроссберга и Р.Липпмана [16,69,70], а также фундаментальную монографию Т.Ханна [6]. Естественно, авторы не могут претендовать на включение в этот достаточно краткий обзор всей массы работ по искусственным нейронным сетям. В целом следует отметить, что указанную во введении мультидисциплинарность в подходе к ИНС определяют три направления исследований: математическое, физическое и технологическое. Им соответствуют три группы публикаций, из которых самой обширной является группа с технологической направленностью. Помимо сведений, приведенных в разд.4, можно еще указать на подробный русскоязычный обзор технологических и программных разработок ИНС по пятнадцати североамериканским корпорациям, содержащийся в статье Широкова В.Ф. в сборнике [68]. Информацию об аппаратных параллельных реализациях клеточных автоматов можно найти в книгах [54,71].

В настоящем обзоре, может быть, недостаточно внимания уделено литературе, посвященной физической трактовке ИНС как больших термодинамических систем, хотя плодотворность использования теории среднего поля, дающей эффективное решение задачи нахождения глобального минимума в модели Хопфилда, подчеркнута в разд.3 достаточно отчетливо. К этой же группе работ по физической трактовке ИНС следует, по-видимому, отнести работы, использующие метод имитационного отжига [31,72].

Основное внимание в настоящем обзоре уделено литературе по математическим моделям, возможно, в силу того, что они дали наибольшее

число удачных приложений ИНС в обработке данных физических экспериментов, в частности, в физике высоких энергий. Обещающими тенденциями в физических приложениях математических моделей ИНС, помимо быстро прогрессирующих возможностей применения нейрочипов с заранее разработанной логикой в различных триггерных системах [44]—[49], являются работы [32,35,73], развивающие метод деформируемых образцов и эластичной руки, а также «динамический» персептрон [74], в котором статическое определение характеристических функций множеств, определяющих область действия нейронов, заменено на динамическое. Такая изменяемая топология позволяет успешно применять этот динамический персептрон для быстрого распознавания трековой информации. Ряд полезных рекомендаций и готовых программ для нейровычислений дан в книге [75]. В последнее время в библиотеке CPC появилась также ИНС программа-классификатор [76].

С точки зрения физических приложений ИНС, на авторов большое влияние оказали работы, выполненные рядом автором совместно с К.Петерсоном [59,32], которому мы выражаем благодарность за внимание, полезные обсуждения и присылку свежих оттисков своих статей. Нам также приятно поблагодарить проф. Л.Занелло, предоставившую нам копии многих работ, использованных нами в данном обзоре.

СПИСОК ЛИТЕРАТУРЫ

1. **Scientific American**, 1979, vol.241, p.3. Русский перевод: Мозг. М.: Мир, 1984.
2. **McCulloch W.S., Pitts W.H.** — Bull. Math. Biophys. 1943, vol.5, p.115.
3. **Hebb D.O.** — The organization of Behavior. N.Y., 1949.
4. **Кохонен Т.** — Ассоциативные запоминающие устройства. М.: Мир, 1982.
5. **Sutton R., Barto A.** — Psychol. Rev., 1981, vol.88, p.135.
6. **Khanna T.** — Foundation of Neural Networks. N.Y.: Addison-Wesley, 1989.
7. **Widrow B., Hoff M.E.** — Adaptive Switching Circuits. IRE WESTON Convention Record, 1960, 4, p.96.
8. **Rosenblatt F.** — Psychol. Rev., 1958, vol.65, p.386.
9. **Rosenblatt F.** — Principles of Neurodynamics. Washington, D.C.: Spartan, 1962. **Розенблатт Ф.** — Принципы нейродинамики. М.: Мир, 1965.
10. **Hopfield J.J.** — Proc. Nat. Acad. Sci. USA, 1982, vol.79, p.2554.
11. **Hopfield J.J.** — Proc. Nat. Acad. Sci. USA, 1984, vol.81, p.3088.
12. **Hopfield J.J.** — Proc. Nat. Acad. Sci. USA, 1987, vol.84, p.8429.
13. **Amari S.** — IEEE Trans. Syst., Man. Cybern., 1983, vol.13, p.741.
14. **Anderson J.A.** — IEEE Trans. Syst., Man. Cybern., 1983, vol.13, p.799.
15. **Kohonen T.** — Self-organization and Associative Memory. 3-d Edition. Berlin: Springer-Verlag, 1990.
16. **Kohonen T.** — Neural Networks, 1988, vol.1, p.3.
17. **Grossberg S.** — Math. Biosci., 1969, vol.4, p.201.
18. **Grossberg S.** — J. Stat. Phys., 1971, vol.3, p.95.
19. **Grossberg S.** — J. Stat. Phys., 1969, vol.1, p.319.

20. Carpenter G.A. — J. Diff. Eqns., 1977, vol.23, p.335.
21. Carpenter G.A. — J. Math. Anal. Appl., 1977, vol.58, p.152.
22. Grossberg S., Ellias A. — Biol. Cybern., 1975, vol.20, p.69.
23. Hopfield J.J., Tank D.W. — Biol. Cybern., 1985, vol.52, p.141; Scince, 1986, vol.233, p.625.
24. Kinzel W. — Z. Phys. B, 1985, vol.60, p.205.
25. Marks II R.J., Oh S., Atlas L.E. — IEEE Trans. Circuits Syst., 1989, vol.36, p.846.
26. Verleysen M. et al. — IEEE Trans. Circuits Syst., 1989, vol.36, p.762.
27. Palm G. — In: Brain Theory (Eds. Palm G., Aertsen A.) Berlin: Springer-Verlag, 1986, p.211. Palm G. — Science, 1987, vol.235, p.1227.
28. Bruce A.D., Gardner E.J., Wallace D.J. — J. Phys., 1987, vol.A20, p.2909.
29. Kirkpatrik S., Sherrington D. — Phys. Rev., 1978, vol.B17, p.4384.
30. Bogolubov N.N., Jr — A Method for Studying Model Hamiltonians. Oxford — N.Y.: Pergamon Press, 1972. Боголюбов Н.Н. (мл.) — Метод исследования модельных гамильтонианов. М.: Наука, 1974.
31. Peterson C. — Neural Computation, 1990, vol.2, p.261.
32. Ohlsson M., Peterson C., Yuille A.L. — Comput. Phys. Commun., 1992, vol.71, p.77.
33. Huber P.J. — Robust Statistics. John Wiley and Sons, N.Y., 1981.
34. Durbin R., Willshaw D. — Nature, 1987, vol.326, p.689.
35. Gyulassy M., Harlander H. — Comput. Phys. Commun., 1991, vol.66, p.31.
36. Lonnblad L. et al. — Preprint LU TP 91-4, Lund, 1991.
37. Ачасова С.М. — Программирование, 1992, № 2, с.40.
38. Thakoor A.P. et al. — Appl. Opt., 1987, vol.26, p.5085.
39. Abu-Mostafa Y., Psaltis D. — Scientific American, 1987, vol.256, p.66.
40. Farhat N. et al. — Appl. Opt., 1985, vol.24, p.1469.
41. Psaltis D., Farhat N. — Opt. Lett., 1985, vol.10, p.98.
42. Proc. of the Int. Conf. on Comp. in High Energy Physics. CERN 92-07, Geneve, 1992.
43. Stimpfl-Abele G. — Ibid. p.642.
44. Denby B. et al. — Ibid., p.674.
45. Handler T., Neis E. — Ibid., p.650.
46. Andree H. — Ibid., p.654.
47. Propriol J. et al. — Ibid., p.652.
48. Cosmo G. et al. — Ibid., p.665.
49. Innocente V. et al. — Ibid., p.669.
50. Ososkov G.A. — Proc. Int. Conf. on Probability and Math. Statistics PROBASTAT '91. Bratislava: Univ. Press, 1991, p.353.
51. Grote M. — Preprint CERN DD/87/3, 1987.
52. Иванов В.В. и др. — Препринт ОИЯИ Р10-92-156, Дубна, 1992.
53. Будагов Ю.А. и др. — Сообщение ОИЯИ Р10-93-140, Дубна, 1993.
54. Wolfram (ed.) — Theory and Applications of Cellular Automata. World Scientific, 1986.
55. Gardner M. — Mathematical Games. Scientific American, 1970, vol.223, p.4.
56. Glazov A.A., Kisel I.V., Konotopskaya E.V., Ososkov G.A. — JINR Commun. E10-91-507, Dubna, 1991.
57. Baranov V.A. et al. — J. Phys. G.: Nucl. Part. Phys., 1991, vol.17, p.S57—S70.
58. Баранов В.А. и др. — ЯФ, 1992, т.55, вып.11, с.2940.
59. Peterson C. — Nucl. Instr. and Meth., 1986, vol.A279, p.537.
60. Denby B. — Comput. Phys. Commun., 1988, vol.49, p.429.
61. Peterson C., Soderberg B. — Int. J. of Neural Syst., 1989, vol.1, p.3.
62. Peterson C. — Lund Preprint LU TP 90-6, 1990.
63. Glazov A.A., Kisel I.V., Konotopskaya E.V. et al. — JINR Commun., E10-92-352, Dubna, 1992.
64. Kisel I.V., Ososkov G.A. — In: CHEP92 — Conf. on Computing in High Energy Physics, CERN, Annecy, 1992, p.646.

65. **Baginyan S., Kisel I., Konotopskaya, Ososkov G.** — JINR Commun. E10-93-86, Dubna, 1993.
66. **Ososkov G.** — In: Proc. 2-d Int. Tampere Conf. in Statistics. Tampere Univ. Press. 1987, p.615.
67. **Автоматы.** М.: Мир, 1956.
68. **Итоги науки и техники.** М.: ВИНИТИ, 1990, т.1,2.
69. **Grossberg S.** — Neural Networks, 1988, vol.1, p.17.
70. **Lippman R.** — IEEE ASSP Magazine, 1987, vol.4, p.4.
71. **Toffoli T., Margolus N.** — Cellular Automata Machines: A New Environment for Modelling. MIT Press, Cambridge, Mass, 1987.
72. **Kirkpatrik S., Gelatt C.D., Vecchi M.P.** — Science, 1983, vol.220, p.671.
73. **Yuille A.L.** — Neural Computation, 1990, vol.2, p.1.
74. **Perrone A. et al.** — SPIE Proc. Series, 1992, Washington, vol.1711, p.470.
75. **Necht-Nielsen R.** — Neurocomputing.: Addison-Wesley, N.Y., 1990.
76. **Cherubino A. et al.** — Comp. Phys. Commun., 1992, vol.72, p.79.