Глава
5

Рекуррентные сети: Ассоциативная память

Сеть Хопфилда и спиновые стекла. Энергия и динамика сети. Ассоциативная память: запись и воспроизведение. Емкость памяти: термодинамический подход. Чувствительность к огрублениям и повреждениям связей. Повышение емкости памяти: разобучение. Запоминание последовательностей образов. Сеть Хопфилда с точки зрения теории информации. Выделение прототипов и предсказание новых классов.

> 🕮 Самое худшее случилось - сказал сэр Дональд Акер, когда в Массачусетском Технологическом Институте соединили входы с выходами
> **С.Лем, *Не буду прислуживать***

Мы уже познакомились с сетями, обучаемыми с учителем, задающим образцы правильных ответов, и обучаемыми без учителя, которые адаптируют свою структуру к данным не требуя дополнительной информации о принадлежности их к тому или иному классу. Однако до сих пор мы ограничивались сетями без обратных связей. Такие сети, будучи обучены, выдают ответ сразу после прохождения черех них входного сигнала. Каждый нейрон, при этом срабатывает лишь однажды. Соответственно, достаточно глубокая, многостадийная обработка данных подразумевает наличие многих слоев, что усложняет обучение. Естественным обобщением таких однопроходных схем служат т.н. *рекуррентные* сети, выходы которых возвращаются обратно на их входы. Тем самым, информация пропускается через одну и ту же сеть многократно.

Новое качество, присущее рекуррентным сетям, - динамическая обработка информации.

Одной из наиболее известных моделей такого рода, которая оказала важнейшее воздействие на возрождение интереса к нейронным сетям в восьмидесятые годы, является сеть Хопфилда. В данной главе мы рассмотрим структуру и свойства этой сети, делающие ее таким привлекательным объектом как теоретических, так и прикладных исследований.

Исторический поворот в 1982 году

В 1982 году в докладах Американской академии наук была опубликована статья американского физика, специалиста в области физики твердого тела из Калифорнийского Технологического Института, Джона Хопфилда (Hopfield, 1982a). С этой работы начался бурный процесс возрождения интереса к искусственным нейронным сетям, на который так негативно повлияла в конце шестидесятых книга Минского и Пейперта. В работе Хопфилда впервые было обращено внимание на аналогию, которая существует между сетями с симметричными связями и давно известными физикам объектами - спиновыми стеклами. Кроме того, стало ясно, что такие сети служат прекрасной основой для построения моделей содержательно-адресованной памяти. И наконец, обнаружилось, что нейронные сети могут быть успешно исследованы с помощью

методов теоретической физики, в частности, статистической механики. Результатом этого обстоятельства явилось массовое внедрение физиков и физических методов в эту новую область знания.[1]

Спиновые стекла

В кристаллической решетке атомы, обладающие магнитными моментами, могут взаимодействовать друг с другом различными способами. Если связи между моментами таковы, что стремятся сориентировать их параллельно, то в основном состоянии (состоянии минимальной энергии) все атомы в решетке ориентируют свои моменты параллельно. Такие вещества называются ферромагнетиками. Связи между атомами описываются при этом одинаковыми положительными числами и называются также ферромагнитными. Если, напротив, все связи отрицательны, то такие вещества называются антиферромагетиками. В антиферромагнетиках соседние спины ориентируются в противоположных направлениях. А вот если связи между магнитными моментами атомов имеют *случайные* значения знаков, то соответствующие системы называются спиновыми стеклами (см. Рисунок 1). Основная особенность системы связей в спиновых стеклах такова, что система в целом оказывается *фрустрированной.*

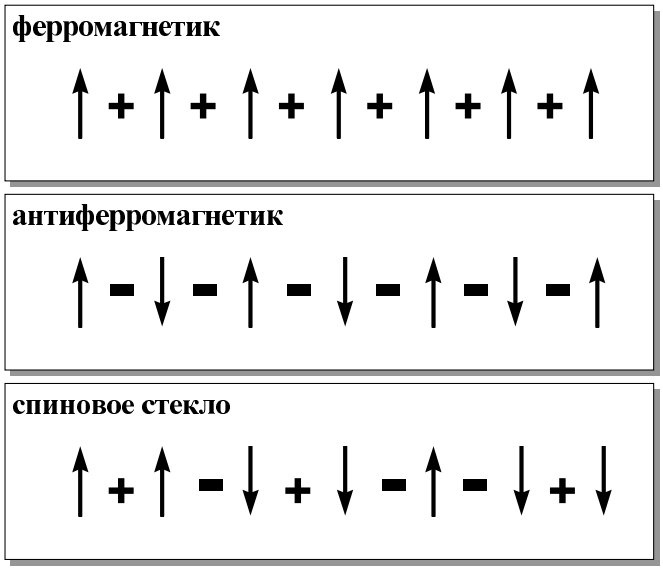

Рисунок 1. Знаки связей между спинами в ферромагнетике, антиферромагнетике и спиновом стекле

Фрустрация (“разочарование”) означает, что как бы ни сориентировались отдельные магнитные моменты атомов в спиновом стекле, всегда найдутся такие пары из них, в которых взаимодействие вносит положительный (разочаровывающий) вклад в энергию состояния (см. Рисунок 2).

[1] Вообще говоря, еще задолго до этого, в 1954г. Крэгг и Темперли указали на аналогию между стационарной активностью нейронных сетей и коллективными состояниями в системах магнитных диполей, а в 1974 году Литтл также провел аналогию между нейронными сетями и спиновыми системами и указал на аналогию шума и температуры. Но ряд обстоятельств, в частности, связанных с другим характером динамики нейронов, другим типом возникающих в сети аттракторов и, главным образом, недостаточно четкая физическая аналогия, не позволили этим исследователям оказать на развитие теории нейронных сетей того влияния, какое оказала на него работа Хопфилда.

Фрустрированность системы обусловливает огромное вырождение ее основного состояния. Спиновое стекло может "*замерзнуть*" в любом из возможных основных состояний системы, отличающемся от множества других аналогичных состояний с практически такой же энергией лишь конфигурацией системы магнитных моментов. Хопфилд предположил, что аналогичное явление может лежать в основе существования огромного числа состояний памяти, характерного для мозга. Действительно, можно рассмотреть модель полносвязной нейронной сети с рекуррентными симметричными связями между нейронами. В такой модели возбуждающим связям будут соответствовать ферромагнитные связи в спиновом стекле, а тормозным - антиферромагнитные связи.

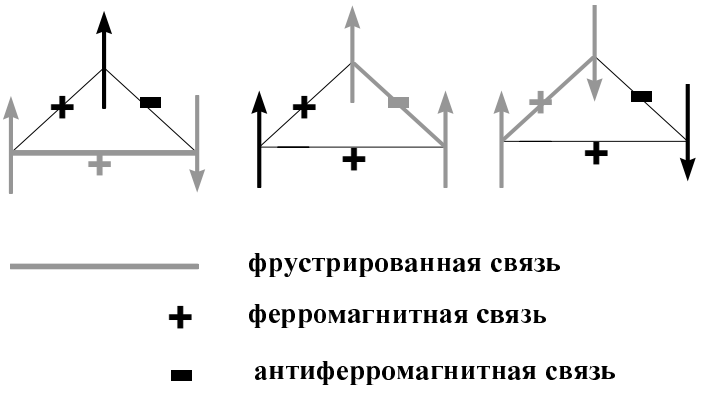

Рисунок 2. Фрустрированная система трех взаимодействующих спинов. При любых их ориентациях всегда находится такая связь , знак которой противоречит взаимной ориентации пары, что приводит к нежелательному положительному вкладу в полную энергию системы.

Подобно спиновым стеклам, такие сети будут иметь множество стационарных конфигураций активностей нейронов, являющихся *аттракторами* (от англ. *attract* - притягивать), т.е. такими состояниями, к которым сходится динамика нейросети. Именно введенная Хопфилдом динамика изменений состояний нейронов наряду с симметричностью связей между нейронами определили новизну описываемой модели.

Сеть Хопфилда как ассоциативная память

Симметричность связей

В Хопфилдовской сети матрица связей между нейронами $\mathbf{w}$ является полной и симметричной ($w_{ij} = w_{ji}$) а самовоздействие нейронов считается отсутствующим ($w_{ii} = 0$). Подобные свойства определяют тесную связь модели со спиновыми стеклами. Критики отмечают, что подобная ориентация на физические системы делает модель несостоятельной с физиологической точки зрения (хотя в мозге существуют некоторые структурные единицы - колонки, связи между нейронами в которых не так уж далеки от симметричных). Однако, самое

главное в таком подходе то, что простота архитектуры сети облегчает имитацию с ее помощью богатого спектра явлений, которые могут быть соотнесены с реальными свойствами мозга.

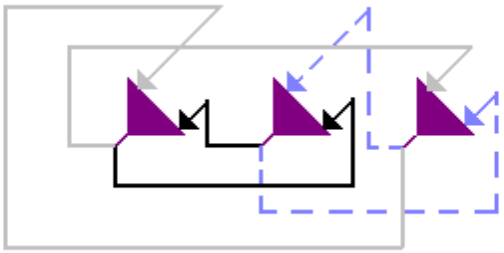

Рисунок 3. Архитектура сети Хопфилда. Связи с одинаковым весом обозначены одинаковыми линиями. Матрица соединений полносвязанная и симметричная. Самовоздействие нейронов отсутствует.

Асинхронная динамика

Нейроны в модели Хопфилда, подобно спиновым переменным, могут принимать два состояния $s_i \in \{-1,+1\}$, а динамика состояний сети носит *асинхронный* характер (т.н. *Глауберова динамика*). В дискретные моменты времени $t=1,2,...$ случайным образом выбирается *один* нейрон (*k-ый*) для которого вычисляется значение *потенциала*

$$h_k = \sum_j w_{kj} s_j$$

При выполнении условия $h_k s_k < 0$ состояние нейрона изменяется на противоположное: $s_k \to -s_k$.

В другом варианте - *последовательной динамике* - перебор нейронов производится не случайным образом а циклически, но в каждый момент времени также может изменяться состояние лишь одного нейрона. Эти два варианта качественно отличаются от *параллельной динамики*, подразумевающей одновременное изменение состояний всех тех нейронов, для которых выполняется условие $h_k s_k < 0$ (такова, например, динамика модели Литтла). Синхронизация моментов обновления состояний нейронов делает такую динамику подверженной "зацикливаниям".

В отличие от многослойных сетей, в которых входные и выходные нейроны пространственно разделены в модели Хопфилда все нейроны одновременно являются и входными, и скрытыми, и выходными. Роль входа в таких сетях выполняет начальная конфигурация активностей нейронов, а роль выхода - конечная стационарная конфигурация их активностей.

Метрика пространства состоянний

Расстояние между состояниями сети можно измерять в т.н. *метрике Хэмминга*. Если два вектора $\mathbf{b}^1$ и $\mathbf{b}^2$ бинарные, то Хэммингово расстояние между ними определяется как количество различающихся компонент. Так, если векторы имеют вид $\mathbf{b}^1$=(1,0,0,0,1) и

$\mathbf{b}^2$=(1,1,0,0,0), то Хэммингово расстояние между ними $\|\mathbf{b}^1 - \mathbf{b}^2\|$ будет равно двум, поскольку в точности две компоненты этих векторов (вторая и пятая) имеют различные значения. Формально, Хэммингово расстояние для таких (*Булевых*) векторов может быть определено как

$$\|\mathbf{b}^1 - \mathbf{b}^2\| = \sum_i (b_i^1 - b_i^2)^2$$

В случае *спиновых* переменных, $s_i^{1,2} = 2b_i^{1,2} - 1$, принимающих значения ±1, расстояние Хэмминга может быть переписано в виде

$$\|\mathbf{s}^1 - \mathbf{s}^2\| = \frac{1}{2}\left(N - \sum_i s_i^1 s_i^2 \right) = \frac{1}{2}\left(N - \mathbf{s}^1\mathbf{s}^2 \right)$$

где $\mathbf{s}^1\mathbf{s}^2$ - скалярное произведение, или *перекрытие* между векторами $\mathbf{s}^1$ и $\mathbf{s}^2$. Таким образом, минимальное Хэммингово расстяние между векторами со спиновыми переменными соответствует максимальному перекрытию между ними.

Энергия состояния

Нетрудно показать, что описанная выше асинхронная динамика сети сопровождается уменьшением *энергии* сети, которая определяется следующим образом:

$$E = -1/2 \sum_{i,j} w_{ij} s_i s_j \ .$$

Действительно, при изменении состояния одного k-го нейрона его вклад в энергию изменяется с $E_k(t) = -s_k(t)\sum_{j\neq k} w_{kj} s_j(t) = -s_k(t)h_k(t)$ на $E_k(t+1) = -s_k(t+1)h_k(t)$.

Следовательно,

$$E_k(t+1) = -\operatorname{sgn}[h_k(t)]h_k(t) = -|h_k(t)| \le -s_k(t)h_k(t) = E_k(t) \ .$$

В случае, когда нейрон имеют ненулевые *пороги* активации ϑ_i, энергия состояния приобретает вид $E = -1/2 \sum_{,j} w_j s\ s_j + \sum_i \vartheta_i s_i$, но вышеприведенный вывод остается в силе.)

Поскольку число нейронов в сети конечно, функционал энергии ограничен снизу. Это означает, что эволюция состояния сети должна закончиться в стационарном состоянии, которому будет соответствовать локальный минимум энергии. В Хопфилдовской модели стационарные конфигурации активностей нейронов являются единственным типом аттракторов в пространстве состояний сети. Мы можем представить динамику сети, сопоставив ее состояние с шариком, движущимся с большим трением в сложном рельефе со множеством локальных

минимумов. Сами эти минимумы будут устойчивыми состояниями памяти, а окружающие точки на склонах - переходными состояниями.[2]

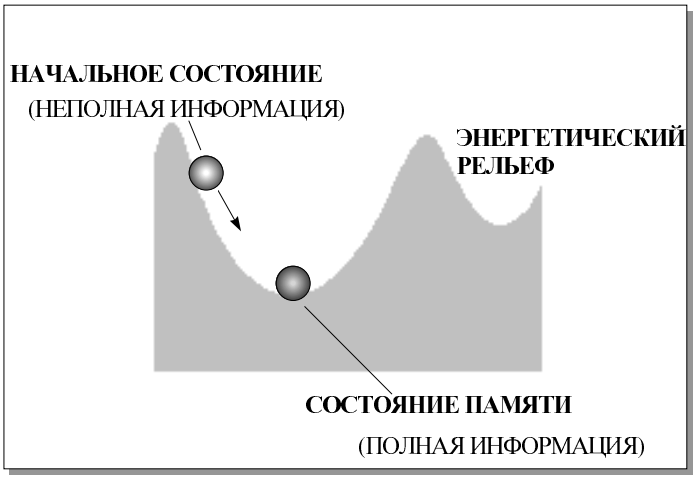

Рисунок 4. Поведение состояния в сети Хопфилда аналогично движению шарика, скатывающегося со склона в ближайшую лунку. Начальное состояние шарика соответствует вектору, содержащему неполную информацию об образе памяти, которому отвечает дно лунки.

Такая динамика определяет главное свойство сети Хопфилда - способность восстанавливать возмущенное состояние равновесия - "вспоминать" искаженные или потерянные биты информации. Восстановление полной информации по какой-либо ее части - вспоминание по ассоциации - наделяет модель Хопфилда свойством ассоциативной памяти. (Далее в этой главе мы продемонстрируем, и более общие возможности сети Хопфилда.)

Ассоциативная память

Аттракторами сети Хопфилда являются стационарные состояния. Если начальная конфигурация $\mathbf{s}$ мало отличается от одного из таких аттракторов сети $\mathbf{s}^*$ (т.е. $\|\mathbf{s}-\mathbf{s}^*\| << N$), то она быстро эволюционирует к этому ближайшему аттрактору, изменив состояния небольшого числа нейронов. Такой переход можно записать в виде $\mathbf{s} \to \mathbf{s}^*$. Можно проинтерпретировать это явление так, что состояние $\mathbf{s}$ содержит частичную, неполную информацию, которая, однако, достаточна для восстановление полной информации, кодируемой состоянием $\mathbf{s}^*$. Например, мы способны восстановить название города по неполному набору букв **В*нец*я.** Такая память, в которой информация ищется не по формальному адресу (подобно поиску книги в библиотеке по ее шифру), а на основе частичной информации о ее содержании, называется адресованной по содержанию. Таким образом модель Хопфилда может использоваться для имитации *содержательно-адресованной* или, иными словами, *ассоциативной* памяти.

[2] При появлении модели Хопфилда многие нейрофизиологи были смущены подобным применением понятия энергии к моделям нейронных сетей. Поэтому, иногда можно встретить более нейтральный термин - функция Ляпунова. В математике так называют функцию состояния динамической системы, которая меняется монотоно (не убывает или, напротив, не возрастает) в процессе эволюции системы

Важным свойством такой памяти, представленной набором аттракторов сети, является ее *распределеннсть*. Это означает, что все нейроны сети участвуют в кодировании всех состояний памяти. Поэтому небольшие искажения значений отдельных весов не сказываются на содержании памяти, что повышает устойчивость памяти к помехам.

Конечно, ассоциативная память может быть реализована и без использования нейронных сетей. Для достаточно с помощью обычного компьютера осуществить *последовательное* сравнение внешнего стимула со всеми предварительно запомненными образами, выбрав из них тот, для которого Хэммингово расстояние до входного сигнала минимально. Однако, сеть Хопфилда позволяет *исключить перебор состояний памяти* и осуществить эту процедуру *параллельным* способом, при котором время выборки из памяти не увеличивается с ростом числа запомненных образов.

Обучение сети. Правило Хебба

Описанная сеть действительно стала использоваться для моделирования ассоциативной памяти, поскольку уже в первой своей работе Хопфилд указал конструктивный метод построения синаптических связей между нейронами, который в некоторых случаях позволял запомнить любые заранее заданные состояния сети.

Например, полезной была бы сеть, аттракторы которой, соответствовали бы векторам, кодирующим бинарные изображения подписей различных людей на чеке. Поскольку практически невозможно одинаково расписаться дважды, подобная сеть была бы незаменима при распознавании подписи, несмотря на ее естественные вариации. Если число различных типов подписей, которые должна распознавать сеть, равно P и *образцы* в некотором смысле типичных, наиболее вероятных или усредненных подписей различных людей кодируются векторами σ^n, $n = 1,\ldots, P$, то желательно, чтобы именно эти векторы кодировали и аттракторы сети, которую мы собираемся использовать для классификации.

Хопфилд предложил использовать для решения этой задачи Хеббовское правило построения межнейронных связей.[3]

$$w_{ij} = \frac{1}{N}\sum_n \sigma_i^n \sigma_j^n, \quad i \neq j \ , \quad w_{ii} = 0, \ i, j=1,\ldots,N.$$

Это правило действительно гарантирует стационарность произвольно выбранных векторов σ^n в случае, когда их число P не превосходит примерно 5% от общего числа нейронов N. При больших значениях P некоторые из запоминаемых векторов σ^n теряют свойство стационарности, а при превышении некоторого критического значения - *емкости памяти* - ($P \cong 0.14N$) стационарные состояния сети теряют всякую связь с ними, и сеть переходит из режима запоминания в режим спинового стекла, для которого характерно наличие очень большого числа аттракторов, далеких от любых запоминаемых векторов. Эти свойства модели Хопфилда были открыты с использованием математического аппарата статистической физики. Заинтересованный читатель может ознакомиться с этим подходом более подробно в последней, дополнительной, главе этой книги.

Аттракторам, не совпадающим с векторами σ^n, часто присваиваются такие негативные названия, как *ложная* или *паразитная память, химеры, русалки* и даже *мусорная куча*. Подобное отношение вызвано тем, что при релаксации начального состояния сети в одно из состояний ложной памяти интерпретировать результат распознавания становится затруднительно. Однако само по себе появление таких непредвиденных аттракторов является замечательным свойством модели Хопфилда и свидетельствует о том, что она способна не просто на ассоциативную выборку запомненной информации, но также и на *синтез* новых образов. Можно сказать, что сеть активно преобразует исходную информацию, а не является пассивным хранилищем образов. Ниже мы покажем, как можно интерпретировать все аттракторы сети единым образом, и приведем примеры, когда т.н. ложная память играет позитивную роль.

Модель Крика - Митчисона. Разобучение

В 1983 году в журнале *Nature* одновременно появились две публикации (Hopfield, Feinstein & Palmer 1983 и Crick & Mitchison, 1983), в которых была описана процедура уменьшения доступа к состояниям ложной памяти и ее возможная биологическая интерпретация. Эта процедура, названная *разобучением*, применяется к уже обученной сети, в пространстве которой есть ложные состояния. Она предполагает многократное предъявление сети в качестве начальных состояний случайно сгенерированных векторов и прослеживание их эволюции вплоть до стационарного состояния σ^*, которое может принадлежать как истинной, так и ложной памяти. После этого связи в сети модифицируются следующим образом: $\delta w_{ij} = -\varepsilon \sigma_i^* \sigma_j^*, \quad i \neq j$, где $\varepsilon > 0$ - небольшая константа.

Хопфилд с коллегами установили, что применение такой процедуры к сети, обученной по правилу Хебба на наборе случайных векторов, приводит к увеличению и выравниванию доступности состояний, соответствующих запоминаемым образам, и снижению доступности состояний ложной памяти. Эти явления они объяснили тем, что в рассматриваемом случае состояниям ложной памяти соответствуют гораздо более "мелкие" энергетические минимумы, чем состояниям, соответствующим запоминаемым образом. Поэтому ложные состояния сильнее подвержены разобучению, которое выражается в "*закапывании*" энергетических минимумов, в которые попадает система. Выравнивание доступности состояний памяти объясняется тем, что состояния с большими областями притяжения чаще притягивают случайный стимул и их область притяжения уменьшается быстрее, чем у состояний с меньшими сферами притяжения.

Крик и Митчисон, кроме того, предположили, что процесс, аналогичный разобучению, происходит в мозгу человека и животных во время фазы быстрого (парадоксального) сна, для которого характерны фантастические сюжеты (составленные из аналогов ложных образов). В этот период кора головного мозга постоянно возбуждается случайными воздействиями ствола мозга, и возникающие картины далеки от тех, которые дает сенсорный опыт. Разобучение при этом эффективно приводит к забыванию подобных парадоксальных картин и к увеличению доступа к образам, соответствующим объектам внешнего мира. Гипотеза о роли быстрого сна была сформулирована Криком и Митчисоном в виде афоризма: "Мы грезим, чтобы забыть".

---

[3] Ранее мы определяли обучение по Хеббу как такое, при котором изменение веса w_{ij} пропорционально *j* - му входу и выходу *i*-го нейрона. В рекуррентной сети Хопфилда состояниие *j*-го нейрона как раз и является *j*-м входом для остальных нейронов.

Идея разобучения затем была развита другими исследователями. В одном из ее вариантов в качестве начальных состояний сети предъявляются не случайные стимулы, а зашумленные случайным шумом запоминаемые образы. При этом, помимо разобучения сети финальным аттрактором, она слегка подучивается запоминаемым образом σ^n

$$\delta w_{ij} = -\varepsilon(\sigma_i^* \sigma_j^* - \sigma_i^n \sigma_j^n), \quad i \neq j.$$

То есть, если образ памяти восстанавливается без ошибки, синаптические связи не модифицируются. Подобная модификация процедуры разобучения может существенно увеличить емкость памяти (с $P \cong 0.14N$ до $P \cong N$).

Активная память

Выделение сигнала из шума

Разобучение действительно улучшает запоминание *случайных* образов. Однако, например, для коррелированных образов доводы, приведенные в предыдущем разделе теряют свое значение. Действительно, если эти образы, например, являются слегка зашумленными вариантами одного образа-прототипа σ^{pro}. Нетрудно показать, что в этом случае единственной зеркальной парой аттракторов в сети с Хеббовскими связями окажется пара $\pm\sigma^{pro}$. Это означает, что вся память, которой обладает сеть, оказывается ложной. Отсюда следует, в частности, что состояниям ложной памяти далеко не всегда соответствуют неглубокие энергетические минимумы.

Этот пример показывает, что ложная память иногда не бесполезна, а преобразуя заучиваемые векторы, дает нам некоторую важную информацию о них. В данном случае сеть как бы очищает ее от случайного шума. Подобное явление характерно и для обработки информации человеком. В известном психологическом опыте людям предлагается запомнить изображения, каждое из которых представляет собой обязательно искаженный равносторонний треугольник. При контрольной проверке на значительно более широком наборе образов, содержащийся в них идеальный равносторонний треугольник опознается испытуемыми как ранее виденный. Такое явление называется *выработкой прототипа*. Именно эта аналогия использовалась нами при введении обозначения σ^{pro}.

Минимальный базис

Состояния ложной памяти могут иметь и другие, не менее интересные формы. Рассмотрим, например, вариант модели Хопфилда, в котором состояния нейронов принимают значения 0 или 1. Подобная модель легко переформулируется в оригинальную, для которой состояниями являются *спиновые* переменные ± 1, путем переопределения порогов. Мы, однако, будем считать, что в нашей сети пороги всех нейронов отрицательны и бесконечно малы. Иначе говоря, динамика состояния нейрона определяется соотношениями

$$v_i = \begin{cases} 1, & \sum_j w_{ij} v_j \geq 0 \\ 0, & \sum_j w_{ij} v_j < 0 \end{cases}$$

Рассмотрим следующий набор векторов:

$\mathbf{v}^1$ = (0, 0, **1**, **1**, **1**, 0, **1**), $\mathbf{v}^2$ = (0, **1**, 0, **1**, 0, 0, **1**), $\mathbf{v}^3$ = (**1**, 0, 0, **1**, 0, **1**, **1**),

который используем для построения Хеббовской матрицы связей

$$w_{ij} = \sum_{n=1}^{3}(2v_i^n - 1)(2v_j^n - 1);\ i \neq j;\quad w_{ii} = 0;\ i, j = 1,\ldots,7.$$

$$\mathbf{w} = \begin{Vmatrix} 0 & -1 & -1 & -1 & -1 & 3 & -1 \\ -1 & 0 & -1 & -1 & -1 & -1 & -1 \\ -1 & -1 & 0 & -1 & 3 & -1 & -1 \\ -1 & -1 & -1 & 0 & -1 & -1 & 3 \\ -1 & -1 & 3 & -1 & 0 & -1 & -1 \\ 3 & -1 & -1 & -1 & -1 & 0 & -1 \\ -1 & -1 & -1 & 3 & -1 & -1 & 0 \end{Vmatrix}$$

сети Хопфилда. Если найти все аттракторы этой сети (что нетрудно сделать в виду небольшой размерности пространства его состояний $2^7=128$), то обнаружится, что помимо векторов $\mathbf{v}$, $\mathbf{v}^2$, $\mathbf{v}^3$ стационарными являются состояния, описываемые векторами

$\mathbf{b}^1$ = (**1**, 0, 0, 0, 0, **1**, 0), $\mathbf{b}$ = (0, **1,** 0, 0, 0, 0, 0), $\mathbf{b}$ = (0, 0, **1**, 0, **1**, 0, 0), $\mathbf{b}^4$ = (0, 0, 0, **1**, 0, 0, **1**).

Векторы $\mathbf{b}^l$ сами по себе замечательны. Их единичные компоненты помечают *кооперированные нейроны*, то есть те из них, которые одновременно активны или одновременно пассивны во всех запоминаемых векторах $\mathbf{v}^n$. Если считать, что компоненты векторов $\mathbf{v}^n$ кодируют некоторые признаки, то кооперированность некоторых нейронов означает, что некоторые признаки избыточны и могут быть заменены одним. Например, если в нашем примере первый нейрон кодирует такое свойство, как пол, а шестой - наличие бороды, то практически со стопроцентной вероятностью они могут быть заменены одним нейроном, о чем сигнализирует вектор $\mathbf{b}^1$.

Векторы $\mathbf{b}$, кроме того, образуют так называемый минимальный базис. А именно, это минимальное число векторов, с помощью линейной комбинации которых могут быть представлены все запоминаемые векторы

$$\mathbf{v}^n = \sum_{l=1}^{L}\alpha_{nl}\mathbf{b}^l,\quad n = 1,\ldots,P.$$

Кроме того, все стационарные состояния сети, в Хеббовские связи которых записаны векторы $\mathbf{v}^n$, также обязательно должны разлагаться по векторам минимального базиса. Это означает, что если некоторые нейроны кооперированы в векторах $\mathbf{v}^n$, то они должны кооперироваться и во всех аттракторах сети.

Используя векторы минимального базиса можно получить новый вид недиагональных элементов Хеббовской матрицы связей

$$w_{ij}(\mathbf{v}) = P\sum_{l=1}^{L} b_i^l b_j^l + \sum_{k=1}^{L}\sum_{m=1}^{L} w_{km}(\alpha) b_i^k b_j^m ,$$

где

$$w_{km}(\alpha) = \sum_{n=1}^{P} (2\alpha_{nk} - 1)(2\alpha_{nm} - 1), \quad k \neq m; \quad w_{kk}(\alpha) = 0 .$$

С помощью этого представления можно получить необходимые условия стационарности состояний сети. В частности, условие того, что сеть будет генерировать в качестве аттракторов векторы минимального базиса, легко формулируется в терминах матричных элементов $w_{km}(\alpha)$. Именно, $l-$му вектору базиса $\mathbf{b}$ будет соответствовать стационарное состояние тогда и только тогда, когда все недиагональные элементы $l-$й строки матрицы $w_{km}(\alpha)$ будут строго отрицательными.

Для рассмотренного нами выше примера эта матрица имеет вид

$$\mathbf{w}(\alpha) = \begin{Vmatrix} 0 & -1 & -1 & -1 \\ -1 & 0 & -1 & -1 \\ -1 & -1 & 0 & -1 \\ -1 & -1 & -1 & 0 \end{Vmatrix},$$

из которого с очевидностью следует стационарность *всех* векторов минимального базиса.

Метод Кинцеля. Уничтожение фрустрированных связей.

“Ложная память” имеет интересный нетривиальный смысл и в случае использования других правил обучения, минимизирующих энергию нейронных сетей.

Одно из них было предложено в 1985 году Кинцелем, который основывал свои рассуждения на реальном наблюдении, согласно которому у ребенка в первые несколько лет жизни отмирает большое число синапсов, хотя именно в это время он учится и усваивает огромное количество информации (Kinzel, 1985). Подобное явление подсказало Кинцелю следующий метод обучения. Возьмем полностью неорганизованную сеть нейронов $\sigma_i \in \{\pm 1\}, i = 1,\ldots,N$ с нулевыми порогами и связями, величины которых имеют Гауссово распределение с нулевым средним, и ликвидируем в ней все фрустрированные в векторах памяти соединения. То есть для всех запоминаемых векторов $\sigma^n, n = 1,\ldots,P$ обнуляются все связи, для которых $w_{ij}\sigma_i^n\sigma_j^n < 0$. В результате получается сеть, в которой все состояния кодируемые векторами σ^n, очевидно, будут стационарными.

Требование нефрустрированности каждой связи для всех запоминаемых векторов, конечно, очень сильное. Для слабо коррелированных образов приходится уничтожать так много межнейронных соединений, что в полученной слабосвязанной сети почти все состояния оказываются стабильными, т.е. появляется большое число “ложных” образов. (Если нейроны вообще не связаны - $\forall w_{ij} = 0$, то *все* возможные состояния сети стационарны). Положение

улучшается, если запоминаемые векторы коррелированы друг с другом. Количество стационарных состояний при этом уменьшается, что было продемонстрировано Кинцелем в ходе компьютерного моделирования. Тем не менее, полное число стационарных состояний не может быть уменьшено до набора запоминаемых векторов. Минимальная память в этой сети представляет собой все возможные комбинации векторов минимального базиса, за исключением тех из них, в которых коррелируют состояния нейронов, антикоррелирующие в запоминаемых векторах. Сеть с такой минимальной памятью может быть получена с помощью простой модификации метода уничтожения фрустрированных связей, который стартует с сети, у которой величины всех синаптических связей положительны и равны между собой, и не уничтожает, а инвертирует знак связи, фрустрированной во всех запоминаемых состояний. В примере, иллюстрируемом приводимым ниже рисунком,

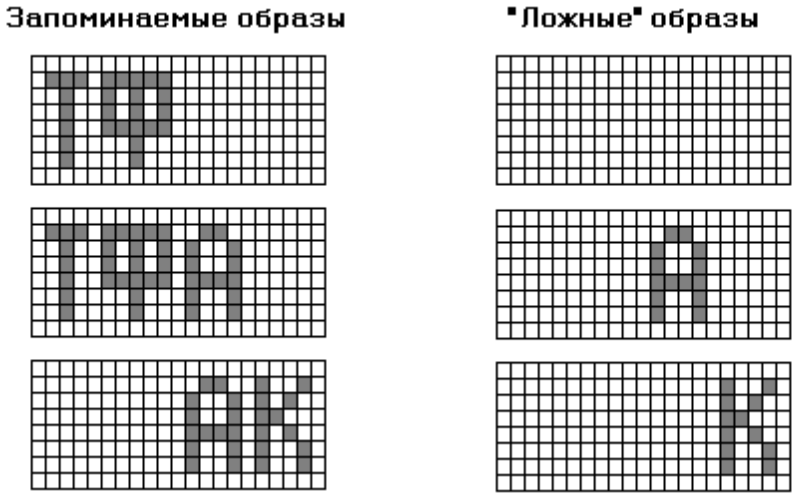

Рисунок 5. Слева - состояния, запоминаемые в сети Кинцеля . Справа - "ложные" образы.

в сети из 168 нейронов, организованных в двумерную структуру, запоминаются три образа: (**ТФ**__) (**ТФА**_) и (__**АК**). "Ложными" образами для сети с минимальной памятью будут при этом: пустое поле (____); (__**А**_); (___**К**) и их негативы. Невозможно раздельное появление в образе памяти (**Т**___) и (_**Ф**__), так как им соответствует один вектор минимального базиса. Невозможно также появление стационарного состояния (**ТФ_К**), так как в заучиваемых образах присутствие (**ТФ**__) исключает присутствие (___**К**) и наоборот.

Неустранимость ложной памяти. Запрещенные наборы,

Мы рассмотрели Хеббовское и Кинцелевское правила построения синаптических связей и убедились, что соответствующие сети демонстрируют нетривиальное отображение множества заучиваемых образов на множество аттракторов сети. В частности, ряд аттракторов далеки от заучиваемых образов и квалифицируются как ложная память. Возникает естественный вопрос о существовании такого метода обучения, который вообще бы устранял дополнительную память.

Оказывается, что ответ на него в общем случае отрицательный. Имеются такие наборы образов, что какую бы матрицу синаптических связей и пороги нейронов, гарантирующие их стационарность, мы не выбрали, в сети с неизбежностью возникнут иные аттракторы.

В частности, уже в сети из трех нейронов невозможно обеспечить стационарность только следующих четырех состояний: (0,0,0), (1,1,0), (1,0,1) и (0,1,1) или симметричного набора состояний. Такие наборы векторов, которые не могут составлять и исчерпывать память сети,

называют *запрещенными*. Можно показать, что для сети из трех нейронов два приведенных выше множества векторов исчерпывают все запрещенные наборы образов.

В сети из четырех нейронов не реализуемы уже 40 наборов векторов, но все они могут быть получены всего из двух независимых наборов *преобразованием однотипности* - перестановками переменных и инверсией.[4] Такая тенденция является обнадеживающей с точки зрения возможностей сетей к запоминанию образов, поскольку доля не реализуемых функций падает. Однако сети, аттракторы которых сконструированы заранее, могут имитировать только ассоциативную память, не создающую новой информации. Нас же сейчас интересует как раз эффект *обобщения*, присущий рекуррентным сетям, так же как и обычным персептронам.

Версии протитипа

Итак, структура аттракторов в модели Хопфилда может допускать различные содержательные интерпретации. В том случае, когда она совпадает со структурой запоминаемых образов мы говорим об ассоциативной памяти (пассивной). Если, напротив, в сети формируется единственный аттрактор, в каком-то смысле являющийся прототипом этих образов, то проявляется способность сети к *обобщению* (generalization). В общем же случае структура аттракторов сети настолько сложна, что на первый взгляд не допускает какой-либо наглядной трактовки. Действительно, такая трактовка должна быть настолько универсальной, чтобы включать режимы запоминания и обобщения в качестве предельных случаев. Тем не менее она возможна и опирается на рассуждения, которые приводятся в данном разделе.

Начнем с рассмотрения сети Хопфилда, в память которой, согласно правилу Хебба, записан только один образ σ^1. В этом случае синаптические связи определяются выражением

$$w_{ij} = \sigma_i^1 \sigma_j^1, \; i \neq j; \quad w_{ii} = 0; \quad i, j = 1, \ldots, N$$

У такой сети есть только два зеркально симметричных стационарных состояния $\pm \sigma^1$. Если она перейдет в одно из них, то величина энергии в минимуме составит

$$E = -\frac{1}{2} \sum_{i \neq j} w_{ij} \sigma_i^1 \sigma_j^1 = -\frac{1}{2} \sum_{i \neq j} (\sigma_i^1 \sigma_j^1)(\sigma_i^1 \sigma_j^1) = -\frac{1}{2} N(N-1)$$

Заметим, что все связи в сети дают в энергию одинаковый отрицательный вклад и поэтому являются не фрустрированными. Напомним, что условием фрустрации связи в состоянии сети σ является неравенство $w_{ij} \sigma_i \sigma_j < 0$.

Именно это условие *не выполняется* ни для одной связи в сети с записанным единственным образом. Мы можем трактовать подобную ситуацию так, что сеть с одним записанным в нее образом точно воспроизводит его в виде своего аттрактора (с точностью до зеркального отражения), и *если мы выберем в этой сети случайную связь, то вероятность ее фрустрации будет равна нулю.*

Таким образом, сеть Хопфилда идеально приспособлена для хранения единственного образа.

[4] Всего существует 402 типа булевых функций четырех переменных, к которым сводится все множество из 65536 функций.

Рассмотрим теперь следующую систему (см. Рисунок 6). Пусть в Хопфилдовской сети-передатчике (слева) записан единственный образ σ^{pro}, который нам неизвестен. Этот образ многократно передается в Хопфилдовскую сеть-приемник (справа) в виде сообщения через канал с шумом. При его прохождении образ σ^{pro} искажается так, что некоторые компоненты кодирующего его вектора меняют свой знак на противоположный.

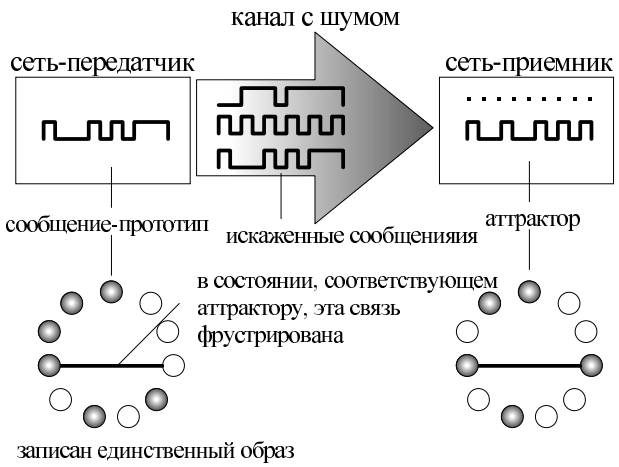

Рисунок 6. Вверху: интерпретация стационарных состояний в сети Хопфилда как локально наиболее правдоподобных версий сообщения, многократно переданного сетью-передатчиком в сеть-приемник через канал с шумом. Внизу: сети с записанным единственным сообщением прототипом (слева) и со всеми искаженными версиями этого сообщения (справа).

Задача сети-премника состоит в том, чтобы имея P полученных сообщений $\sigma^n,\ n=1,\ldots,P$ восстановить исходное сообщение σ^{pro}. Исходя из полученных сообщений, оценим вероятность того, что в исходном сообщении компоненты σ_i^{pro} и σ_j^{pro} имеют одинаковое или противоположные значения. Для этого нужно просто подсчитать, в скольких из P сообщений произведения $\sigma_i^n\sigma_j^n$ положительны или отрицательны и отнести это число к полному числу сообщений. Формально эти вероятности можно записать как

$$\Pr_{ij}^{\pm} = \frac{1}{2P}\sum_{n=1}^{P}(1 \pm \sigma_i^n\sigma_j^n)\,.$$

Вспоминая выражение для правила Хебба, убеждаемся что если сообщения $\sigma^n,\ n=1,\ldots,P$, полученные сетью-приемником, сформируют свои связи в соответствие с ним, то тогда

$$w_{ij} = \Pr_{ij}^{+} - \Pr_{ij}^{-},\quad i \neq j\,.$$

Используя последнее соотношение, преобразуем выражение для энергии состояния σ в сети-приемнике к виду

$$E(\sigma) = -\frac{N(N-1)}{2} + \frac{1}{2}\Big(\sum_{i=1}^{N}\sum_{j\neq i}^{N} \mathrm{Pr}_{ij}^{-}(1+\sigma_i\sigma_j) + \sum_{i=1}^{N}\sum_{j\neq i}^{N} \mathrm{Pr}_{ij}^{+}(1-\sigma_i\sigma_j) \Big).$$

Поскольку мы не знаем точного вида сообщения σ^{pro}, записанного в связях сети-передатчика, то мы не знаем и величин этих связей. Однако, мы можем задаться следующим вопросом: если состояние сети-передатчика совпадает с состоянием сети-приемника σ, то какова вероятность, что случайно выбранная связь в сети-передатчике окажется фрустрированной. Легко увидеть, что эта вероятность равна

$$\mathrm{Pr}_{random(i,j)}^{frust} = \frac{1}{2N(N-1)}\left(\sum_{i=1}^{N}\sum_{j\neq i}^{N} \mathrm{Pr}_{ij}^{-}(1+\sigma_i\sigma_j) + \sum_{i=1}^{N}\sum_{j\neq i}^{N} \mathrm{Pr}_{ij}^{+}(1-\sigma_i\sigma_j) \right).$$

Таким образом, энергия состояния сети-приемника с точностью до постоянных множителя и слагаемого совпадает с вероятностью фрустрации случайно выбранной связи в сети-передатчике, оцененной по полученным от нее сообщениям.

Однако в сети-передатчике записано лишь одно сообщение, и вероятность фрустрации связей в ней равна нулю. Но поскольку ни сообщение, ни соответствующие ему связи сети-передатчика нам не известны, мы можем лишь пытаться найти такое состояние сети-приемника, которое хотя бы локально минимизирует эту вероятность. Подобные состояния были бы локально наилучшими версиями сообщения, посылаемого сетью-передатчиком. А так как вероятность нахождения фрустрированной связи в передатчике связана с энергией состояния в приемнике, то такими наилучшими версиями как раз и окажутся состояния, соответствующие энергетическим минимумам сети-приемника. Таким образом все аттракторы сети Хопфилда, связи которой сформированы согласно правилу Хебба, исходя из набора обучающих векторов σ , $n = 1,\ldots,P$, могут трактоваться как наиболее вероятные версии некоторого сообщения, переданного P раз через канал с шумом и представленных заучиваемыми векторами.

Подобный подход устраняет деление состояний памяти на истинные и ложные, давая им единую интерпретацию. В такой трактовке функционирование сети Хопфилда в качестве пассивной памяти соответствует случаю, когда шум в канале очень велик, т.е. все принимаемые сетью сообщения некоррелированы. Это не дает ей возможности выделить из них сообщения и, рассматривая их как равноправные его версии, сеть генерирует аттрактор в каждой точке N-мерного пространства $\sigma^n, n = 1,\ldots,P$. Если же, напротив, шум в канале невелик, т.е. все запоминаемые векторы мало отличаются от передаваемого сообщения, в сети вырабатывается его единственная версия.

Хотя первоначально сеть Хопфилда привлекалась для объяснения свойств ассоциативной памяти, можно привести множество различных примеров ее применения и для выделения зашумленного сигнала-прототипа. В качестве одного из таких примеров мы рассмотрим один - поиск промоторов в ДНК

Пример: поиск промоторов в ДНК

Промоторами называются области четырехбуквенной последовательности ДНК (построенной из нуклеотидов A,T,G,C), которые предшествуют генам. Эти области состоят из 50-70 нуклеотидов и распознаются специальным белком РНК-полимеразой. Полимераза связывается с промотором и *транскрибирует* ее (расплетает на две нити). У кишечной палочки, например, обнаружено около трехсот различных промоторов. Несмотря на различие, эти области имеют

некоторые похожие участки, которые представляют собой как бы искажения некоторых коротких последовательностей нуклеотидов (например, *бокс Гильберта* - TTGACA и *бокс Прибноу* - TATAAT). Поэтому основные методы распознавания промоторов основываются на представлении о консенсус-последовательности: некотором *идеальном* промоторе, искажениями которого являются *реальные* промоторы. Близость некоторой последовательности к консенсус-последовательности оценивается по значению некоторого индекса гомологичности. Очевидно, что представление о версии-прототипа в теории минимизирующих энергию нейронных сетей прямо соответствует представлению о консенсус-последовательности.

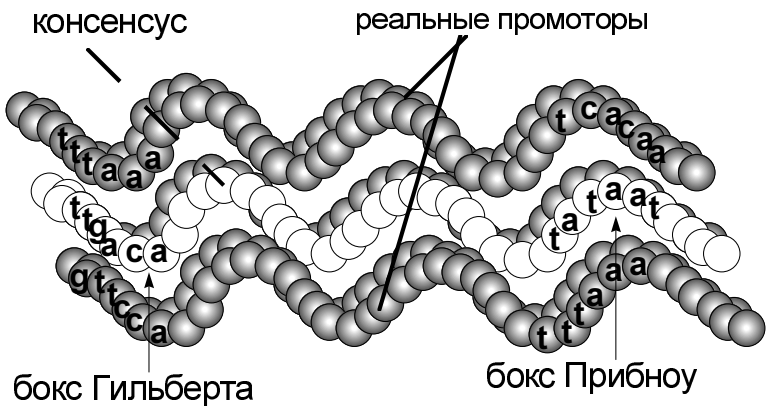

Рисунок 7. Идеальный промотор - консенсус-последовательность (в середине) является аналогом единственной версии прототипа - аттрактора в сети Хопфилда, выработанной в ней при записи зашумленных сообщений (аналогов реальных промоторов: сверху и снизу). Аналогом гомологического индекса, определяющего близость реальных промоторов к консенсус-последовательности, является энергия состояния сети.

Поэтому сеть Хопфилда, например, может непосредственно использоваться для ее поиска. Более того, оказывается, что энергия состояния сети может использоваться в качестве аналога гомологического индекса при оценке близости последовательности промотора к консенсус-последовательности. Такой подход позволил создать новый, весьма эффективный нейросетевой метод поиска промоторов. Аналогичный подход может использоваться для поиска скрытых повторов в ДНК и реконструкции эволюционных изменений в них.

Хотя молекулярная генетика представляет собой достаточно специфическую область применения методов обработки информации, она часто рассматривается как показательный пример приложений такой информационной технологии, как Извлечение Знаний из Данных (Data Mining). Применение для этих целей нейросетевых методов мы рассмотрим более подробно в отдельной главе.

Пустые классы

Активная кластеризация

Итак, мы установили, что преобразование информации *рекуррентными* нейронными сетями минимизирующими энергию может приводить к появлению в их пространстве состояний аттракторов, далеких по форме от образов внешнего для сети окружения. Таким образом, в отличие от рассмотренной в прошлой главе кластеризации, осуществляемой сетями без обратных связей, появляется возможность использовать рекуррентные сети для *активной кластеризации*, при которой сеть "творчески" относится к входным векторам, осуществляя нетривиальные обобщения поступающих на ее вход сигналов.

Теоретическим основанием такой активной кластеризации является отмеченное выше наблюдение, что *все* устойчивые состояния сети Хопфилда могут быть проинтерпретированы единым образом, как локально наилучшие версии одного сообщения, переданного через канал с шумом. Если предъявить сети эти сообщения, использованные для формирования ее связей, в качестве начальных состояний, она расклассифицирует их, отнеся к той или иной версии прототипа. (Такая классификация при асинхронной динамике будет в общем случае нечеткой - одно и то же начальное состояние в разных попытках может эволюционировать к разным аттракторам).

Однако, если исследовать *все* пространство состояний сети, предъявляя ей не ранее заучиваемые, а случайно сгенерированные векторы, то в нем могут обнаружиться такие аттракторы, которые не притягивают ни одного вектора из заучиваемого набора. Подобные аттракторы можно назвать *пустыми классами*, сформированными сетью. Понятие пустого класса не совпадает с понятием ложного состояний памяти. Последние не всегда описывает пустой класс. Например, когда в сети на основе слегка искаженных сообщений генерируется единственная версия сообщения, не совпадающая ни с одним из них (ложное состояние), но притягивающее все полученные сообщения (не пустой класс).

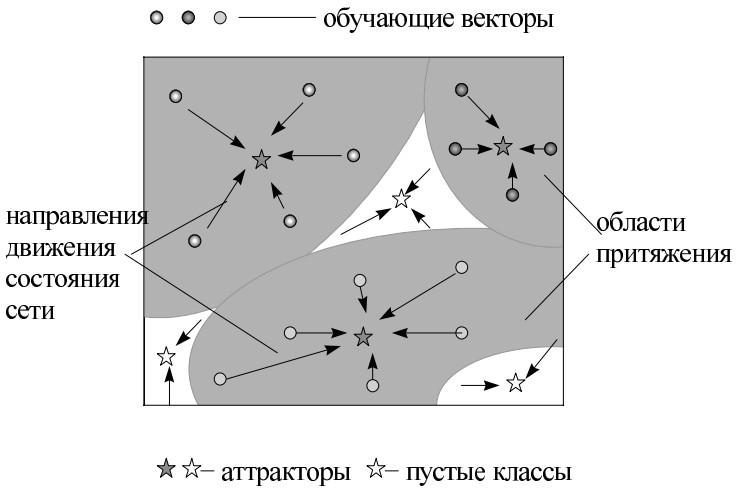

Рисунок 8. Пространство состояний сети с пустыми классами.

Аттракторы, являющиеся центрами притяжения состояний, относящихся к пустым классам, предоставляют нам совершенно новую информацию. Действительно, они *предсказывают* существование новых классов объектов, которые не имеют своих представителей в полученных сетью сообщениях.

Известным примером такой предсказательной категоризации является периодическая система элементов Менделеева, в которой изначально были определены три *пустые* клетки для впоследствии обнаруженных новых химических элементов. Итак, минимизирующие энергию нейронные сети типа сети Хопфилда могут использоваться для предсказания существования новых классов объектов.

Анализ голосований

В качестве иллюстрации приведем результаты кластеризации данных по голосованиям в ООН в 1969-1970гг. В данном примере анализировались голосования по 14 резолюциям для 19 стран. Сеть, производившая кластеризацию стран по степени схожести их голосований, состоит из $N = 14$ нейронов, состояния которых представляют картину голосования одного из участников по отобранным 14 резолюциям (*да* и *нет* соотносились с бинарными состояниями нейронов). Этой сети предъявлялись результаты голосований 19 стран - членов ООН, которые сформировали матрицу связей сети по правилу Хебба. Результаты категоризации входных векторов (а тем самым - и соответствующих стран), этой нейронной приведены в таблице:

Таблица 1. Кластеризация результатов голосований в ООН

Группа 1	**ранг**	**Группа 2**	**ранг**	**Группа 3**	**ранг**	**Группа 4**	**ранг**
США	0	Югославия	2	Болгария	4	???	0
Новая Зеландия	2	Кения	2	СССР	5		
Великобритания	3	ОАР	2	Сирия	6		
Албания	4	Дагомея	9	Танзания	7		
Бразилия	5	Сенегал	9				
Норвегия	5						
Мексика	6						
Швеция	7						
Венесуэла	8						
Франция	9						

Все используемые при обучении страны разделились на три легко интерпретируемых класса (условно: "*капиталистические*", "*неприсоединившиеся*" и "*социалистические*"), то есть кодирующие их голосования векторы-состояния эволюционируют к одному из трех стационарных состояний (локально наилучших версий прототипа "*страна - член ООН*"). Хэммингово расстояние от соответствующих состояний до притягивающих их аттракторов приведено в колонках "*ранг*".

У сети, однако, имеется и четвертое стационарное состояние, не притягивающее ни один из 19 образов, используемых при построении матрицы связей сети. Это состояние может рассматриваться как описывающее совершенно новую группу стран, в которую не входят ни одна из рассматриваемых. Мы можем описать эту группу, изучив вид соответствующего аттрактора - центра пустого четвертого класса. Действительно, такое изучение легко выявляет тот факт, что представители этого нового класса должны были бы иметь по сравнению с учтенными странами совершенно особое мнение при голосовании по корейскому вопросу. Учитывая то, что ни Южная, ни Северная Корея до сих пор не представлены в ООН, интерпретация этого класса является прозрачной. Очевидно, что подобный подход может использоваться при анализе экономических, социологических, демографических и других данных. В частности он может использоваться для поиска новых потенциальных и свободных мест на рынках, в политическом спектре и пр.

ЛИТЕРАТУРА

Crick, F. & Mitchison G. (1983). "The function of dream sleep". *Nature*, **304**, 111.

Diderich, S. & Opper, M. (1987) “Learning of Correlated Patterns in Spin-Glass Networks by Local Learning Rules”. *Phys.Rev.Lett.*, **58**, 949.

Ezhov, A.,A., Kalambet, Yu.,A. & Knizhnikova, L.A. (1990) “Neural networks: general properties and particular applications”. In: A.Holden & V.Kryukov (Eds.) *Neural Networks - Theory and Architecture*, Manchester, Manchester University Press, 39.

Ezhov, A.,A.. (1994) “Empty classes, predictive and clustering thinking networks”, *Neural Network World*, **4**, 671.

Ezhov,A.,A. & Vvedensky V.L. (1997) “Object generation with neural networks (when spurious memories are useful)”. *Neural Networks*, 9, 1491.

Hassoun M.H. ed; *Associative Neural Memories: Theory and Implementation.* Oxford, 1995.

Hopfield,J.,J. (1982a) “Neural Networks and Physical Systems with Emergent Collective Computational Abilities”, *Proc.Natl.Acad.Sci. USA*, **79**, 2554.

Hopfield, J.,J. (1982b) “Neurons with Graded Response Have Collective Computational Properties Like Those of Two-State Neurons”, *Proc.Natl.Acad.Sci. USA*, **81**, 3088.

Hopfield, J.,J., Feinstein, D.,I., & Palmer, R.G. (1983) “Unlearning has a stabilizing effect in collective memories”. *Nature*, **304**, 158.

Kinzel, W. (1985) “Learning and pattern recognition in spin glass models”. *Z. Phys. B. Condensed Matter*, **60**, 205.

Kohonen, T. *Self-organization and Associative Memory.* Springer-Verlag, 1989.

Muller, B., Reinhardt, J, & Strikland M.,T. (1995) *Neural Networks. An Introduction.* 2nd edition, Springer.

Vedenov, A.,A., Ezhov, A.,A., Kamchatnov, A.,M., Knizhnikova, L.,A., & Levchenko, E.,B. (1990) “Neural networks: general properties and particular applications”. In: A.Holden & V.Kryukov (Eds.) *Neural Networks - Theory and Architecture*, Manchester, Manchester University Press, 169.