Тема 9

Ассоциативные запоминающие нейронные сети

Ранее, был описан класс рекуррентных нейронных сетей. Это сети с наличием обратных связей между различными слоями нейронов. Наличие обратных связей добавляет в процесс функционирование динамические зависимости между слоями. При этом различные начальные наборы параметров сети могут приводить как к стабилизации нейросети на некотором этапе функционирования (то есть, начиная с определённой итерации, отсутствуют значительные изменения параметров сети), так и к дестабилизации и скачковым изменениям параметров.

Сети с обратными связями

В общем случае может быть рассмотрена нейронная сеть, содержащая произвольные обратные связи, по которым переданное возбуждение возвращается к данному нейрону, и он повторно выполняет свою функцию. Нейродинамика в таких системах становится итерационной, то есть критерий останова НС выглядит как логическое условие, выполнение которого означает конец работы сети. Это свойство существенно расширяет множество типов нейросетевых архитектур, но одновременно приводит к появлению новых проблем.

Обратные связи могут приводить к возникновению *неустойчивостей*, подобно тем, которые возникают в динамических системах при наличии положительной обратной связи. В нейронных сетях неустойчивость проявляется в циклической или блуждающей смене состояний и выходов нейронов. В общем случае ответ на вопрос об устойчивости динамики произвольной системы с обратными связями крайне сложен и до настоящего времени является открытым.

Далее мы рассмотрим некоторые классы сетей, функционирующих в качестве ассоциативных запоминающих устройств. Ассоциативная память играет роль системы, определяющей взаимную зависимость векторов. Главная задача ассоциативной памяти сводится к запоминанию обучающей выборки таким образом, чтобы при предъявлении входного вектора НС смогла сгенерировать ответ – какой из запомненных векторов выборки наиболее близок к поданному на вход.

Модель Хопфилда

Модель Хопфилда [2] занимает особое место в ряду нейросетевых моделей. В ней было предложено связать математический аппарат нелинейных динамических систем с нейронными сетями. Образы памяти сети соответствуют устойчивым предельным точкам (аттракторам) динамической системы. При этом появилась возможность теоретически оценить объём памяти сети Хопфилда, определить область параметров сети, в которой достигается наилучшее функционирование. Кроме того, была доказана теорема [1], описавшая подмножество сетей с обратными связями, выходы которых в конце концов достигают устойчивого состояния. Сеть Хопфилда является одной из первых и наиболее изученных архитектур рекуррентных нейронных сетей.

Определение. Динамическую систему будем называ

Функционирование сети Хопфилда

Рассмотрим сеть из *N* формальных нейронов, в которой выход каждого из нейронов y_i, *i=1..N*, может принимать только два значения {-1, +1}. Любой нейрон имеет связь со всеми остальными нейронами y_i, которые в свою очередь связаны с ним. Силу связи от *i*-го к *j*-му нейрону обозначим как w_{ij}.

В модели Хопфилда предполагается условие *симметричности* связей $w_{ij} = w_{ji}$, с нулевыми диагональными элементами $w_{ii} = 0$. Это условие имеет весьма отдаленное отношение к известным свойствам биологических сетей, в которых, как правило, не соблюдается симметрия связей между нейронами. Однако именно симметричность связей, как будет ясно из дальнейшего, существенно влияет на устойчивость динамики.

Выход каждого нейрона y_i в модели Хопфилда в режиме функционирования происходит по известному правилу для формальных нейронов МакКаллока и Питтса.

$$h_j(t) = \sum_{i \neq j} w_{ij} y_i(t-1), \qquad (9.1)$$

$$y_j(t) = \varphi(h_j(t)) := \begin{cases} -1, & \text{åñëè } h_j(t) < 0, \\ +1, & \text{åñëè } h_j(t) > 0_j, \\ y_j(t-1), & \text{åñëè } h_j(t) = 0. \end{cases} \qquad (9.2)$$

Таким образом, компактно формулы функционирования НС Хопфилда можно записать следующим образом:

$$\begin{aligned} \overline{h}(t) &= \overline{y}(t-1)\cdot W - \overline{T}, \\ \overline{y}(t) &= \varphi(\overline{h}(t)). \end{aligned} \qquad (9.3)$$

Изменение состояний возбуждения всех нейронов может происходить одновременно, в этом случае говорят о *параллельной* динамике. Рассматривается также и *последовательная* нейродинамика, при которой в данный момент времени происходит изменение состояния только одного нейрона. Многочисленные исследования показали, что свойства памяти нейронной сети практически не зависят от типа динамики. При моделировании нейросети на обычном компьютере удобнее последовательная смена состояний нейронов. В аппаратных реализациях нейросетей Хопфилда применятся параллельная динамика.

Алгоритм функционирования сети Хопфилда.

1. На вход сети подаётся вектор $\overline{y}(0)$.

2. Сеть функционирует в соответствии с формулами:

$$\begin{aligned} \overline{h}(t) &= \overline{y}(t-1)\cdot W - \overline{T}, \\ \overline{y}(t) &= F(\overline{h}(t)), \end{aligned}$$

где $F(\overline{x}) = (\varphi(x_1), \varphi(x_2), ..., \varphi(x_n))$.

3. Шаг 2 повторяется пока сеть не придёт (с достаточной точностью) в некоторой стандартный режим, т. е. выход сети в следующий момент времени не будет отличаться от выхода в предыдущий момент: $\overline{y}(t+1) = \overline{y}(t)$.

4. Этот стандартный режим соответствует тому образу, который вспоминает сеть.

Нейродинамика приводит к изменению вектора выхода $\overline{y}(t)$. Вектор выхода описывает траекторию на множестве выходов нейросети. Это множестве для сети с двумя уровнями возбуждения каждого нейрона,

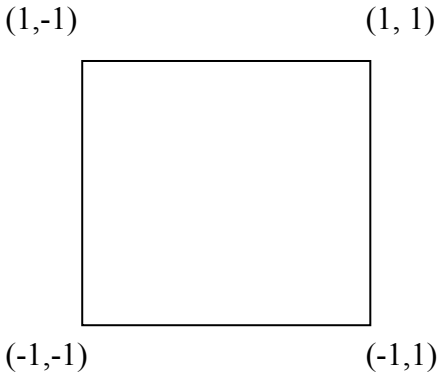

Рис. 9.1. 2-х мерный куб. Указанные на рисунке четыре точки являются возможными состояниями сети из 2-х нейронов.

очевидно, представляет собой множество вершин гиперкуба размерности, равной числу нейронов *N*. Возможные наборы значений координат вершин гиперкуба (см. рис.9.1) и определяют возможные значения вектора состояния.

Рассмотрим теперь проблему устойчивости динамики изменения выходов.

Определение. Назовем динамическую систему *диссипативной* [4], если производная энергии её по времени всегда отрицательна или равна нулю (в равновесном состоянии). В диссипативных системах на каждом шаге происходит необратимое уменьшение энергии. В установившемся режиме ($t=\infty$) все состояния таких систем сосредотачиваются на некотором подмножестве Y^* множества возможных состояний.

$$\lim_{t\to\infty}\bar{y}(t)=Y^*. \qquad (9.4)$$

Определение. Множество Y^*, к которому стремится состояние диссипативной динамической системы, называется *аттрактором*.

Определение. Аттрактор, для которого в установившемся режиме динамической системы $\bar{y}(t+1)=\bar{y}(t)$, для всех t, начиная с некоторого, называется *устойчивой предельной точкой*. Аттрактор, для которого $\bar{y}(t+k)=\bar{y}(t)$, для всех t, начиная с некоторого, называется *устойчивым предельным циклом*.

Из формул (9.3) следует, что на каждом временном шаге некоторый нейрон *i* изменяет свой выход в соответствии со знаком величины $h_i(t)-T_i$, то приведенное ниже соотношение всегда неположительно:

$$\Delta E_i(t)=-(y_i(t+1)-y_i(t))\cdot(h_i(t)-T_i)\le 0, \qquad (9.5)$$

где $E_i(t)=-y_i(t)\cdot(h_i(t)-T_i)$.

Таким образом, соответствующая величина *E(t)*, являющаяся суммой отдельных значений $E_i(t)$, может только убывать, либо сохранять свое значение в процессе нейродинамики. Определим «энергию сети» *E(t)* следующим образом:

$$E(t)=-\frac{1}{2}\sum_i\sum_j w_{ij}y_i(t)y_j(t)+\sum_i y_i(t)T_i \qquad (9.6)$$

или, в матричной форме

$$E(t) = -\frac{1}{2}\bar{y}(t)\cdot W\cdot \bar{y}^T(t) + \bar{y}(t)\cdot \bar{T} \qquad (9.7)$$

Введенная таким образом величина *E(t)* является функцией состояния $E(t) = E(\bar{y}(t))$ и называется энергетической функцией (энергией) нейронной сети Хопфилда. Поскольку она обладает свойством невозрастания при динамике сети, то одновременно является для нее функцией Ляпунова [5].

Далее будем полагать, что в ИНС Хопфилда нулевой вектор порогов активации: $\bar{T} = 0$. Все доказанные ниже утверждения можно обобщить и на случай ненулевого вектора $\bar{T}$.

Следующие утверждения являются очевидными [4].

Утверждение 9.1. Нейронная сеть Хопфилда сходится к устойчивым стационарным точкам, если в установившемся режиме, когда $\bar{y}(t+1) = \bar{y}(t)$, изменение энергии $\Delta E(t)$ равняется нулю.

Утверждение 9.2. Сеть Хопфилда сходится к предельному циклу длины два, если в установившемся режиме, когда $\bar{y}(t+2) = \bar{y}(t)$, изменение энергии $\Delta E(t)$ равняется нулю.

Утверждение 9.3. В сети Хопфилда достигнувшей предельного цикла длины два, т. е. при $\bar{y}(t+2) = \bar{y}(t)$, выполняется равенство $\bar{y}(t+1) = -\bar{y}(t)$.

Теорема 9.1. Аттракторами нейронной сети Хопфилда являются устойчивые стационарные точки и предельные циклы длины два.

Доказательство. Энергия сети Хопфилда определяется следующей формулой:

$$E(t) = -\bar{h}(t)\cdot \bar{y}^T(t). \qquad (9.8)$$

Изменение энергии $\Delta E(t+1) = -\left(\bar{h}(t+1)\bar{y}^T(t+1) - \bar{h}(t)\bar{y}^T(t)\right)$.

Попадание сети в устойчивую предельную точку означает что $\bar{y}(t+1) = \bar{y}(t)$, $\bar{h}(t+1) = \bar{h}(t)$ и, следовательно $\Delta E(t+1) = 0$. Попадание сети в предельный цикл длины два означает что $\bar{y}(t+1) = -\bar{y}(t)$, $\bar{h}(t+1) = -\bar{h}(t)$ и, следовательно $\Delta E(t+1) = 0$.

Теорема доказана.

Таким образом, поведение такой динамической системы устойчиво при любом входном векторе $\bar{y}(0)$ и при любой симметричной матрице связей $W = (w_{ij})$ с нулевыми диагональными элементами. Динамика при этом заканчивается в одном из минимумов функции Ляпунова, причем активности всех нейронов будут совпадать по знаку с входными сигналами.

Поверхность энергии $E(\bar{y})$ в пространстве состояний имеет весьма сложную форму с большим количеством локальных минимумов. Стационарные состояния, отвечающие минимумам, могут интерпретироваться, как *образы* памяти нейронной сети. Эволюция к такому образу соответствует процессу извлечения из памяти. При произвольной матрице связей W образы также произвольны. Для записи в память сети какой-либо осмысленной информации требуется определенное значение весов *W*, которое может получаться в процессе обучения.

Обучение ИНС Хопфилда. Правило обучения Хебба

Правило обучения для сети Хопфилда опирается на исследования Дональда Хебба [3], который предположил, что синаптическая связь, соединяющая два нейрона, будет усиливаться, если в процессе обучения оба нейрона согласованно испытывают возбуждение либо торможение. Простой алгоритм, реализующий такой механизм обучения, получил название *правила Хебба*. Рассмотрим его подробно.

Пусть задана обучающая выборка образов $\xi^{(\alpha)}$, $\alpha = 1 \ldots p$. Требуется построить процесс получения матрицы связей W, такой, что соответствующая нейронная сеть будет иметь в качестве стационарных состояний образы обучающей выборки (значения порогов нейронов T обычно полагаются равными нулю): $y(t) = y(t-1) \cdot W$, $y(t) := \xi$.

В случае одного обучающего образа правило Хебба, замена $y(t)$ на ξ приводит к требуемой матрице:

$$w_{ij} = \xi_i \cdot \xi_j . \qquad (9.9)$$

Покажем, что состояние $\bar{y} = \bar{\xi}$ является стационарным для сети Хопфилда с указанной матрицей. Действительно, для любой пары нейронов i и j энергия их взаимодействия в состоянии $\bar{\xi}$ достигает своего минимально возможного значения

$$E_{ij} = -\frac{1}{2}\xi_i\xi_j\xi_i\xi_j = -\frac{1}{2}. \qquad (9.10)$$

При этом E - полная энергия равна $E = -\frac{1}{2}N^2$, что отвечает глобальному минимуму.

Для запоминания других образов применяется итерационный процесс:

$$w_{ij}^{(\alpha)} = w_{ij}^{(\alpha-1)} + \xi_i^{(\alpha)} \cdot \xi_j^{(\alpha)}, \quad w_{ij}^{(0)} = 0, \quad \alpha = 1 \ldots p, \qquad (9.11)$$

который приводит к полной матрице связей в форме Хебба:

$$w_{ij} = \sum_{\alpha=1}^{p} \xi_i^{(\alpha)} \cdot \xi_j^{(\alpha)}. \qquad (9.12)$$

Устойчивость совокупности образов не столь очевидна, как в случае одного образа. Ряд исследований показывает, что нейронная сеть, обученная по правилу Хебба, может в среднем, при больших размерах сети N, хранить не более чем $p = \frac{N}{4\ln N}$ различных образов. Устойчивость может быть показана для совокупности ортогональных образов, когда

$$\frac{1}{N}\sum_{k=1}^{N} \xi_j^{(\alpha)} \cdot \xi_j^{(\beta)} = \delta_{\alpha\beta} := \begin{cases} 1, \alpha = \beta, \\ 0, \alpha \neq \beta. \end{cases} \qquad (9.13)$$

В этом случае для каждого состояния $\xi^{(\alpha)}$ произведение суммарного входа i-го нейрона h_i на величину его активности $s_i = \xi_i^{(\alpha)}$ оказывается положительным, следовательно, само состояние $\xi^{(\alpha)}$ является состоянием притяжения (*устойчивым аттрактором*):

$$h_i \cdot \xi_i^{(\alpha)} = \sum_j \left(\left(\sum_\beta \xi_i^{(\beta)} \xi_j^{(\beta)} \right) \xi_j^{(\alpha)} \right) \cdot \xi_i^{(\alpha)} = N > 0. \qquad (9.14)$$

Таким образом, правило Хебба обеспечивает устойчивость сети Хопфилда на заданном наборе относительно небольшого числа ортогональных образов. В следующем пункте мы остановимся на особенностях памяти полученной нейронной сети.

Ассоциативность памяти и задача распознавания образов

Динамический процесс последовательной смены состояний нейронной сети Хопфилда завершается в некотором стационарном состоянии, являющемся локальным минимумом энергетической функции $E(\bar{y})$. Невозрастание энергии в процессе динамики приводит к выбору такого локального минимума $\bar{y}$, в область притяжения которого попадает начальное состояние (исходный, предъявляемый сети образ $\bar{y}_0$). В этом случае также говорят, что выход $\bar{y}_0$ находится в чаше минимума $\bar{y}$.

При последовательной динамике в качестве стационарного состояния будет выбран такой образ $\bar{y}$, который потребует минимального числа изменений состояний отдельных нейронов. Поскольку для двух двоичных векторов минимальное число изменений компонент, переводящее один вектор в другой, является расстоянием Хемминга $\rho_H(\bar{y},\bar{y}_0)$, то можно заключить, что динамика сети заканчивается в ближайшем по Хеммингу локальном минимуме энергии.

Пусть состояние $\bar{y}$ соответствует некоторому идеальному образу памяти. Тогда эволюцию от состояния $\bar{y}_0$ к состоянию $\bar{y}$ можно сравнить с процедурой постепенного восстановления идеального образа $\bar{y}$ по его искаженной (зашумленной или неполной) копии $\bar{y}_0$. Память с такими свойствами процесса считывания информации является *ассоциативной*. При поиске искаженные части целого восстанавливаются по имеющимся неискаженным частям на основе ассоциативных связей между ними.

Ассоциативный характер памяти сети Хопфилда качественно отличает ее от обычной, адресной, компьютерной памяти. В последней извлечение необходимой информации происходит по *адресу* ее начальной точки (ячейки памяти). Потеря адреса (или даже одного бита адреса) приводит к потере доступа ко всему информационному фрагменту. При использовании ассоциативной памяти доступ к информации производится непосредственно по ее *содержанию*, т.е. по частично известным искаженным фрагментам. Потеря части информации или ее зашумление не приводит к катастрофическому ограничению доступа, если оставшейся информации достаточно для извлечения идеального образа.

Поиск идеального образа по имеющейся неполной или зашумленной его версии называется задачей *распознавания образов*. В нашей лекции особенности решения этой

задачи нейронной сетью Хопфилда будут продемонстрированы на примерах, которые получены с использованием модели сети на персональной ЭВМ.

Несмотря на интересные качества, нейронная сеть в классической модели Хопфилда далека от совершенства. Она обладает относительно скромным объемом памяти, пропорциональным числу нейронов сети N, в то время как системы адресной памяти могут хранить до $2N$ различных образов, используя N битов. Кроме того, нейронные сети Хопфилда не могут решить задачу распознавания, если изображение смещено или повернуто относительно его исходного запомненного состояния. Эти и другие недостатки сегодня определяют общее отношение к модели Хопфилда, скорее как к теоретическому построению, удобному для исследований, чем как повседневно используемому практическому средству.

Литература

1. Cohen M. A., Grossberg S. G. 1983. Absolute stability of global pattern formation and parallel memory storage by compatitive neural networks. IEEE Transactions on Systems, Man and Cybernetics 13:815-26.
2. Horfield J. J. 1982. Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Science 79:2554-58.
3. Hebb D. O. 1949. The organization of behavior. New lork: Wiley.
4. Головко В. А. Нейронные сети: обучение, организация и применение. – М.: ИПРЖР, 2001.
5. Ляпунов А. М. Общая задача об устойчивости движения. – М.: Гостехиздат, 1952.