ISSN 0868–5886 НАУЧНОЕ ПРИБОРОСТРОЕНИЕ, 2005, том 15, № 1, с. 29–45

ОБЗОРЫ

УДК 621.391+519.21+519.245

© Г. Ф. Малыхина, А. В. Меркушева

ЭЛЕМЕНТЫ СТАТИСТИЧЕСКОЙ КОНЦЕПЦИИ ОБУЧЕНИЯ НЕЙРОННОЙ СЕТИ И ПРОГНОЗИРОВАНИЕ ТОЧНОСТИ ЕЕ ФУНКЦИОНИРОВАНИЯ

Обучение нейронной сети (НС) для ряда задач (распознавание образов, нелинейная регрессия, идентификация распределения вероятности) анализируется в обобщенной форме на основе концепции, включающей вероятностную трактовку передаточной функции НС вход—выход, и базовых понятий элементов статистической теории обучения. Это — понятия, имеющие математически формализованную основу: мера многообразия (множества) отображений НС и изоморфного ему множества функций потерь; характеристика этого многообразия на основе энтропии и размерности Вапника—Червоненкиса; функционал риска (ФР) и условие, допускающее его оценку функционалом эмпирического риска (ФЭР); граница отличия величины фактического ФР от ФЭР. Описанные элементы статистической теории обучения обепечивают возможность прогноза и корректирования ("управления") показателя функционирования НС после обучения, т. е. при тестировании сети на данных, не участвовавших в обучении.

ВВЕДЕНИЕ

Расширению области приложений методов обработки информации на основе нейронных сетей (НС) как наиболее мощного средства аппроксимации многопараметрических зависимостей (многомерных функций) может способствовать более полное представление о статистических концепциях и принципах обучения НС. Корректно формализованная и логически адекватная основа процесса обучения строится на элементах статистической теории обучения [1–5] и позволяет учитывать вероятностный тип зависимости вход—выход НС, т. е. вероятностный тип зависимости передаточной функции сети, которая связана с ее структурой и величиной синаптических весов [6].

В практической реализации алгоритмов обучения НС особенно трудной остается задача оценки соотношения между доступным размером обучающей выборки, достигнутом при обучении показателем точности работы НС (выполнения желаемого вида многопараметрического отображения) и ожидаемым показателем точности преобразования на данных, не использовавшихся при обучении. Такую проверку нейросети называют тестированием, и показатель качества выполнения требуемого преобразования сетью определяется достигнутой (за счет обучения) способностью НС к обобщению.

Основные понятия, концепции и некоторые аспекты статистической теории обучения рассматриваются применительно к НС с прямым распространением сигнала [6–7], супервизорной ("с учителем") формой обучения и вероятностным представлением как входных данных (векторов $\mathbf{x}$ с распределением $P(\mathbf{x})$), так и выхода НС — вектора $\mathbf{y}$ с условным распределением $P(\mathbf{y}|\mathbf{x})$.

Супервизором (учителем) выдается сети набор одинаково и независимо распределенных векторов $\mathbf{x}$ из распределения $P(\mathbf{x})$ с соответствующими значениями выхода $\mathbf{y}$ из распределения $P(\mathbf{y}|\mathbf{x})$. Этим создается обучающая выборка образцов — примеров:

$$\{\mathbf{x}_1,\mathbf{y}_1;\ \mathbf{x}_2,\mathbf{y}_2;\ \ldots;\ \mathbf{x}_n,\mathbf{y}_n\}. \qquad (1)$$

Считается, что распределения $P(\mathbf{x})$ и $P(\mathbf{y}|\mathbf{x})$ вполне определенные, но неизвестные, а доступной информацией служит только обучающая выборка $\{\mathbf{x}_1,\mathbf{y}_1;\ \mathbf{x}_2,\mathbf{y}_2;\ \ldots;\ \mathbf{x}_n,\mathbf{y}_n\}$. Обучаемая НС за счет выбора значений ее параметров (совокупности α синаптических весов из некоторой допустимой области определения Λ) способна выполнять набор функций отображения $\{f(\mathbf{x},\alpha),\ \alpha\in\Lambda\}$. Задача обучения состоит в выборе некоторой функции, которая принадлежит множеству $\{f(\mathbf{x},\alpha),\ \alpha\in\Lambda\}$ и которая предсказывает (наилучшим образом) ответы супервизора. Отбор такой функции основывается на обучающем множестве (1), состоящем из n случайных и независимых, одинаково распределенных (НОР) наблюдений, извлекаемых в соответствии с вероятностью $P(\mathbf{x}, \mathbf{y}) = P(\mathbf{x})P(\mathbf{y}|\mathbf{x})$. Выбор лучшего из доступных приближений к желаемому отображению (т. е. к откликам супервизора) осуществляется минимизацией риска. Это значит, что нужно выполнить три этапа.

1. Найти подходящую меру расхождения (так называемую функцию потерь) $L(\mathbf{y}, f(\mathbf{x},\alpha))$ между откликом супервизора $\mathbf{y}$ и откликом, который обеспечивается обучаемой НС.

2. После этого на основе вероятности $P(\mathbf{x},\mathbf{y})$ следует получить функционал риска (ФР) в виде ожидаемой функции потерь $R(\boldsymbol{\alpha})$[1)]:

$$R(\boldsymbol{\alpha}) = \int [L(\mathbf{y}, f(\mathbf{x},\boldsymbol{\alpha}))] \mathrm{d}P(\mathbf{x},\mathbf{y}). \qquad (2)$$

3. Найти функцию $f(\mathbf{x},\boldsymbol{\alpha}_0)$, которая минимизирует функционал риска $R(\boldsymbol{\alpha})$ по классу функций $\{f(\mathbf{x},\boldsymbol{\alpha}),\ \boldsymbol{\alpha}\in\Lambda\}$ в условиях, где распределение совместной вероятности $P(\mathbf{x},\mathbf{y})$ неизвестно и доступна только информация, содержащаяся в обучающем наборе (множестве) (1).

Рассматриваемая модель обучения НС, принцип минимизации функционала риска, его компоненты и этапы реализации для получения лучшего отображения, аппроксимирующего желаемое отображение (задаваемое супервизором на обучающей выборке), охватывает все основные задачи, которые решаются средствами НС. Это — задачи распознавания образов, оценки нелинейной регрессии и выбора максимально правдоподобной плотности вероятности [8–10].

- При бинарном распознавании образов выход $\mathbf{y}$, определяемый супервизором, принимает два значения $\mathbf{y} = \{0, 1\}$, а $\{f(\mathbf{x},\boldsymbol{\alpha}),\ \boldsymbol{\alpha}\in\Lambda\}$ — это набор функций-индикаторов (т. е. функций, которые принимают только два значения — нуль или единицу). В качестве функции потерь принимается выражение

$$L(\mathbf{y}, f(\mathbf{x},\boldsymbol{\alpha})) = \begin{cases} 0, & \text{если } \mathbf{y} = f(\mathbf{x},\boldsymbol{\alpha}); \\ 1, & \text{если } \mathbf{y} \neq f(\mathbf{x},\boldsymbol{\alpha}). \end{cases} \qquad (3)$$

Для этой функции потерь функционал (2) обеспечивает вероятность ошибки классификации (т.е. когда ответы $\mathbf{y}$, даваемые супервизором, и ответы, даваемые функцией-индикатором $f(\mathbf{x},\boldsymbol{\alpha})$, отличаются). Поэтому задача состоит в том, чтобы найти функцию, которая минимизирует вероятность ошибки классификации. При этом мера вероятности $P(\mathbf{x},\mathbf{y})$ неизвестна, но имеются обучающие данные (1).

- В задаче оценки регрессии ответы супервизора $\mathbf{y}$ и набор $\{f(\mathbf{x},\boldsymbol{\alpha}),\ \boldsymbol{\alpha}\in\Lambda\}$, который содержит функцию регрессии $f(\mathbf{x},\boldsymbol{\alpha}_0)$, связаны соотношением $f(\mathbf{x},\boldsymbol{\alpha}_0) = \int \mathbf{y}\, \mathrm{d}_y P(\mathbf{x},\mathbf{y})$. Причем известно, что для $f(\mathbf{x},\boldsymbol{\alpha}) \in \mathbf{L}_2$ регрессией является функция, которая минимизирует функционал (2) с функцией потерь в форме (4):

$$L(\mathbf{y}, f(\mathbf{x},\boldsymbol{\alpha})) = (\mathbf{y} - f(\mathbf{x},\boldsymbol{\alpha}))^2. \qquad (4)$$

Так что задача оценки регрессии — это задача минимизации функционала риска (2) с функцией потерь (4) в ситуации, где распределение вероятности $P(\mathbf{x},\mathbf{y})$ неизвестно, но имеются обучающие данные (1).

- В задаче оценки плотности распределения вероятности из набора плотностей $\{p(\mathbf{x},\boldsymbol{\alpha}),\ \boldsymbol{\alpha}\in\Lambda\}$ в качестве функции потерь может использоваться выражение

$$L(p(\mathbf{x},\boldsymbol{\alpha})) = -\log(p(\mathbf{x},\boldsymbol{\alpha})). \qquad (5)$$

Желаемая плотность минимизирует функционал (2) с функцией потерь (5). Так что снова, чтобы оценить плотность, исходя из данных (1), нужно минимизировать функционал риска при условии, что распределение вероятности $P(\mathbf{x},\mathbf{y})$ неизвестно, а данные $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ независимы и одинаково распределены.

Развитие введенных выше исходных представлений о статистической основе обучения позволяет:

- ввести понятие эмпирического риска $R_{\text{эмпир.}}(\boldsymbol{\alpha})$ в виде среднего (по обучающей выборке) от функции потерь;
- ввести формализованное выражение для фактического риска (взвешенной по вероятности функции потерь), который характеризует уровень обобщения НС;
- установить правило индукции принципа минимизации эмпирического риска (принципа МЭР), согласно которому при увеличении размера обучающей выборки $R_{\text{эмпир.}}(\boldsymbol{\alpha}) \to R(\boldsymbol{\alpha})$ (эмпирический риск стремится к его фактическому значению [1, 3, 5, 6, 11].

Обоснование справедливости принципа МЭР использует понятие энтропии $H(n)$, характеризующей многообразие набора функций $\{f(\mathbf{x},\boldsymbol{\alpha}),\ \boldsymbol{\alpha}\in\Lambda\}$, и понятие размерности Вапника—Червоненкиса, определяющей (для того же набора функций) способность реализовать разделение набора обучающих точек (векторов) дихотомией различного вида [2)]. В упрощенной формулировке возможность применять правило принципа МЭР определяется условием выполнения соотношения $\frac{H(n)}{n} \underset{n\to\infty}{\to} 0$, которое верно при сильном возрастании размера обучающей выборки.

[1)] Функционалом принято называть скаляр, величина которого зависит от некоторых функций. Здесь это функционал риска $R(\boldsymbol{\alpha})$, однако в функционале (2) остается "сквозная" переменная $\boldsymbol{\alpha}$, которая делает этот функционал зависящим от $\boldsymbol{\alpha}\in\Lambda$. Параметр $\boldsymbol{\alpha}$ определяет разнообразие функций отображения $\{f(\mathbf{x},\boldsymbol{\alpha}),\boldsymbol{\alpha}\in\Lambda\}$, которые могут быть реализованы обучаемой нейронной сетью. Функционал риска $R(\boldsymbol{\alpha})$ определен как математическое ожидание функции потерь $L(\mathbf{y}, f(\mathbf{x},\boldsymbol{\alpha}))$ по вероятностной мере $\mathrm{d}P(\mathbf{x},\mathbf{y})$.

[2)] Дихотомия набора векторов (точек) $\mathbf{z}_1,\ldots,\mathbf{z}_n$ — это разделение их на две группы из несовпадающих точек.

Следующим этапом является оценка скорости сходимости $R_{\text{эмпир.}}(\boldsymbol{\alpha}) \to R(\boldsymbol{\alpha})$ эмпирического риска к фактической его величине (ожидаемой на фазе тестирования). Такая оценка получается в форме верхней границы возможного различия фактической функции риска от $R_{\text{эмпир.}}(\boldsymbol{\alpha})$. Эта граница $|R(\boldsymbol{\alpha}) - R_{\text{эмпир.}}(\boldsymbol{\alpha})|$ зависит от объема обучающей выборки и размерности Вапника—Червоненкиса и может быть определена как для конкретной задачи с фиксированной функцией распределения $P(\mathbf{x},\mathbf{y})$, так и в толерантной форме, т. е. в форме границы, которая справедлива при любой функции распределения.

Наличие таких границ позволяет (еще на стадии обучения сети) с помощью размера обучающей выборки и меры сложности набора отображений (характеризуемой РВЧ) влиять (и корректировать) на показатели обобщения НС на стадии ее тестирования, а затем и функционирования [11–14].

Особенно продуктивным оказывается так называемый метод структурной минимизации эмпирического риска в задаче бинарного распознавания образов, в которой используются разделяющие плоскости (или их нелинейные отображения в пространство размерности выше, чем размерность входных векторов). Эти плоскости выбираются по критерию наибольшей величины минимального отстояния от нее разделяемых точек (векторов) обучаемой выборки и называются оптимальными разделяющими плоскостями. Векторы, помещающиеся на границах плоского слоя, окружающего разделяющую плоскость и свободного от разделяемых точек, называют "векторами поддержки". Этот метод порождает новый класс алгоритмов, основанных на векторах поддержки, а НС, обучаемые с помощью этого метода, называют сетями с векторами поддержки [3)].

АНАЛИТИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ ОСНОВНЫХ КОНЦЕПЦИЙ СТАТИСТИЧЕСКОЙ ТЕОРИИ ОБУЧЕНИЯ НС

Анализ элементов статистической теории обучения [15] удобно проводить, используя более компактную (чем выше) форму обозначений. Пара векторов $(\mathbf{x},\mathbf{y})$ — вход и выход НС — обозначается одной буквой $\boldsymbol{z}$, тогда роль распределения $P(\mathbf{x},\mathbf{y})$ займет вероятностная мера $P(\boldsymbol{z})$. Таким образом, общая форма задачи обучения основана на понятии вероятностной меры $P(\boldsymbol{z})$, определенной на пространстве $\boldsymbol{Z}$, и наборе функций потерь $\{Q(\boldsymbol{z},\boldsymbol{\alpha}), \boldsymbol{\alpha}\in\Lambda\}$. Цель обучения достигается минимизацией функционала риска

$$R(\boldsymbol{\alpha}) = \int Q(\boldsymbol{z},\boldsymbol{\alpha})\mathrm{d}P(\boldsymbol{z}) \qquad (6)$$

при условии, что вероятностная мера $P(\boldsymbol{z})$ неизвестна, но имеется обучающая выборка в форме набора независимых одинаково распределенных (НОР) данных

$$\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_n. \qquad (7)$$

Функция потерь $Q(\boldsymbol{z},\boldsymbol{\alpha})$ строится на основе функции отображения, реализуемого нейронной сетью (при текущем наборе значений $\boldsymbol{\alpha}$ ее синаптических весов), поэтому два набора: набор функций потерь $\{Q(\boldsymbol{z},\boldsymbol{\alpha}), \boldsymbol{\alpha}\in\Lambda\}$ и набор функций отображения НС $\{f(\boldsymbol{z},\boldsymbol{\alpha}), \boldsymbol{\alpha}\in\Lambda\}$ — имеют взаимно однозначное соответствие (изоморфны), а численность их совпадает [4)]. В связи с этим описываемые ниже характеристики этих наборов (энтропия, функция роста, размерность Вапника—Червоненкиса) относятся в равной мере одновременно к обоим наборам (множествам).

Чтобы минимизировать функционал риска (6) при неизвестной вероятностной мере $P(\boldsymbol{z})$, используются возможности принципа МЭР. Ожидаемый функционал риска [5)] $R(\boldsymbol{\alpha})$ заменяется функционалом эмпирического риска (8), образованным на основе обучающего множества (7):

$$R_{\text{эмпир.}}(\boldsymbol{\alpha}) = \frac{1}{n}\sum_{i=1}^{n} Q(\boldsymbol{z}_i,\boldsymbol{\alpha}). \qquad (8)$$

Принцип МЭР имеет общий характер и связан с методами решения ряда задач обучения (оценка регрессии с помощью метода наименьших квадратов (МНК), метод максимального правдоподобия для оценки плотности вероятности). Так, в задаче регрессии вводится $(n+1)$-мерная переменная, используется функция потерь (4) и функционал эмпирического риска (8) в виде $R_{\text{эмпир.}}(\boldsymbol{\alpha}) = (1/n)\sum_{i=1}^{n}(\mathbf{y}_i - f(\mathbf{x}_i,\boldsymbol{\alpha}))^2$, который следует минимизировать. Эта процедура соответствует МНК. Для выбора функции плотности вероятности из данного набора $\{p(\mathbf{x},\boldsymbol{\alpha}), \boldsymbol{\alpha}\in\Lambda\}$ при подстановке

[3)] Такого типа алгоритмы и НС, допускающие обучение на их основе, в американской терминологии называют SVM-алгоритмами (от *support vector machine*) и SV-нейронными сетями.

[4)] Выражаясь вполне строго, следует говорить о мощности множества $\{Q(\boldsymbol{z},\boldsymbol{\alpha}), \boldsymbol{\alpha}\in\Lambda\}$, так же как и о мощности множества $\{f(\boldsymbol{z},\boldsymbol{\alpha}), \boldsymbol{\alpha}\in\Lambda\}$ функций отображения, которые могут быть реализованы НС (за счет выбора значений $\boldsymbol{\alpha}$ ее синаптических весов из допустимого множества этих значений Λ). Использование понятия *набор* вместо *множество* сделано для простоты и в связи с тем, что эти множества (наборы) можно считать дискретными и перечислимыми или конечными по численности значений.

[5)] Под ожидаемым имеется ввиду математическое ожидание — усреднение по вероятностной мере $P(\boldsymbol{z})$.

функции потерь (5) в (8) получается метод максимального правдоподобия, и, чтобы найти аппроксимацию плотности распределения, нужно минимизировать $R_{\text{эмпир.}}(\boldsymbol{\alpha}) = -(1/n)\sum_{i=1}^{n} \ln[P(\mathbf{x}_i, \boldsymbol{\alpha})]$.

К прикладным аспектам статистической теории обучения относится формализованная трактовка следующих этапов [1, 2, 6, 13–14].

- Обоснование использования принципа МЭР для оценки фактического функционала риска (ФР) $R(\alpha)$ и его минимального значения, которое может быть получено на НС с доступным ей набором отображений $\{f(z,\alpha),\ \alpha\in\Lambda\}$ (и соответственно с набором функций потерь $\{Q(z,\alpha),\ \alpha\in\Lambda\}$). Это значит, что требуется определить условия, при которых принцип МЭР может служить начальным звеном процедуры оценки фактического ФР [6)], т. е. показателя точности обобщения НС. Таким образом, рассмотренные ниже условия обеспечивают правомерность следующей цепочки соотношений [7)]:

$$\boldsymbol{\alpha}_n = \arg\left\{\min_{\boldsymbol{\alpha}\in\Lambda}\left[R_{\text{эмпир.}}(\boldsymbol{\alpha}) = \frac{1}{n}\sum_{i=1}^{n} Q(z_i, \boldsymbol{\alpha}\right]\right\};$$

$$R(\boldsymbol{\alpha}_n)_{n\to\infty} \xrightarrow{\text{Вер.}} R(\boldsymbol{\alpha}_0), \qquad (9)$$

где $\boldsymbol{\alpha}_0$ значение, которое дает $\min_{\boldsymbol{\alpha}\in\Lambda} R(\boldsymbol{\alpha})$:

$$\boldsymbol{\alpha}_0 = \arg\left\{\min_{\boldsymbol{\alpha}\in\Lambda} R(\boldsymbol{\alpha})\right\};$$

$$R_{\text{эмпир.}}(\boldsymbol{\alpha}_n)_{n\to\infty} \xrightarrow{\text{Вер.}} R(\boldsymbol{\alpha}_0). \qquad (10)$$

Здесь и далее "Вер." — вероятность. Выражение (9) показывает, что решение, найденное с использованием МЭР, сходится к лучшему решению, которое может реализовать НС, а (10) показывает, что величина эмпирического риска сходится к наименьшему риску.

- Установление, насколько быстро наименьшее значение $R_{\text{эмпир.}}(\boldsymbol{\alpha})$ сходится (при возрастании n) к наименьшей величине фактического риска R.
- Получение соотношений для границы различия ЭФР $R_{\text{эмпир.}}$ и ФР R, которые зависят от размера обучающей выборки и меры многообразия отображений нейронной сети. Эта граница различия позволяет прогнозировать достижимый показатель обобщения НС.

Обоснование использования принципа МЭР

Обоснование использования принципа МЭР состоит в получении условия равномерной сходимости (т. е. сразу для всего набора $\{Q(z,\alpha),\ \alpha\in\Lambda\}$) эмпирического риска к действительному риску $R(\alpha)$ [12]:

$$\lim_{n\to\infty} \text{Вер.}\left\{\max_{\boldsymbol{\alpha}\in\Lambda}\left[R(\boldsymbol{\alpha}) - R_{\text{эмпир.}}(\boldsymbol{\alpha})\right] > \varepsilon\right\} = 0 \qquad (11)$$

для любого малого ε,

где n относится к объему обучающей выборки, по которой формируется $R_{\text{эмпир.}}$ [8)].

Значимость этого условия связана с содержащимся в нем утверждением, что любой анализ адекватности использования принципа МЭР должен предусматривать наименее благоприятный ("наихудший" относительно $\alpha\in\Lambda$) случай соотношения $R(\alpha)$ и $R_{\text{эмпир.}}$.

Логическая схема получения условия равномерной сходимости основана на концепциях, играющих важную роль в статистической теории обучения нейронных сетей. Это прежде всего относится к понятию энтропии для набора функций $\{Q(z,\alpha),\ \alpha\in\Lambda\}$) и одновременно для набора $\{f(z,\alpha),\ \alpha\in\Lambda\}$ функций отображений, реализуемых нейронной сетью. Понятие энтропии вводится в два этапа: сначала для функций-индикаторов, а затем — для функций общего вида.

Энтропия набора (множества) функций-индикаторов. Энтропия набора функций-индикаторов $\{Q(z,\alpha),\ \alpha\in\Lambda\}$) (т. е. функций, принимающих только два значения 0 или 1) характеризует меру разнообразия этого набора (на выборке обучающих векторов $z_1, z_2, \ldots, z_n$) величиной $N^{\Lambda}(z_1, z_2, \ldots, z_n)$, представляющей число различных способов разделений (дихотомий) этой выборки, которые могут быть получены с использованием функций заданного набора. Величину $H^{\Lambda}(z_1, z_2, \ldots, z_n) = \ln[N^{\Lambda}(z_1, z_2, \ldots, z_n)]$, называют случайной энтропией, поскольку она образована с использованием (случайной) обучающей выборки, формируемой на основе распределения

---

6) Далее без дополнительной детализации будут использоваться сокращения ФР $R(\alpha)$ — для фактического риска (функции потерь, усредненной по вероятностной мере) и ЭФР $R_{\text{эмпир.}}(\boldsymbol{\alpha})$ — для эмпирического функционала риска (среднего по обучающей выборке от функции потерь). ФР $R(\alpha)$ соответствует величине риска для НС с параметрами α, т.е. при отображении $f(z,\alpha)$.

7) Сходимость по вероятности означает (например, в (9)), что Вер $[|R(\alpha_n) - R(\alpha_0)|_{n\to\infty} > \varepsilon\,] = 0$ при любом малом ε. Или более строго: для любых малых чисел $\eta > 0$ и $\varepsilon > 0$ существует такое n_0, что при $n > n_0$ с вероятностью не менее $1-\eta$ выполняется соотношение $\|R(\alpha_n) - R(\alpha_0)\| < \varepsilon$.

8) Этот тип сходимости называют равномерной *односторонней* сходимостью. Здесь и далее более строго следовало бы использовать символ sup (супремум) вместо max, однако, как правило, в практике Λ является дискретным и конечным перечнем параметра α.

$P(z_1, z_2, \ldots, z_n)$. Математическое ожидание — вероятностное среднее (обозначаемое символом E) называют просто энтропией $H^{\Lambda}(n)$ набора функций-индикаторов $\{Q(z,\alpha), \alpha\in\Lambda\}$ на обучающей выборке размера n:

$$H^{\Lambda}(n) = \mathrm{E}\{H^{\Lambda}(z_1, z_2, \ldots, z_n)\} = \\ = \mathrm{E}\{\ln N^{\Lambda}(z_1, z_2, \ldots, z_n). \quad (12)$$

Энтропия $H^{\Lambda}(n)$ описывает ожидаемое разнообразие данного набора функций-индикаторов на обучающей выборке размера n.

Энтропия набора функций общего вида. Совокупность $\{Q(z,\alpha), \alpha\in\Lambda\}$ функций общего вида, значения которых находятся в ограниченных пределах $A \le Q(z,\alpha) \le B$, ограничена n-мерным кубом со стороной $B-A$. Она может трактоваться как совокупность точек в этом кубе или как совокупность n-векторов $\mathbf{q}(\alpha)=[Q(z_1,\alpha), Q(z_2,\alpha),\ldots, Q(z_n,\alpha)]^{\mathrm{T}}$, каждый из которых определяется значением $\alpha\in\Lambda$. Известно, что из такой совокупности можно выделить (минимальную по численности) ε-сеть векторов [9)], число которых удобно обозначить $N^{\Lambda}(\varepsilon; z_1, z_2, \ldots, z_n)$, т. к. это число зависит от Λ, определяющего набор $\{Q(z,\alpha), \alpha\in\Lambda\}$, от величины ε и от самой обучающей выборки $z_1, z_2, \ldots, z_n$ (поскольку последняя определяет совокупность векторов $\mathbf{q}(\alpha)$: $\{\mathbf{q}(\alpha)=[Q(z_1,\alpha), Q(z_2,\alpha),\ldots, Q(z_n,\alpha)]^{\mathrm{T}}, \alpha\in\Lambda\}$.

Логарифм величины $N^{\Lambda}(\varepsilon; z_1, z_2, \ldots, z_n)$ (которая является случайной, как и обучающая выборка $z_1, z_2, \ldots, z_n$) называют *случайной* ε-энтропией Вапника—Червоненкиса:

$$H^{\Lambda}(\varepsilon; z_1, z_2, \ldots, z_n) = \ln(N^{\Lambda}(\varepsilon; z_1, z_2, \ldots, z_n)).$$

Ее математическое ожидание $H^{\Lambda}(\varepsilon, n) = \mathrm{E}\{H^{\Lambda}(\varepsilon; z_1, z_2, \ldots, z_n)$ чаще всего называется просто энтропией Вапника—Червоненкиса или VC-энтропией. Форма написания VC-энтропии $H^{\Lambda}(\varepsilon, n)$ соответствует тому, что она характеризует меру разнообразия набора $\{Q(z,\alpha), \alpha\in\Lambda\}$ функций общего вида (а более точно — конечную ε-сеть этого набора) с точки зрения ожидаемого количества дихотомий выборки размера n из совокупности данных с распределением $P(z_1, z_2, \ldots, z_n)$.

Понятия энтропии имеют ту же направленность, но более конструктивную форму, что и условия равномерной сходимости типа (11), обеспечивающие правомерность использования принципа МЭР. Так, в задачах распознавания образов средствами НС применяются индикаторные функции потерь. В этом случае условие равномерной сходимости даже в более сильной (двусторонней) форме (13) обеспечивается при выполнении соотношения (14) [16]:

$$\lim_{n\to\infty} \mathrm{Bep.}\left\{\max_{\alpha\in\Lambda}\left|R(\alpha) - R_{\text{эмпир.}}(\alpha)\right| > \varepsilon\right\} = 0 \quad (13)$$

для любого малого ε;

$$\lim_{n\to\infty} \frac{H^{\Lambda}(n)}{n} = 0. \quad (14)$$

Для задач более широкой постановки, в которых в качестве набора функций потерь требуется применение функций общего вида (не относящихся к функциям-индикаторам), условие двусторонней равномерной сходимости, обеспечивающее справедливость использования принципа МЭР для прогноза обобщения НС и оценки фактического риска, выполняется одновременно с соотношением (15) для энтропии Вапника—Червоненкиса:

$$\lim_{n\to\infty} \frac{H^{\Lambda}(\varepsilon, n)}{n} = 0 \quad (15)$$

для любого малого ε.

Энтропия Вапника—Червоненкиса для набора функций общего вида строится, как показано выше, с использованием ε-сети ограниченной совокупности n-векторов $\{\mathbf{q}(\alpha)=[Q(z_1,\alpha), Q(z_2,\alpha),\ldots, Q(z_n,\alpha)]^{\mathrm{T}}, \alpha\in\Lambda\}$.

Таким образом, обоснованность практического применения логической последовательности соотношений (9) и (10) определяется установленной в статистической теории обучения импликацией [11–13]:

$$\lim_{n\to\infty} \frac{H^{\Lambda}(\varepsilon, n)}{n} = 0, \ \forall\varepsilon \ \Rightarrow \\ \Rightarrow \lim_{n\to\infty} \mathrm{Bep.}\left\{\max_{\alpha\in\Lambda}\left|R(\alpha) - R_{\text{эмпир.}}(\alpha)\right| > \varepsilon\right\} = 0, \ \forall\varepsilon. \quad (16)$$

Границы различия рисков и прогноз обобщения нейронной сети

Условия для адекватности применения принципа МЭР, выраженные в форме предельных соотношений для энтропии, носят асимптотический характер и, строго говоря, могут использоваться

[9)] Набор векторов $\{\mathbf{q}(\alpha), \alpha\in\Lambda\}$ имеет минимальную ε-сеть $\mathbf{q}(\alpha_1), \mathbf{q}(\alpha_2),\ldots, \mathbf{q}(\alpha_n)$, если существует $N = N^{\Lambda}(\varepsilon; z_1, z_2, \ldots, z_n)$ векторов $\mathbf{q}(\alpha_1), \mathbf{q}(\alpha_2),\ldots, \mathbf{q}(\alpha_N)$, таких что для любого вектора $\mathbf{q}(\alpha^*)$, $\alpha^*\in\Lambda$ среди этих векторов может быть найден вектор $\mathbf{q}(\alpha_r)$, который ε-близок к $\mathbf{q}(\alpha^*)$. Это значит, что $\rho(\mathbf{q}(\alpha^*), \mathbf{q}(\alpha_r)) = \min_{1\le i\le n} |Q(z,\alpha^*) - Q_{\text{эмпир.}}(z,\alpha_i)| \le \varepsilon$, где ρ — евклидово расстояние между векторами $\mathbf{q}(\alpha^*)$ и $\mathbf{q}(\alpha_r)$.

только при очень больших размерах обучающих выборок. Поэтому представляет интерес оценка скорости сходимости минимума эмпирического риска к оптимально достижимому фактическому ФР, условие для получения этой оценки, установление такого условия в общей форме, которая пригодна для совокупности задач с различными видами вероятностной меры $P(z_1, z_2, \ldots, z_n)$, и получение границ различия эмпирического и фактического рисков, которые позволяют прогнозировать уровень обобщения нейронной сети после обучения.

Существующий подход к решению первой группы перечисленных вопросов (т. е. кроме границ различия рисков) удобно проследить на задаче распознавания образов, решаемой нейросетевыми средствами, где в качестве набора функций потерь $\{Q(z,\alpha),\ \alpha\in\Lambda\}$ используются индикаторные функции. Получение результатов здесь базируется на модификации и некотором развитии рассмотренной выше концепции энтропии, которая отражает меру многообразия набора функций потерь $\{Q(z,\alpha),\ \alpha\in\Lambda\}$ и ту же меру для набора $\{f(\mathbf{x},\alpha),\ \alpha\in\Lambda\}$ функций отображения нейронной сети.

Помимо рассмотренной ранее энтропии набора индикаторных функций

$$H^{\Lambda}(n) = \mathrm{E}\{H^{\Lambda}(z_1, z_2, \ldots, z_n)\} = \\ = \mathrm{E}\{\ln N^{\Lambda}(z_1, z_2, \ldots, z_n)$$

вводится модифицированная VC-энтропия (MVCE) $H^{\Lambda}_{\mathrm{MVCE}}(n)$ и так называемая функция роста $G^{\Lambda}(n)$:

$$H^{\Lambda}_{\mathrm{MVCE}}(n) = \ln\{\mathrm{E}[H^{\Lambda}(z_1, z_2, \ldots, z_n)]\},$$
$$G^{\Lambda}(n) = \ln\left\{\max_{z_1,\ldots,z_n} N^{\Lambda}(z_1,\ldots,z_n)\right\}. \qquad (17)$$

Модифицированная VC-энтропия $H^{\Lambda}_{\mathrm{MVCE}}(n)$ представляет логарифм ожидаемого (взвешенного по распределению вероятности) значения случайной энтропии $N^{\Lambda}(z_1, z_2, \ldots, z_n)$, а функция роста $G^{\Lambda}(n)$ является логарифмом случайной энтропии, максимизированной по возможным вариантам значений в обучающей выборке. Эти функции определены таким образом, что для любого значения n справедливо неравенство

$$H^{\Lambda}(n) \le H^{\Lambda}_{\mathrm{MVCE}}(n) \le G^{\Lambda}(n).$$

На основе определения функций модифицированной VC-энтропии и функции роста могут быть даны основные положения элементов статистической теории обучения, относящиеся к группе поставленных выше вопросов.

Скорость сходимости эмпирической оценки риска к его фактическому значению

В обязательном условии (необходимом и достаточном) для применимости принципа МЭР, которому должна удовлетворять любая НС, использующая этот принцип, нет информации о скорости сходимости минимального эмпирического риска к величине ФР НС при обобщении. Условием быстрой сходимости ФР при значении вектора α весов у НС, минимизирующего $R_{\text{эмпир.}}$, к оптимальному значению ФР в обобщении служит соотношение

$$\lim_{n\to\infty} H^{\Lambda}_{\mathrm{MVCE}}(n) = 0.$$

При этом быстрая [10] сходимость гарантирует экспоненциальное убывание вероятности превышения разностью рисков любого малого числа ε:

$$\text{Вер.}\{R(\alpha_n) - R(\alpha_0) > \varepsilon\} < \exp(-c\varepsilon^2 n),$$

где c — некоторая положительная постоянная.

Надо заметить, что как соотношение, описывающее условие применимости принципа МЭР, так и условие быстрой сходимости справедливы только для данной вероятностной меры, т.е. для того распределения $P(z_1, z_2, \ldots, z_n)$, которое входит в формирование энтропий $H^{\Lambda}(n)$ и $H^{\Lambda}_{\mathrm{MVCE}}(n)$. Однако наиболее важно построить НС для решения многих различных задач — для различных вероятностных мер. Другими словами, желательно установить, при каких условиях принцип МЭР является адекватным и обеспечивается быстрой сходимостью независимо от вероятностной меры $P(z)$, т.е. независимо от вида функции совместного распределения данных входа—выхода, используемых для обучения НС и для ее последующей работы на новых данных.

Таким условием для применимости принципа МЭР при любом виде распределения $P(z)$ служит выполнение соотношения для функции роста

$$\lim_{n\to\infty} \frac{G^{\Lambda}(n)}{n} = 0.$$

Условие в этой форме обеспечивает также быструю сходимость.

Описанные основы понятий и концепций прикладной теории обучения НС позволяют рассмотреть

[10] Принято считать, что сходимость происходит быстро, если для любого $n > n_0$ выполняется условие: Вер.$\{R(\alpha_n) - R(\alpha_0) > \varepsilon\} < \exp(-c\varepsilon^2 n)$, где $c > 0$ — положительная постоянная. Т.е. вероятность отличия $R(\alpha_n)$ от $R(\alpha_0)$ (при n больше некоторого значения n_0) убывает быстрее экспоненты $\exp(-c\varepsilon^2 n)$.

метод получения границ для разницы ФР (при значении вектора α, минимизирующего $R_{\text{эмпир.}}$) и оптимального значения ФР в обобщении. Эти границы более строго определяют скорость сходимости и устанавливаются сначала для вполне определенной функции распределения, а затем это ограничение снимается и определяются "глобальные" оценки границы, ориентированные на любой вид распределения. Глобальные оценки границы как более общие несколько шире.

Ряд преобразований (описанных ниже) позволяет получить неасимптотические оценки, которые ориентированы на объемы обучающих выборок, реально имеющихся в прикладных задачах при решении их средствами нейронных сетей. Таким образом, оценки для скорости обучения и показателей достижимых уровней обобщения НС будут основываться на различного типа границах, которые оценивают пределы этих показателей для фиксированного количества элементов обучающей выборки, позволяют их прогнозировать и в известной степени держать под контролем.

Оценка скорости обучения НС

Получение неасимптотической (т.е. для заданного размера обучающей выборки) границы на скорость равномерной сходимости связано с введением нового понятия — размерности Вапника—Червоненкиса (РВЧ) [11]. РВЧ служит для определения конструктивной границы на функцию роста $G^{\Lambda}(n)$. Показано [5], что функция роста может либо выражаться соотношением $G^{\Lambda}(n) = n\ln 2$, либо быть ограничена величиной

$$G^{\Lambda}(n) < h\left(\ln\frac{n}{h}+1\right),$$

где h — это такое целое число, для которого $G^{\Lambda}(n) = h\ln 2$ и *одновременно* $G^{\Lambda}(h+1) \neq (h+1)\ln 2$. Иначе говоря, функция роста может быть либо линейной функцией от n, либо быть ограниченной и иметь верхнюю границу в виде логарифмической функции.

Считается, что РВЧ набора функций-индикаторов $\{Q(z,\alpha),\ \alpha\in\Lambda,\ Q(z,\alpha)\in(0,1)\}$ будет конечной, если функция роста для этого набора является линейной. Кроме того, считается, что РВЧ набора функций-индикаторов является конечной и равной h, если функция роста ограничена логарифмической функцией с коэффициентом h.

Конечность РВЧ набора функций-индикаторов (которые в качестве отображения могут быть реализованы нейронной сетью) является необходимым и достаточным условием для адекватного использования принципа МЭР независимо от вероятностной меры. Конечность значения РВЧ обеспечивает также и быструю сходимость.

РВЧ имеет и несколько другую трактовку. РВЧ набора *функций-индикаторов* $\{Q(z,\alpha),\ \alpha\in\Lambda,\ Q(z,\alpha)\in(0,1)\}$ — это максимальное число h векторов $z_1,\ldots,z_h$, которые могут быть разделены на две части всеми 2^h возможными способами путем использования функций этого набора. Если такое разделение возможно для любого числа n векторов, то РВЧ равно бесконечности.

Для набора функций *общего* вида, имеющих границы a и A: $\{a\le Q(z,\alpha)\le A,\ \alpha\in\Lambda\}$, РВЧ определяется с помощью специальным способом образованного набора индикаторных функций. Вместо конечной функции общего вида создается функция-индикатор

$$I(z,\alpha,\beta)=\theta\{Q(z,\alpha)-\beta\},\quad \alpha\in\Lambda, \qquad (18)$$

где β — некоторая постоянная; θ — ступенчатая функция (функция Хэвисайда), принимающая значение 1, если ее аргумент (выражение в скобках) положителен, и принимающая значение 0, если аргумент менее нуля. Другими словами, функция θ определяется выражением

$$\theta(u)=\begin{cases}0, & \text{если } u<0;\\ 1, & \text{если } u\ge 0.\end{cases}$$

При этом в качестве РВЧ набора функций общего вида принимается РВЧ набора соответствующих функций-индикаторов (18).

Следствием этого правила определения РВЧ служат два положения, которые полезны в практическом приложении к нейронным сетям.

- Для набора линейных *индикаторных* функций (в n-мерном пространстве $z_1\ldots z_n$), которые имеют вид $Q(z,\alpha)=\theta\{\Sigma_{k=1}^{n}\alpha_k z_k+\alpha_0\}$, РВЧ равна $h=n+1$, т. к., используя функции этого набора, можно разделить по крайней мере $n+1$ векторов.
- Для набора линейных функций *общего вида* $Q(z,\alpha)=\Sigma_{k=1}^{n}\alpha_k z_k+\alpha_0$ (в n-мерном пространстве $z_1\ldots z_n$) РВЧ также равна $h=n+1$, поскольку этой величине $(n+1)$ равна РВЧ соответствующих индикаторных функций (если использовать $\alpha_0-\beta$ вместо α_0, что не изменит набора индикаторных функций).

Можно, например, рассмотреть плоскость $\mathbf{w}^{*\mathrm{T}}\mathbf{x}-b=0$, $\|\mathbf{w}^*\|=1$, называемую Δ-разделяющей при условии, если она классифицирует векторы $\mathbf{x}$ следующим образом:

$$y=\begin{cases}1, & \text{если } \mathbf{w}^{*\mathrm{T}}\mathbf{x}-b\ge\Delta;\\ -1, & \text{если } \mathbf{w}^{*\mathrm{T}}\mathbf{x}-b\le-\Delta.\end{cases}$$

[11] РВЧ не имеет ничего общего с обычным понятием размерности вектора, матрицы или пространства.

Тогда для некоторой совокупности векторов **x**, принадлежащих шару радиуса R, набор Δ-разделяющих плоскостей имеет РВЧ h, величина которой ограничена в соответствии с неравенством

$$h \le \min\left(\left[\frac{R^2}{\Delta^2} \right], n \right) + 1 .$$

Это показывает, что, хотя в общем случае РВЧ набора плоскостей равна n+1 (где n — размерность входного пространства), величина РВЧ набора Δ-разделяющих плоскостей при большой величине Δ может быть меньше, чем n+1.

Как отмечено выше, величина РВЧ ограничивает функцию роста $G^{\Lambda}(n)$ и, следовательно, дает форму условия адекватности использования принципа МЭР вне зависимости от распределения вероятностей. Тем не менее, справедливость этого условия пока гарантирована только для очень больших n, т. е. носит асимптотический характер. С точки зрения приложений НС желательно получение границ для различия минимального значения $R_{\text{эмпир.}}$ от функции риска при обобщении $R(\alpha)$ для фактически реализуемых размеров обучающей выборки. Такие границы установлены в двух модификациях: свободные от типа распределения (толерантные) границы и границы, соответствующие определенному распределению, связанному со спецификой решаемой задачи. Имея в виду зависимость границ от n, их называют также границами для скорости сходимости процесса обучения НС.

Свободные от типа распределения (толерантные) границы для скорости сходимости процесса обучения получены Вапником [12], [13]. Для набора функций $\{Q(z,\alpha),\ \alpha \in \Lambda\}$, имеющих конечное значение РВЧ и ограниченных как целое:

$$0 \le Q(z,\alpha) \le B, \quad \alpha \in \Lambda, \ B \text{ — константа}, \qquad (19)$$

выполняется условие в виде неравенства (20), которое связывает фактический риск $R(\alpha)$ и его эмпирическую оценку $R_{\text{эмпир.}}(\alpha)$. Неравенство (20) дает предел возможного превышения $R(\alpha)$ своей оценки $R_{\text{эмпир.}}(\alpha)$. С вероятностью не менее $1-\eta$ одновременно для всех функций (19) выполняется ограничение (20):

$$R(\alpha) \le R_{\text{эмпир.}}(\alpha) + \frac{B\varepsilon}{2}\sqrt{1 + \frac{4R_{\text{эмпир.}}(\alpha)}{B\varepsilon}}, \qquad (20)$$

где h — значение РВЧ; ε определяется выражением (21):

$$\varepsilon = 4\frac{h(\ln\frac{2n}{h} + 1) - \ln\eta}{n}. \qquad (21)$$

Для функций потерь НС в виде набора функций-индикаторов, используемых в задаче (бинарного) распознавания образов, постоянная B равна единице, так что в этом случае правая часть выражения (20) приобретает более простой вид.

Точные (зависящие от распределения) границы для сходимости процесса обучения определяют границы для степени отличия фактического риска от его эмпирической оценки $R_{\text{эмпир.}}(\alpha)$ и учитывают информацию о вероятностной мере. При анализе задачи получения таких границ используется метод, основанный на так называемом теоретико-множественном подходе [2, 10].

- Допускается (по априорной информации), что вероятность $P(z)$ относится к набору (множеству) $\boldsymbol{P}$ вероятностных мер, который является частью большего набора $\boldsymbol{P}_0$, т. е. $P(z) \in \boldsymbol{P} \subset \boldsymbol{P}_0$.
- Используется расширенное (обобщенное) определение функции роста

$$G^{\Lambda}_{\boldsymbol{P}}(\varepsilon, n) = \ln\left\{ \max_{P(z)\in\boldsymbol{P}} \mathrm{E}_{P(z)} N^{\Lambda}(\varepsilon; z_1, \ldots, z_n) \right\}. \qquad (22)$$

Для функций-индикаторов $\{Q(z,\alpha),\ \alpha \in \Lambda,\ Q(z,\alpha) \in (0, 1)\}$ и для экстремального случая, когда $\boldsymbol{P} = \boldsymbol{P}_0$, расширенное определение $G^{\Lambda}_{\mathbf{P}}(\varepsilon, n)$ совпадает с простой функцией роста $G^{\Lambda}(n)$. Для другого крайнего случая, когда $\boldsymbol{P}$ содержит только $P(z)$, обобщенная функция роста совпадает с модифицированной VC-энтропией $H^{\Lambda}_{\text{MVCE}}(n)$, выражение которой дано в (17).

В общем случае для ограниченного (константами A и B) набора функций потерь $\{Q(z,\alpha),\ A \le Q(z,\alpha) \le B,\ \alpha \in \Lambda\}$ при больших n выполняется соотношение [10, 16]:

$$\text{Вер.}\left\{ \max_{\alpha\in\Lambda} \left| \int Q(z,\alpha)\mathrm{d}P(z) - \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\alpha) \right| > \varepsilon \right\} \le$$

$$\le 12\exp\left\{ \left(\frac{G^{\Lambda}_{\boldsymbol{P}}(\varepsilon/6(B-A); 2n)}{n} - \frac{\varepsilon^2}{B-A} + \frac{\ln(n)}{n} \right) n \right\}.$$

Показано, что из этого соотношения может быть получена другая форма различия фактического риска и его эмпирической оценки. Для достаточно большого n с вероятностью не менее $1-\eta$ одновременно для всех $\alpha \in \Lambda$ (включая то α, которое минимизирует $R_{\text{эмпир.}}$) справедлива граница различия фактического риска и его эмпирической оценки, определяемая выражением:

$$\int Q(z,\boldsymbol{\alpha})\mathrm{d}P(z) \le$$

$$\le \frac{1}{n}\sum_{i=1}^{n} Q(z_i,\boldsymbol{\alpha}) + \sqrt{\frac{G_{\boldsymbol{P}}^{\Lambda}(\varepsilon/6(B-A);2n) - \ln\eta/12}{n}}. \quad (23)$$

К сожалению, эта граница не конструктивна, поскольку нет метода для оценки обобщенной функции роста. Чтобы эти границы стали практически полезными и точными, нужна оценка обобщенной функции роста для данных набора функций потерь и набора $\boldsymbol{P}$ вероятностных мер, но метод получения оценки обобщенной функции роста пока окончательно не разработан.

Прогноз и контроль показателя обобщения НС, реализуемого сетью после обучения, может основываться на рассмотренных выше границах. Так, при больших значениях размера обучающей выборки n второе слагаемое в правой части выражения (20) становится близким к нулю. Тогда функционал эмпирического риска становится хорошей оценкой ФР при обобщении, который отражает показатель обобщения НС и либо косвенно, либо непосредственно (как в случае задачи распознавания образов) характеризует процент количества ошибок, среднеквадратичную ошибку аппроксимации и другие показатели обобщения НС.

Принцип структурной минимизации риска

Элементы прикладной теории, связанные с прогнозированием, контролем и "управлением" показателем обобщения обученной НС, включают условие адекватности применения принципа минимизации функционала эмпирического риска [12)], которое учитывает размер обучающей выборки и соответствует такому ее объему, каким практически располагает исследователь. Формализованное обоснование принципа МЭР, использующее ряд модификаций концепции энтропии набора функций $\{Q(z,\boldsymbol{\alpha}),\ \boldsymbol{\alpha}\in\Lambda\}$, приводит к получению границ предельного различия ФЭР и ФР фазы обобщения нейронной сети с учетом размера обучающей выборки. Таким образом, эти результаты получены для малых объемов обучающих данных, обычно доступных при решении прикладных задач. Тем не менее, следует отметить некоторое несовершенство рассмотренного метода. Если при использовании соотношения (20) для границы скорости сходимости (предела различия минимума эмпирического риска и риска при обобщении) величина отношения n/h велика, то второе слагаемое в правой части (20) будет незначительно, и вследствие этого фактический риск $R(\boldsymbol{\alpha})$ очень близок к $R_{\text{эмпир.}}(\boldsymbol{\alpha})$, а малая величина $R_{\text{эмпир.}}(\boldsymbol{\alpha})$ обеспечивает малую величину фактического (ожидаемого) риска. Однако когда n/h мало́, то даже малое значение $R_{\text{эмпир.}}(\boldsymbol{\alpha})$ не гарантирует малости реального риска. В этом случае минимизация $R(\boldsymbol{\alpha})$ требует нового принципа, который может быть получен минимизацией одновременно обоих слагаемых в (20). Одно из них зависит от величины $R_{\text{эмпир.}}$, а второе зависит от РВЧ набора функций $\{Q(z,\boldsymbol{\alpha}),\ \boldsymbol{\alpha}\in\Lambda\}$. При этом необходимо найти метод, который наряду с минимизацией $R_{\text{эмпир.}}$ контролирует и "управляет" РВЧ обучаемой сети. Такой метод строится на основе принципа структурной минимизации риска (СМР) [5, 9].

Принцип СМР состоит в минимизации функционала риска относительно эмпирического риска и РВЧ набора фунций $\{Q(z,\boldsymbol{\alpha}),\ \boldsymbol{\alpha}\in\Lambda\}$ (являющегося отражением множества функций отображения, реализуемых НС). В наборе $\boldsymbol{S}$ функций $\{Q(z,\boldsymbol{\alpha}),\ \boldsymbol{\alpha}\in\Lambda\}$ вводится некоторая структура, состоящая из последовательности расширяющихся наборов (подмножеств) $\boldsymbol{S}_k$ функций $\{Q(z,\boldsymbol{\alpha}),\ \boldsymbol{\alpha}\in\Lambda_k\}$, таких что их объединение заполняет общий набор (множество) функций:

$$\boldsymbol{S}_1 \subseteq \boldsymbol{S}_2 \subseteq \ldots \subseteq \boldsymbol{S}_n \subseteq \ldots \subseteq \boldsymbol{S}^* = \bigcup_{(k)} \boldsymbol{S}_k\ ,$$
$$\boldsymbol{S}^* \cong \boldsymbol{S}, \quad (24)$$

где символ $\cong$ означает, что объединение $\boldsymbol{S}^*$ "плотно" в множестве $\boldsymbol{S}$.

При этом к допустимым относятся структуры, обладающие тремя свойствами:

- $\boldsymbol{S}^*$ везде плотно в $\boldsymbol{S}$, т.е. в $\boldsymbol{S}^*$ может быть найдена функция $Q(z,\boldsymbol{\alpha})$, достаточно близкая от функции, выбранной (любым образом) в $\boldsymbol{S}$.
- РВЧ h любого подмножества $\boldsymbol{S}_k$ — конечная величина.
- Каждый элемент $\boldsymbol{S}_k$ структуры ограничен в целом (некоторой константой B_k):

$$0 \le Q(z,\boldsymbol{\alpha}) \le B_k \quad \text{при} \quad \boldsymbol{\alpha}\in\Lambda_k .$$

Принцип СМР предполагает, что для данной обучающей выборки $\{z_1, z_2,\ldots, z_n\}$(численностью n) выбирается элемент структуры $\boldsymbol{S}_l$, $l = l(n)$ и выбирается такая функция из $\boldsymbol{S}_l$, для которой гарантированный риск (20) является минимальным. Принцип предполагает существующее противоречие между качеством аппроксимации и сложностью аппроксимирующей функции (фактически сложностью структуры НС). При возрастании n минимум $R_{\text{эмпир.}}$ снижается, однако слагаемое, ответственное за доверительный интервал (второе слагаемое в (20)), возрастает. Принцип СМР принимает во внимание оба фактора.

[12)] Иногда его называют принципом индукции минимизации эмпирического риска, имея в виду логический вывод правомерности его применения для прогнозирования функционала риска на фазе обобщения нейронной сети.

Метод СМР обеспечивает для любой функции распределения сходимость к лучшему решению с вероятностью единица [5, 9]. Этот метод является достаточно общим, независимым от распределения условием адекватности сходимости эмпирического риска к риску при обобщении. Функции $Q(z,\alpha_n^{l(n)})$, которые минимизируют риск $R(\alpha_n^{l(n)})$ на элементе $\boldsymbol{S}_l$ структуры, сходятся к функции, минимизирующей риск на всем множестве функций $\{Q(z,\alpha),\ \alpha\in\Lambda\}$. При достаточно большой обучающей выборке (при $n\to\infty$) асимптотическая скорость сходимости $R(\alpha_n^{l(n)})$ к общему минимуму $R(\alpha)$ на всем множестве $\boldsymbol{S}$ определяется выражением [13)]

$$V(n)=r_{l(n)}+B_{l(n)}\sqrt{\frac{h_{l(n)}\ln(n)}{n}} \qquad (25)$$

при условии, если изменение $l=l(n)$ таково, что

$$\lim_{n\to\infty}\frac{B_{l(n)}^2 h_{l(n)}\ln(n)}{n}\to 0. \qquad (26)$$

В (25) B_l — это граница для функций из $\boldsymbol{S}_l$, а $r_l(n)$ — скорость аппроксимации:

$$r_{l(n)}=\min_{\alpha\in\Lambda_l}\int Q(z,\alpha)\,\mathrm{d}P(z)-\min_{\alpha\in\Lambda}\int Q(z,\alpha)\,\mathrm{d}P(z).$$

Элементы теории построения алгоритмов обучения НС

Для выполнения процедур принципа СМР в обучающих алгоритмах нужно контролировать два фактора, присутствующие в соотношении (20) для границ:

- величину эмпирического риска;
- слагаемое, определяющее доверительный интервал в оценке (20), путем выбора из структуры элемента $\boldsymbol{S}_l$ с подходящей величиной РВЧ —

и стремиться их оба минимизировать.

Метод удобно проанализировать на задаче распознавания образов, рассматривая обучение нейронных сетей двух типов:

1) НС с прямым распространением сигнала (простого аналога взаимодействия нейронов) и

2) НС "с векторами поддержки", появление которых связывают с определенным этапом развития статистической теории обучения (СТО).

Чтобы следовать этой схеме анализа, требуется введение ряда соотношений, уточняющих описанные выше представления об элементах СТО для задачи распознавания образов.

Метод разделяющих (гипер-) плоскостей [14)] ***и его модификация*** [17–20]. Для минимизации эмпирического риска на наборе линейных индикаторных функций

$$f(\mathbf{x},\mathbf{w})=\theta\left\{\sum_{i=0}^{n}w_i x^i\right\},\quad \mathbf{w}\in W \qquad (27)$$

при обучающей выборке $(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_l,y_l)$, где n-вектор $\mathbf{x}_j=(x_j^1,\ldots,x_j^n)^{\mathrm{T}}$ и $y_j\in\{0,1\}$, $j=1,\ldots,l$, требуется найти вектор параметров НС $\mathbf{w}=(w_1, w_2, \ldots, w_n)^{\mathrm{T}}$, компоненты которого в качестве синаптических весов сети обеспечивают наименьшее значение для $R_{\text{эмпир.}}(\mathbf{w})$:

$$R_{\text{эмпир.}}(\mathbf{w})=\frac{1}{l}\sum_{j=1}^{l}\left[y_j-f(\mathbf{x}_j,\mathbf{w})\right]^2. \qquad (28)$$

К сожалению, в прикладных задачах набор линейных индикаторных функций часто оказывается неспособным обеспечить малое значение эмпирического риска [18]. В качестве возможности увеличения гибкости набора функций применяются два подхода.

- Использование более общего набора индикаторных функций, который является суперпозицией линейных индикаторных функций.
- Отображение входных векторов **x** в пространство более высокой размерности и создание в этом пространстве Δ–разделяющих плоскостей, у которых в слое толщиной Δ с каждой стороны плоскости не содержится разделяемых точек (векторов).

Первый подход связан с обучением НС обычной структуры, второй вариант связан с НС "с векторами поддержки". Как отмечено выше, сети такой структуры и алгоритмы для их обучения сформировались в одном из направлений СТО.

Сигмоидная аппроксимация индикаторных функций НС. Анализ требования минимизации функционала (20) в связи с обучением НС показывает, что непосредственное использование

[13)] Об асимптотической скорости сходимости $V(n)$ случайной величины ξ_n, n=1,2,... к ξ_0 говорят, когда $V(n)^{-1}\,|\,\xi_n-\xi_0\,|_{n\to\infty}\xrightarrow{\text{Вер.}}c$, где c — константа.

[14)] Поскольку рассмотрение проводится в векторном пространстве распознаваемых "точек" **x**, то разделение их осуществляется гиперплоскостями. Этот несколько перегруженный термин сначала указан в форме "гиперплоскость", а далее для простоты будет говориться о плоскости, подразумевается, конечно, везде ее многомерность, т. е. по сути дела —– гиперплоскость.

В формуле (27) и следующих слагаемое с индексом 0 соответствует смещению нейрона. При этом считается, что условная дополнительная компонента x_0=1, а w_0 представляет величину смещения [7].

градиентного метода для набора строго индикаторных функций невозможно, поскольку для них градиент равен или 0 или 1. Поэтому индикаторные функции аппроксимируются сигмоидными функциями[15)]

$$\hat{f}(\mathbf{x},\mathbf{w}) = S\{\sum_{i=0}^{n} w_i x^i\}, \qquad (29)$$

где S — гладкая монотонная функция, для которой $S(-\infty) = 0$ или -1, $S(+\infty) = +1$.

Это — сигмоидные функции типа $S_1(u) = \dfrac{1}{1+\exp(-u)}$ или $S_2(u) = \dfrac{2\,\mathrm{arctg}(u)+\tau}{2\pi}$.

При использовании одного из видов сигмоидной функции функционал

$$R_{\text{эмпир.}}(\mathbf{w}) = \frac{1}{l}\sum_{i=1}^{l}(y_i - \hat{f}(\mathbf{x}_i,\mathbf{w}))^2 \qquad (30)$$

становится гладким по $\mathbf{w}$ (непрерывно дифференцируемым), имеет градиент и поэтому может быть минимизирован с применением градиентных методов. Так, градиентный метод крутого спуска (по поверхности $R_{\text{эмпир.}}(\mathbf{w})$) использует правило обновления $\mathbf{w}$ в форме соотношения: $\mathbf{w}^{(n+1)} = \mathbf{w}^{(n)} - \gamma^{(n)}\ \mathrm{grad}\left[R_{\text{эмпир.}}(\mathbf{w}^{(n)})\right]$, где верхним индексом (n) указан номер итерации обновления; $\gamma^{(n)} \ge 0$ и обычно зависит от номера итерации. Для сходимости метода градиентного спуска достаточно, чтобы $\gamma^{(n)}$ удовлетворяло условию: $\sum_{n=1}^{\infty}\gamma^{(n)} = \infty$ и $\sum_{n=1}^{\infty}[\gamma^{(n)}]^2 < \infty$, т.е. ряд из $\gamma^{(n)}$ расходится, а ряд из $[\gamma^{(n)}]^2$ — сходится.

Таким образом, идея состоит в сигмоидальной аппроксимации индикаторных функций на стадии оценки коэффициентов $\mathbf{w}$ (синаптических весов НС) и использовании индикаторных функций с этой аппроксимацией на стадии распознавания.

Обобщение этой идеи ведет к более общим структурам НС с распространением сигналов вперед (без обратных связей [21, 22]). Так, чтобы увеличить гибкость набора решающих правил при обучении, рассматривается суперпозиция нескольких линейных функций-индикаторов. Такая суперпозиция соответствует сети нейронов, вместо отдельного нейрона, для которого достаточно набора простых индикаторных функций. При этом все функции-индикаторы в этой суперпозиции заменяются сигмоидными функциями.

Метод вычисления градиента эмпирического риска для сигмоидной аппроксимации функции активации нейронов, связанный с алгоритмом обратного распространения [7], введен в работах [4, 5]. Показано, что РВЧ нейронных сетей зависит от вида сигмоидной фнкции и количества синаптических весов в НС. При некоторых общих условиях РВЧ сети ограничена (хотя значение РВЧ обычно очень велико). Если РВЧ не меняется в процессе обучения, то способность НС к обобщению (т.е. показатели точности выполнения требуемого от нее отображения на новой информации с прежними статистическими характеристиками) зависит от того, насколько хорошо НС минимизирует эмпирический риск на достаточно большом обучающем материале.

При минимизации эмпирического риска с использованием метода обратного распространения возникают три проблемы.

1. Функционал эмпирического риска может иметь несколько локальных минимумов, и процедура минимизации гарантирует сходимость к некоторому из них. Поэтому в общем случае функция, найденная с использованием процедуры, основанной на градиенте, может быть далеко не лучшей. Качество полученной аппроксимации зависит от многих факторов и в особенности от начальной величины параметров алгоритма.

2. Сходимость к локальному минимуму может быть довольно медленной из-за высокой размерности пространства синаптических весов НС.

3. Сигмоидная функция имеет масштабирующий параметр, который влияет на качество. Чтобы выбрать этот параметр нужно сбалансировать противоречие между качеством аппроксимации и скоростью сходимости. Поэтому считается, что хорошая минимизация $R_{\text{эмпир.}}$ во многих отношениях зависит от искусства исследователя.

Оптимальные разделяющие плоскости. Для получения структуры НС, альтернативной к НС прямого распространения, следует сначала рассмотреть "оптимальные" разделяющие плоскости (фактически гиперплоскости с плоскопараллельной зоной, свободной от точек обучающей выборки) [23].

В задаче бинарного распознавания обучающие данные $\{(\mathbf{x}_1,y_1), (\mathbf{x}_2,y_2), \ldots, (\mathbf{x}_l,y_l);\ \mathbf{x}\in\mathbf{R}^n,\ y\in\{-1, +1\}\}$ могут быть разделены на два класса плоскостью

$$\mathbf{w}^{\mathrm{T}}\mathbf{x} - b = 0, \qquad (31)$$

причем считается, что выход НС $y = 1$ соответствует $\mathbf{x}\in$ класс1, а выход $y = -1$ соответствует $\mathbf{x}\in$ класс2.

Набор векторов разделяется оптимальной плоскостью (или Δ–разделяющей гиперплоскостью), если безошибочное разделение этого набора на два класса достигается с помощью плоскости при

[15)] Сигмоидной называют функцию активации нейрона в НС (имеющую смысл его передаточной функции) с монотонным ростом от нуля до единицы (как у функции распределения вероятности) или от –1 до +1.

пустом слое с максимальной толщиной Δ с каждой стороны этой плоскости [6].

Свойство разделяющей плоскости указывать, по какую сторону от нее лежит некоторый обучающий вектор $\mathbf{x}_i$, может быть представлено соотношением

$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - b \ge 1, \quad \text{если } y_i = 1;$$
$$\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - b \le -1, \quad \text{если } y_i = -1.$$

Более компактное описание этого свойства дает (эквивалентное по смыслу) выражение

$$y_i[\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - b] \ge 1, \quad i = 1,2,...,l\,. \qquad (32)$$

Показано [6, 24], что при условии (32) плоскость будет оптимальной, если норма вектора $\mathbf{w}$, определяющего нормаль (перпендикуляр) к этой плоскости будет минимальной. Поэтому для определения оптимальной плоскости требуется минимизация функционала (33) $\Phi(\mathbf{w})$ при дополнительном выполнении условия (32):

$$\Phi(\mathbf{w}) = \frac{1}{2}\| \mathbf{w} \|^2 = \frac{1}{2}(\mathbf{w}^{\mathrm{T}}\mathbf{w}). \qquad (33)$$

Решением этой задачи условной минимизации является "седловая" точка функционала Лагранжа (лагранжиана) $L(\mathbf{w}, b, \boldsymbol{\alpha})$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_l)$, который сводит задачу к безусловной минимизации за счет введения дополнительных параметров α_i, называемых множителями Лагранжа:

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}(\mathbf{w}^{\mathrm{T}}\mathbf{w}) - \sum_{i=1}^{l}\alpha_i\{y_i[\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - b] - 1\}, \qquad (34)$$

где α_i — множители Лагранжа.

Поскольку здесь (для удобства) условия с множителями Лагранжа введены в функционал со знаком минус, то этот функционал должен минимизироваться относительно $\mathbf{w}$, b и максимизируется относительно $\alpha_i \ge 0$. Решение $\mathbf{w}_0$, b_0 и $\alpha_i^{(0)}$ ($i = 1,\ldots, l$) удовлетворяет условиям

$$\frac{\partial L(\mathbf{w}_0, b_0, \alpha_i^{(0)})}{\partial b} = 0, \qquad \frac{\partial L(\mathbf{w}_0, b_0, \alpha_i^{(0)})}{\partial \mathbf{w}} = 0.$$

Явный вид этих условий (получаемый подстановкой развернутой формы лагранжиана (34)) позволяет выявить ряд свойств оптимальной гиперплоскости.

- Коэффициенты $\alpha_i^{(0)}$ для оптимальной гиперплоскости удовлетворяют ограничению

$$\sum_{i=1}^{l}\alpha_i^{(0)} y_i = 0, \quad \alpha_i^{(0)} \ge 0, \ i = 1,2,...,l. \qquad (35)$$

- Параметр оптимальной гиперплоскости $\mathbf{w}_0$ является линейной комбинацией векторов обучающего набора

$$\mathbf{w}_0 = \sum_{i=1}^{l}\alpha_i^{(0)} y_i \mathbf{x}_i, \quad \alpha_i^{(0)} \ge 0, \quad i = 1,2,...,l. \qquad (36)$$

- Решение удовлетворяет условию (называемому условием Куна—Таккера)

$$\alpha_i^{(0)}\{[(\mathbf{x}_i\,\mathbf{w}_0) - b_0]y_i - 1\} = 0. \qquad (37)$$

Из этих условий следует, что только некоторые обучающие векторы в выражении (36) — "векторы поддержки" — могут иметь в разложении $\mathbf{w}_0$ ненулевые коэффициенты $\alpha_i^{(0)}$. Векторы поддержки — это векторы $\mathbf{x}_i$, для которых в (32) достигается равенство, т. е. они поддерживают (принадлежат им) плоскости, лежащие с двух сторон от разделяющей плоскости и образующие слой (толщины 2Δ), свободный от обучающих точек. Поэтому получается соотношение:

$$\mathbf{w}_0 = \sum_{\Xi}\alpha_i^{(0)} y_i \mathbf{x}_i, \quad \alpha_i^{(0)} > 0, \qquad (38)$$

где Ξ — множество индексов совокупности векторов поддержки, определяющих $\mathbf{w}_0$.

Подстановка выражений для $\mathbf{w}_0$ обратно в лагранжиан (34) с учетом условия (37) дает функционал

$$W(\boldsymbol{\alpha}) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{ij}^{l}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j\,. \qquad (39)$$

Остается максимизировать этот функционал в квадранте неотрицательных α_i ($\alpha_i \ge 0$, $i = 1,2,\ldots,l$) при ограничении

$$\sum_{i}^{l}\alpha_i y_i = 0. \qquad (40)$$

Подстановка выражения (38) для $\mathbf{w}_0$ в (31) приводит к плоскости в виде выражения, связывающего векторы поддержки:

$$\sum_{i=1}^{l}\alpha_i^{(0)}\mathbf{x}^{\mathrm{T}}\mathbf{x}_i + b = 0\,. \qquad (41)$$

В случае, когда обучающие данные линейно неразделимы, может применяться метод получения квази-оптимальной разделяющей плоскости. Для этого используются новые переменные ξ_i (так называемые переменные бездействия, $\xi_i \ge 0$). Переменые ξ_i служат допустимой величиной погружения некоторой части из обучающих точек в "свободный" слой 2Δ, принадлежащий оптимальной разделяющей плоскости для остальной (большей) части обучающих точек. Чтобы количество и величина нарушений оптимальности были наименьшими, в минимизируемый функционал

вводится регуляризирующая компонента. Конструкция функционала имеет вид:

$$\Phi(\xi) = \mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{i=1}^{l}\xi_i ,$$

где постоянная C играет роль параметра регуляризации (в смысле А.Н. Тихонова). Функционал минимизируется при условии

$$y_i[\mathbf{w}^{\mathrm{T}}\mathbf{x}_i - b] \geq 1-\xi_i, \quad i = 1,2,\ldots, l.$$

Использование описанного выше метода условной оптимизации на основе введения множителей Лагранжа (с переходом к безусловной минимизации по расширенному перечню параметров) приводит к тому, что оптимальная плоскость снова выражается соотношением (41) на векторах поддержки. Коэффициенты α_i определяются путем максимизации того же квадратичного выражения (39), как и в случае линейной разделимости. Однако здесь требуется использовать несколько отличающиеся ограничения в виде условий (42):

$$0 \leq \alpha_i \leq C; \quad i = 1,2,...,l; \quad \sum_i^l \alpha_i y_i = 0. \qquad (42)$$

При отсутствии возможности разделимости анализируемого набора точек (векторов) плоскостью эта задача решается с помощью поверхности общего вида (в n-мерном пространстве). Для этого осуществляется преобразование в так называемое пространство признаков, которое имеет более высокую размерность (сравнительно с исходной размерностью векторов обучающей выборки) и в котором, как доказано Ковером (Cover T.) [27], может быть достигнута разделимость с помощью гиперплоскости. Такой метод при решении задачи распознавания образов используется нейронной сетью с векторами поддержки[16)]. Применяется концепция отображения входных обучающих векторов в пространство **Z** признаков, имеющее более высокую размерность, причем нелинейное преобразование выбирается априорно и "произвольно". В этом новом признаковом пространстве строится оптимальная разделяющая плоскость. Целью является создание ситуации (подобно рассмотренному ранее примеру), при которой для Δ-разделяющих гиперплоскостей РВЧ определяется отношением R^2/Δ^2. Для получения хорошего обобщения у НС следует контролировать РВЧ и уменьшать ее величину путем построения Δ-разделяющей гиперплоскости с максимальным значением Δ-слоя. Собственно, ради повышения Δ и используется пространство высокой размерности.

Бозером и Гуйоном (Boser B., Guyon I.) [19] было отмечено, что для описания оптимальных разделяющих плоскостей в признаковом пространстве и для оценки компонент вектора нормали (39) (представляющей эту разделяющую плоскость) требуется использовать произведение двух векторов $\boldsymbol{z}(\mathbf{x}_1)$ и $\boldsymbol{z}(\mathbf{x}_2)$, которые являются изображениями в признаковом пространстве входных векторов $\mathbf{x}_1$ и $\mathbf{x}_2$. Поэтому, если есть возможность оценить произведение двух векторов в признаковом пространстве $\boldsymbol{z}(\mathbf{x}_1)$ и $\boldsymbol{z}(\mathbf{x}_2)$ в виде функции двух переменных во входном пространстве $\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{z} = K(\mathbf{x},\mathbf{x}_i)$, тогда будет возможно и создать решения, которые эквивалентны оптимальной плоскости в признаковом пространстве. Чтобы получить это решение, следует заменить произведение $\mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j$ в (39) и (41) функцией $K(\mathbf{x}_i, \mathbf{x}_j)$. Другими словами, создаются нелинейные решающие функции, которые во входном пространстве имеют вид:

$$I(\mathbf{x}) = \mathrm{sign}\left(\sum_{\substack{(\text{векторы}\\ \text{поддержки}\ \mathbf{x}_i)}} \alpha_i K(\mathbf{x}_i, \mathbf{x}) + b\right), \qquad (43)$$

которые эквивалентны линейным решающим функциям (33) в признаковом пространстве. Коэффициенты α_i в (43) определяются путем решения уравнения (44) при ограничениях (42):

$$W(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{\substack{i,j=1\\ i\neq j}}^{l}\alpha_i\,\alpha_j y_i\, y_j K(\mathbf{x}_i, \mathbf{x}_j). \qquad (44)$$

В то же время, согласно функциональному анализу, общая форма произведения векторов определяется посредством симметричной, положительно определенной функции $K(x,y)$, удовлетворяющей условию Мерсера [28]: для любого сигнала с конечной энергией ($\int z(t)^2\,\mathrm{d}t \geq 0$) справедливо неравенство $\int K(x,y)z(x)z(y)\mathrm{d}x\mathrm{d}y \geq 0$. Поэтому любая функция $K(x,y)$, удовлетворяющая условию Мерсера, может быть использована для получения правила (43), что эквивалентно созданию оптимальной разделяющей плоскости в некотором признаковом пространстве.

Обучаемую НС, реализующую отображения в виде (43), называют нейронной сетью с векторами поддержки [29]. Использование различных выражений для внутреннего произведения в форме $K(\mathbf{x},\mathbf{x}_i)$ позволяет создавать различные НС этого типа с произвольным типом решающих поверхностей (нелинейных во входных пространствах) [30–32].

Например, сеть с радиальными базисными функциями (РБФ-сеть) [6, 33] и решающими функциями типа

[16)] По американской терминологии это — Support vector neural network. Реже используются термины Support vector machine (SVM) или SVM-type neural network.

$$f(\mathbf{x}) = \mathrm{sign}\left(\sum_{i=1}^{l} y_i \alpha_i \exp\left\{ - \frac{\| \mathbf{x} - \mathbf{x}_i \|^2}{\sigma^2} \right\} \right)$$

(где α_i, $i = 1, \ldots, l$ и σ — параметры РБФ-сети) может быть выполнена с использованием функций вида $K(\mathbf{x}, \mathbf{x}_i) = \exp\left\{ - \frac{\| \mathbf{x} - \mathbf{x}_i \|^2}{\sigma^2} \right\}$. В этом случае НС при обучении будет находить как центры $\mathbf{x}_i$, так и соответствующие веса α_i. Такая НС обладает некоторыми полезными свойствами:

- задача оптимизации этой НС имеет единственное решение;
- процесс обучения идет довольно быстро;
- использование введенного вида решающего правила позволяет в процессе обучения сети определить набор векторов поддержки;
- получение нового набора решающих функций достигается простым изменением только функций (ядра $K(\mathbf{x},\mathbf{x}_i)$), которые определяют скалярное произведение в признаковом пространстве $\mathbf{Z}$.

Способность к обобщению у нейронной сети прямого распространения (НС_ПР) и сети с векторами поддержки. Способность к обобщению у НС_ПР и НС с векторами поддержки (SVM) основана на рассмотренных элементах статистической теории обучения и на полученных оценках для скорости сходимости эмпирического риска к его действительной величине. Кроме того, чтобы гарантировать высокие показатели обобщения обучаемой сети, нужно построить структуру $\mathrm{S}_1 \subset \mathrm{S}_2 \subset \ldots \subset \mathrm{S}_n \subset \ldots \subset \mathrm{S}$ на наборе решающих функций $\boldsymbol{S} = \{Q(z,\alpha), \alpha \in \Lambda\}$ и затем выбрать как подходящий элемент $\boldsymbol{S}_l$ в структуре, так и функцию $Q(z,\alpha_n^l) \in \boldsymbol{S}_l$ в этом элементе, которая минимизирует границу (20). Граница (16) может быть переписана в простой форме:

$$R(\alpha_n^l) \le R_{\text{эмпир.}}(\alpha_n^l) + \Omega\left(\frac{n}{h_{l(n)}} \right), \qquad (45)$$

где первый член — это оценка риска, а второй является доверительным интервалом для этой оценки.

При создании НС определяется набор допустимых функций с некоторым значением РВЧ h^*. Для данного размера n обучающей выборки величина h^* определяет доверительный интервал $\Omega(n/h^*)$. Поэтому формирование НС связано с выбором структуры, подходящей для данного обучающего набора. В период обучения НС минимизируется первый член в границе (45) (количество ошибок на обучающем наборе).

Если при построении НС она будет выбрана слишком сложной (относительно доступного набора обучающих данных), то доверительный интервал $\Omega(n/h^*)$ будет большим. В случае если даже возможно минимизировать эмпирический риск до нуля, то количество ошибок на тестирующем наборе (т. е. при обобщении) может оказаться все же большим. Этот случай называют переподгонкой или избыточной подгонкой (под тонкую статистически случайную структуру обучающей выборки). Чтобы избежать избыточной подгонки (и получить малый доверительный интервал), следует стремиться создать НС с малой величиной РВЧ. Поэтому для получения хорошего обобщения у НС нужно, во-первых, предложить подходящую архитектуру НС и, во-вторых, настройкой параметров НС получить функцию отображения, которая минимизирует число ошибок на обучающих данных. Совместное решение этих задач для НС осуществляется на эвристической основе, или, по-простому говоря, с помощью интуиции и искусства исследователя.

В методах сетей с векторами поддержки можно управлять обоими параметрами: в случае задачи распознавания с разделимостью обучающих точек получается единственное решение, которое минимизирует эмпирический риск (возможно, вплоть до нуля) путем использования Δ–разделяющих гиперплоскостей с максимальным Δ-слоем (т.е. на основе получения набора отображений НС с наименьшей величиной РВЧ).

В общем случае для той же задачи получается единственное решение, когда выбирается сбалансированная величина параметра C в минимизируемом функционале $\Phi(\xi)$ с регуляризирующей компонентой, т. к. от C зависит предпочтительное соотношение между оценкой ошибки обобщения и ее доверительным интервалом.

ЗАКЛЮЧЕНИЕ

В рамках прикладной статистической теории обучения показан единообразный способ формализации группы задач, решаемых средствами нейронных сетей (НС) супервизорным методом ("с учителем"): распознавание образов, нелинейная регрессия и оценка плотности распределения вероятности. При этом применено вероятностное описание по входу и выходу НС с требованием ориентироваться не на сами вероятностные меры, а только на известные данные обучающей выборки. Три указанные задачи рассмотрены в терминах понятий: набор функций отображения НС, функция потерь и функционал риска. Все они параметризованы вектором α, компоненты которого представляют совокупность настраиваемых синаптических весов НС.

Следуя работам [1–5], используется нетрадиционное компактное представление обучающей выборки в форме $z_1, z_2, \ldots, z_n$ (где z_i объединяет

вход сети $\mathbf{x}_i$ и ее выход y_i), многообразия наборов (множеств) функций отображения НС $\{f(z,\alpha), \alpha\in\Lambda\}$, функций потерь $\{Q(z,\alpha), \alpha\in\Lambda\}$, фукционалов эмпирического риска $\{R_{\text{эмпир.}}(\alpha)\}$(среднего от функции потерь) и соответствующих функционалов риска $\{R(\alpha)\}$(функций потерь, взвешенных по вероятностной мере $P(z)$). В этом представлении даны отмеченные выше три основные задачи, решаемые с помощью НС.

Рассмотрены основные концепции элементов статистической теории обучения.

- Принцип минимизации эмпирического риска (принцип МЭР).
- Условия правомерности его применения в форме наличия сходимости к нулю вероятности максимального (по набору отображений) отличия величины $R(\alpha)$ и $R_{\text{эмпир.}}(\alpha)$. Фактический риск непосредственно или косвенно характеризует ожидаемую частоту ошибок НС при тестировании (ошибок на стадии обобщения).
- Базовые для статистической теории обучения понятия энтропии, VC-энтропии, модифицированной VC-энтропии и функции роста — для обучающей выборки или усредненно по вероятностному распределению обучающих данных, которые разным образом характеризуют меру многообразия набора функций потерь $\{Q(z,\alpha), \alpha\in\Lambda\}$ (или изоморфного ему набора функций отображения НС).
- Условие правомерности расширенного применения принципа МЭР, который использует понятие энтропии и условие быстрой сходимости к нулю отличия $R(\alpha)$ и $R_{\text{эмпир.}}(\alpha)$, выраженные через модифицированную VC-энтропию и функцию роста.
- Для скорости сходимости процесса обучения приведены границы, основанные на размерности Вапника—Червоненкиса и функции роста. Границы на скорость сходимости приспособлены к реальному (небольшому) количеству обучающих данных и получены в двух формах: толерантные границы, справедливые независимо от вида распределения вероятностей, и точные (зависящие от распределения) границы. Эти границы позволяют прогнозировать и в определенной мере влиять на показатели обобщения НС в процессе ее функционирования после завершения обучения.
- Рассмотрен принцип структурной минимизации риска, который предусматривает получение возможно лучших показателей обобщения НС путем одновременной минимизации $R_{\text{эмпир.}}$ и размерности Вапника—Червоненкиса набора функций потерь $\{Q(z,\alpha), \alpha\in\Lambda\}$(или изоморфного ему набора функций отображения НС).
- Изложены элементы теории построения алгоритмов обучения НС. Для задачи распознавания образов рассмотрен метод построения оптимальной разделяющей (гипер-) плоскости, использующий понятие "векторов поддержки". Метод приспособлен для линейно разделимой совокупности точек (образов) и имеет приближенную форму для случая не полностью разделимой совокупности образов.
- Показано, что для неразделимой совокупности образов целесообразно преобразование входного пространства в пространство признаков более высокой размерности, в котором уже может быть реализована линейная (с помощью гиперплоскости) разделимость образов. Для этого достаточно произведение векторов заменить некоторой симметричной функцией ("ядром") $K(\mathbf{x},\mathbf{y})$, удовлетворяющей условию Мерсера.
- Подход на основе построения Δ-разделяющей ("оптимальной") гиперплоскости и перехода в признаковое пространство более высокой размерности применяется в НС с векторами поддержки (support vector neural networks), сетях с радиальными базисными функциями (RBF-сетях) и может быть использован в сетях прямого распространения сигнала общего вида (без обратных связей).

Таким образом, рассмотренные элементы статистической теории обучения показывают, что "абстрактный" анализ помогает раскрытию общей модели обобщения, реализуемого нейронной сетью. Согласно этой модели, способность к обобщению обучаемой НС зависит от меры многообразия отображений у НС. Это понятие более емко, чем просто размерность пространства или число свободных параметров у функции потерь). Оно является основой в оценке границы различия эмпирического риска и ошибки обобщения НС в фазе ее функционирования.

Развитие SVM-методов продолжается в направлении уточнения границ различия, использующих оценки функции роста и РВЧ, расширения области применения структур НС с векторами поддержки (SV-структур НС) и создания ядер $K(\mathbf{x},\mathbf{y})$ с желательными свойствами инвариантности.

СПИСОК ЛИТЕРАТУРЫ

1. *Вапник В.Н., Глазкова Т.Г., Кощеев В.А., Михальский А.И., Червоненкис А.Я.* Алгоритмы и программы восстановления зависимостей. М.: Наука, 1984. 814 с.
2. *Vapnik V.N.* The Nature of Statistical Learning Theory. NY: Springer-Verlag, 1995. 188 p.
3. *Vidyasagar A.* A Theory of Learning and Generalization. NY: Springer-Verlag, 1997. 210 p.
4. *Poggio T., Girosi F.* Networks for approximation and learning // Proceedings of IEEE. 1990. V. 84. P. 1481–1497.

5. *Vapnik V.N.* Statistical Learning Theory. NY: Wiley, 1998. 736 p.
6. *Haykin S.* Neural Networks: A Comprehensive Foundation. Upper Saddle River, NY: Prentice-Hall, 1994. 646 p.
7. *Меркушева А.В.* Применение нейронной сети для текущего анализа нестационарного сигнала (речи). I. Основные принципы // Научное приборостроение. 2003. Т. 13, № 1. С. 64–71.
8. *Ту Дж., Гонсалес Р.* Принципы распознавания образов. М.: Мир, 1978. 411 с.
9. *Devroye L., Giorfi L., Lugosi G.* A Probability Theory of Pattern Recognition. NY: Springer-Verlag, 1996. 210 p.
10. *Вапник В.Н.* Оценка зависимостей на основе эмпирических данных. М.: Наука, 1979. 448 с.
11. *Blumer A., Ehrenfeucht D., Haussler D., Warmuth M.K.* Learning ability and the Vapnik—Chervonenkis dimension / J. ACM. 1989. V. 36, N 4. P. 929–965.
12. *Вапник В.Н.* Необходимые и достаточные условия для сходимости метода минимизации эмпирического риска // Сборник АН СССР по распознаванию, классификации и предсказанию. М.: Наука, 1989. Т. 2. С. 217–249.
13. *Vapnik V.N.* Principles of Risk Minimization for Learning Theory // Advances in Neural Information Processing Systems. San Mateo, CA.,1992. V. 4 / Kaufman. P. 831–838.
14. *Kearns M.J., Vizirani U.V.* An Introduction to Computational Learning. Cambridge, MA: MIT Press, 1994. 183 p.
15. *Alon N., David B., Cesa-Bianchi N., Haussler D.* Scale-Sensitive Dimensions, Uniform Convergence, and Learnability // J. ACM. 1997. V. 44. P. 617–631.
16. *Вапник В.Н.* Необходимые и достаточные условия для равномерной сходимости среднего к его ожиданию // Теория вероятностей и ее приложение. М.: Наука, 1981. Т. 26. С. 532–553.
17. *Burtlett P.L., Long P., Williamson R.C.* Fatt-Shattering and Learnability of Real-Valued Functions // Journ. Comput. Syst. Sci. 1996. V. 52, N 3. P. 434–452.
18. *Burtlett P.L., Shawe-Taylor J.* Generalization Performance on Support Vector Machine and other Pattern Classifiers // Advances in Kernal Methods — Supoprt Vector Learning / Eds.: Sholkopf B., Buges C., Smola A. Cambridge, MA: MIT Press, 1999. 167 p.
19. *Boser B., Guyon I., Vapnik V.* A Training Algorithm for Optimal Margin Classifiers // Proceedings of 5th Annual Workshop on Computation Learning Theory. Pittsburgh, PA: ACM, 1992. P. 144–152.
20. *Opper M.* On the Annealed VC Entopy for Margin Classifier: a Statistical Mechanics Study // Advances in Kernal Methods — Supoprt Vector Learning / Eds.: Sholkopf B., Buges C., Smola A. Cambridge, MA: MIT Press, 1999. 167 p.
21. *Hasson M.M.* Fundamentals of Artificial Neural Networks. Cambridge, MA: MIT Press, 1995. 186 p.
22. *Меркушева А.В.* Применение нейронной сети для текущего анализа нестационарного сигнала (речи). II. Исследование и оптимизация нейронной сети // Научное приборостроение. 2003. Т. 13, № 1. С. 72–84.
23. *Вапник В.Н.* Теория распознавания образов. М.: Наука, 1974. 353 с.
24. *Дуда Р., Харт П.* Распознавание образов и анализ сцен. М.: Мир. 176 с.
25. *Тихонов А.Н., Гончарский А.В., Степанов В.В., Ягола А.Г.* Регуляризующие алгоритмы и априорная информация. М.: Наука, 1983. 198 с.
26. *Морозов В.А., Гребенников А.И.* Методы решения некорректно поставленных задач // Алгоритмические методы. М.: МГУ, 1992. 320 с.
27. *Cover T.M., Thomas J.A.* Elements of Information Theory. NY: Wiley, 1991. 396 p.
28. *Колмогоров А.Н., Фомин С.В.* Элементы теории функционального анализа. М.: Наука, 1989. 79 с.
29. *Girosi F., Jones M., Poggio T.* Regularization Theory and Neural Networks Architectures // Neural Computations. 1995. V. 7, N 2. P. 219–269.
30. *Burges C.G.* Simplified Support Vector Decision Rule // Proceedings of 13th Intern. Conference on Machine Learning. San Matteo, CA, 1996. P. 71–77.
31. *Cortes C., Vapnic V.* Support Vector Networks // Machine Learning. 1995. V. 20. P. 273–297.
32. *Girosi F.* An Equivalence Between Sparse Approximation and Support Vector Machine // Neural Computations. 1998. V. 10, N 6. P. 1455–1480.
33. *Fung C.F.* On Line Adaptive Training Using Radial Basis Functions // Neural Networks. 1996. V. 9, N 9. P. 1597–1618.

Санкт-Петербург

Материал поступил в редакцию 10.12.2004.

ELEMENTS OF THE STATISTICAL LEARNING CONCEPT FOR A NEURAL NETWORK AND ACCURATE PREDICTION OF ITS OPERATION

G. F. Malychina, A. V. Merkusheva

Saint-Petersburg

The learning of neural networks (NN) for many problems (pattern recognition, nonlinear multi-parameter regression, probability distribution identification) is considered in generalized form on the basis of a concept that includes probabilistic interpretation for the NN input—output transfer function and basic notions having a mathematically formalized foundation: diversity (a set) of mapping being realized by NN (and a set of loss functions isomorphic to it); characteristics of that diversity on the basis of entropy and Vapnik—Chervonenkis dimension; risk functional (RF) and a condition allowing RF approximation by means of an empirical risk functional (ERF); the limits of the actual RF departure from ERF. The elements of the leaning statistical theory described here provide prediction and correction ("control") of the NN operation index after leaning, i.e. at the stage of NN testing with the data on not participating in learning.