УДК 621.396.96

В.И. Кошелев, Т.Д. Нгуен

ОБУЧЕНИЕ МНОГОСЛОЙНЫХ НЕЙРОННЫХ СЕТЕЙ НА ОСНОВЕ АЛГОРИТМА ОБРАТНОГО РАСПРОСТРАНЕНИЯ ОШИБОК ПРИ РАСПОЗНАВАНИИ ВОЗДУШНЫХ ОБЪЕКТОВ

Для задачи распознавания воздушных объектов по дальностным портретам выполнен сравнительный анализ различных алгоритмов обучения многослойных нейронных сетей. Показано, что для решения данной задачи в большинстве случаев лучшим является алгоритм сопряженных градиентов. Определено оптимальное значение скорости обучения алгоритмов градиентного спуска.

Введение. Вопросам обучения многослойных нейронных сетей (МНС) в последние годы уделяется большое внимание. Во многом проблема обучения МНС связана с определением метода корректировки весов. Только в 1986 году Румелхарт и др. смогли решить окончательно эту проблему, что дало новый толчок к применению нейронных сетей в практических задачах. Созданный метод корректировки весов называется алгоритмом обратного распространения ошибок (ОРО) [1…3]. Многие модификации алгоритмов созданы на основе метода ОРО. Возможность их практического использования, как правило, связана с особенностями конкретной задачи.

Цель работы. Целью работы является определение оптимального выбора параметров обучения алгоритма ОРО при решении задачи распознавания воздушных объектов.

Математическая интерпретация алгоритма ОРО. Структура МНС представлена на рис. 1. Количество слоев и нейронов в скрытых слоях зависит от сложности решаемой задачи [4]. Слой 0 представляет собой входной слой и играет роль сенсора, принимающего внешние сигналы. Слой M является выходным, число нейронов ровно числу входов нейронных сетей. В частности, при решении задачи классификации оно равно числу классов. Каждый вход слоя 0 и каждый нейрон скрытых слоев связывается со следующими нейронами весовыми коэффициентами. На рис. 1 показан весовой коэффициент синаптической связи w_{kj}^m, соединяющий нейрон k слоя m и нейрон j слоя m-1.

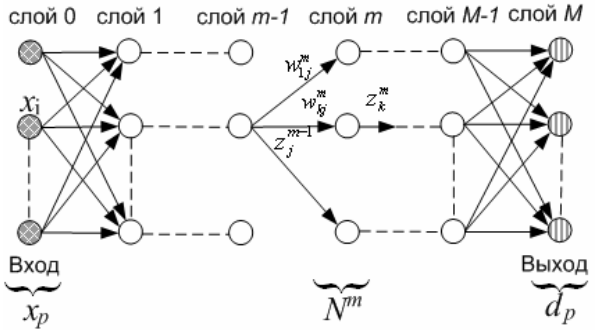

Рис. 1. Структура многослойных нейронных сетей

В каждом нейроне выполняется суммирование взвешенных данных, поступающих на его вход, затем формируется функция активации:

$$z_k^m = f\left(\sum_{j=1}^{N^{m-1}} w_{kj}^m z_j^{m-1}\right). \qquad (1)$$

Такие вычисления выполняются последовательно по каждому слою. Для каждого слоя, кроме нулевого слоя, матрица весов имеет вид

$$W^m = \begin{bmatrix} w_{11}^m & \cdots & w_{1N^{m-1}}^m \\ \vdots & \ddots & \vdots \\ w_{N^m 1}^m & \cdots & w_{N^m N^{m-1}}^m \end{bmatrix}. \qquad (2)$$

Заметим, что все весы, связанные с отдельным нейроном, находятся в одной строке. Выходной вектор (m -1)-го слоя

$$(z^{m-1})^T = (z_1^{m-1} \cdots z_{N^{m-1}}^{m-1}). \qquad (3)$$

Выходной вектор слоя m может быть вычислен как:

$$(z^m)^T = f((a^m)^T) = (f(a_1^m) \cdots f(a_{N^m}^m)), \qquad (4)$$

где $a^m = w^m z^{m-1}$.

Для входного вектора x и соответствующей желаемому выходу d_p ошибка оценивается величиной:

$$E = \frac{1}{2}\sum_{q=1}^{N^M}\left[z_q^M(x) - d_q(x)\right]^2, \qquad (5)$$

где z_q^M – выход нейрона q выходного слоя M.

В общем случае, если на входах нейронных сетей (НС) присутствует все множество p образов, полная ошибка равна:

$$E_\Sigma = \frac{1}{2}\sum_{p=1}^{P}E_p = \sum_{p=1}^{P}\sum_{q=1}^{N^M}\left[z_q^M(x_p) - d_q(x_p)\right]^2. \qquad (6)$$

Вектор градиента $\nabla E = \left\{\partial E / \partial w_{ji}^m\right\}$ показывает направление к максимуму среднего квадрата ошибки (СКО). Значение ∂w_{ji}^m должно изменяться в направлении минимума СКО. Обозначим дискретные моменты времени, соответствующие одному шагу итерационной процедуры, как t. Тогда веса будут корректироваться в момент времени $t+1$ и

$$w_{ji}^m(t+1) = w_{ji}^m(t) - \mu\sum_{p=1}^{P}\left.\frac{\partial E_p}{\partial w_{ji}^m}\right|_{W(t)}, \qquad (7)$$

где μ – коэффициент скорости обучения, $0 < \mu < 1$. Формула (7) в краткой форме имеет вид:

$$W(t+1) = W(t) - \mu\nabla E. \qquad (8)$$

Для каждого слоя (кроме слоя 0) градиентная матрица ошибок равна

$$(\nabla E)^m = \begin{bmatrix} \frac{\partial E}{\partial w_{11}^m} & \cdots & \frac{\partial E}{\partial w_{1N^{m-1}}^m} \\ \vdots & \ddots & \vdots \\ \frac{\partial E}{\partial w_{N^m 1}^m} & \cdots & \frac{\partial E}{\partial w_{N^m N^{m-1}}^m} \end{bmatrix}. \qquad (9)$$

Для каждого слоя (кроме выходного слоя) градиент ошибок, связанный с выходами нейронов, может быть определен как

$$\nabla_{Z^m} E = \left(\frac{\partial E}{\partial z_1^m} \cdots \frac{\partial E}{\partial z_{N^{m-1}}^m}\right). \qquad (10)$$

Для выходного слоя $\nabla_{Z^m} E$ легко определить, так как известны выходы НС $z^M(x_p)$ и вектор цели d_p. Учитывая функцию ошибок, функцию активации $f(\cdot)$ и полную производную $f(\cdot)'$, градиент ошибок можно вычислить по следующей рекурсивной формуле [3]:

$$\nabla_{Z^m} E = (W^{m+1})^T\left[\nabla_{Z^{m+1}} E \otimes f'(a^{m+1})\right], \qquad (11)$$

где знаком $\otimes$ обозначена операция умножения по элементу.

Вычисление проводится от M-1 до 1, для входного 0 и выходного слоя M:

$$(\nabla E)^m = \left[\nabla_{Z^m} E \otimes f'(a^m)\right](z^{m-1})^T, \qquad (12)$$

где $z^0 = x$. Вектор градиента ∇E можно выразить через выходы нейрона j, то есть $z^m{}_j$ как

$$\frac{\partial E}{\partial w_{ji}^m} = \frac{\partial E}{\partial z_j^m}\frac{\partial z_j^m}{\partial w_{ji}^m}, \qquad (13)$$

$$\frac{\partial z_j^m}{\partial w_{ji}^m} = \frac{\partial}{\partial w_{ji}^m}\left[f\left(\sum_{l-1}^{N^{m-1}} w_{jl}^m z_l^{m-1}\right)\right] =$$

$$= f'\left(\sum_{l-1}^{N^{m-1}} w_{jl}^m z_l^{m-1}\right)\frac{\partial}{\partial w_{ji}^m}\left[f\left(\sum_{l-1}^{N^{m-1}} w_{jl}^m z_l^{m-1}\right)\right] =$$

$$= f'(a_j^m).z_i^{m-1}. \qquad (14)$$

В свою очередь, $\partial E / \partial z_j^m$ можно выразить через выходы z_j^{m+1} слоя $m+1$:

$$\frac{\partial E}{\partial z_j^m} = \sum_{l=1}^{N^{m+1}}\frac{\partial E}{\partial z_l^{m+1}}\frac{\partial z_l^{m+1}}{\partial z_j^m} =$$

$$= \sum_{l=1}^{N^{m+1}}\frac{\partial E}{\partial z_l^{m+1}}\frac{\partial}{\partial z_j^m}\left[f\left(\sum_{q=1}^{N^m} w_{lq}^{m+1} z_q^m\right)\right] =$$

$$= \sum_{l=1}^{N^{m+1}}\frac{\partial E}{\partial z_l^{m+1}} f'\left(\sum_{q=1}^{N^m} w_{lq}^{m+1} z_q^m\right)\frac{\partial}{\partial z_j^m}\left(\sum_{q=1}^{N^m} w_{lq}^{m+1} z_q^m\right) =$$

$$= \sum_{l=1}^{N^{m+1}}\frac{\partial E}{\partial z_l^{m+1}} f'(a_l^{m+1} w_{lj}^{m+1}) w_{lj}^{m+1}. \qquad (15)$$

Для слоев 0 и M

$$\frac{\partial E}{\partial w_{ji}^m} = \frac{\partial E}{\partial z_j^m}\frac{\partial z_j^m}{\partial w_{ji}^m} = \left[\sum_{l=1}^{N^{m+1}}\frac{\partial E}{\partial z_l^{m+1}} f'\left(a_l^{m+1} w_{lj}^{m+1}\right)\right] f'\left(a_j^m z_i^{m-1}\right). \qquad (16)$$

На практике в качестве начальных условий для весов выбираются случайные значения в интервале (-1…1). Коэффициент μ определяется экспериментально. При малых значениях μ, обучение будет длительным и может привести к точке локального минимума. Если, напротив, значение μ велико, обучение происходит быстро, но может быть пропущен глобальный минимум. Другой проблемой является осцилляция в окрестности точки минимума.

Экспериментальные исследования. Применим алгоритм ОРО для решения задачи распознавания 5 воздушных объектов (ВО) с различными дальностными портретами (ДП). При этом создаем обучающее и контрольное множество методом математического моделирования [5, 6]. Каждое множество составляет 800 ДП. Используем трехслойную нейронную сеть с одним скрытым слоем. Число нейронов в скрытом слое

выбираем по рекомендации [2] равным половине количества входов и выходов, т.е. равным 50. Число входов равно числу отсчетов ДП, т.е. 90. Число нейронов в выходном слое равно 5 (по количеству ВО). Таким образом, структура нейронной сети описывается числами 90-50-5.

Обучение НС производится с помощью алгоритма градиентного спуска (GD). Зависимость коэффициента, характеризующего скорость обучения μ от ошибки обучения, показана на рис. 2.

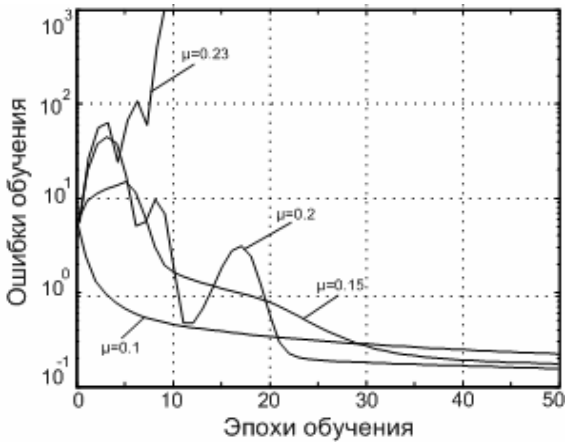

Рис. 2. Выбор скорости обучения

Из рис. 2 очевидно, что при $\mu > 0.1$ возникает осцилляция ошибки в начале этапа обучения, при этом амплитуда осцилляций увеличивается с увеличением μ. При $\mu \geq 0.23$ возрастание ошибки обучения приводит к тому, что процедура расходится. С учетом случайного характера начальных значений весов следует выбирать значение $\mu \leq 0.1$, что обеспечивает сходимость процедуры.

Другой проблемой, решаемой при обучении, является повышение скорости сходимости. Известно, что значение весов прямо пропорционально значению ошибки обучения. С уменьшением значения ошибки обучения уменьшается значение корректировки весов, вследствие этого алгоритм градиентного спуска медленно сходится. Решение этой проблемы привело к созданию следующего ряда алгоритмов.

Алгоритм градиентного спуска с возмущением (GDM)

$$W(t+1) = mcW(t) - (1-mc)\mu\nabla E, \qquad (17)$$

которой отличается от GD наличием параметра *mc* возмущения. При *mc*=0 изменение вектора весов определяется только градиентом, при *mc*=1 текущее приращение равно предшествующему приращению. В табл. 1 приведены результаты GDM для *mc* = 0.9.

Таблица 1

Алгоритм	Время обучения	Число эпох
GD	106 сек.	1913
GDM	131 сек.	2376
GDA	101 сек.	1875
Rprop	5 сек.	46
Флетчер-Ривс	9 сек.	51
Полак-Рибейр	12 сек.	78
Повел-Беал	7 сек.	40

Для алгоритма градиентного спуска с выбором параметра скорости (GDA), если отношение текущего значения ошибки к предыдущему превышает величину *max*, то новые значения настраиваемых параметров не принимаются. При этом μ уменьшается с коэффициентом $k1$. Если новая ошибка меньше предыдущей, то μ увеличивается с коэффициентом $k2$. Результаты приведены в табл. 1 для $k1$=0.7, $k2$=1.2.

Особенностью порогового алгоритма ОРО (Rprop) является то, что значение приращения настраиваемого параметра увеличивается пропорционально значению Δ1, когда градиент ошибки сохраняет знак для двух последовательных эпох. Значение приращения уменьшается пропорционально значению Δ2, когда градиент ошибки изменяет знак по сравнению с предыдущей эпохой. Если градиент ошибки равен нулю, то приращение остается неизменным. В результатах табл. 1 задано Δ1=1.2, Δ2=0.5.

Все алгоритмы метода сопряженных градиентов на первой интеграции начинают поиск в направлении антиградиента

$$P(0) = -\nabla E(0). \qquad (18)$$

Когда выбрано направление, требуется определить оптимальное расстояние (шаг поиска), на величину которого следует изменить настраиваемые параметры:

$$W(t+1) = W(t) + \alpha(t)P(t). \qquad (19)$$

Затем определяется следующее направление поиска как линейная комбинация нового направления наискорейшего спуска и вектора движения в сопряженном направлении:

$$P(t) = -\nabla E(t) + \beta(t)P(t-1). \qquad (20)$$

Различные алгоритмы метода сопряженного градиента отличаются способом вычисления значения $\beta(t)$. Флетчер-Ривс предложил вычислять $\beta(t)$ по следующей формуле:

$$\beta(t) = W(t)^T W(t) / W(t-1)^T W(t), \qquad (21)$$

а Полак-Рибейре по формуле

$$\beta(t) = \Delta W(t)^T W(t) / W(t-1)^T W(t). \qquad (22)$$

В случае, когда возникает условие плохой сходимости или количество итераций превысило число настраиваемых параметров сети, Повел - Беал предложил заново формировать направление антиградиента. Условием для рестарта является:

$$\left\|W(t-1)^T W(t)\right\| \geq 0.2\left\|W(t)\right\|^2. \qquad (23)$$

Алгоритм на основе метода Ньютона вычисляет корректируемые параметры по формуле:

$$W(t+1) = W(t) - H^{-1}(t)\nabla E(t), \qquad (24)$$

где H – матрица Гессе вторых частных производных функционала ошибки по настраиваемым параметрам. Вычисление матрицы Гессе весьма сложно, поэтому разработан ряд квазиньютоновских алгоритмов. В данном эксперименте не удается использовать квазиньютоновские алгоритмы, т.к. они требуют слишком большого объема памяти.

В табл. 1 приведено сравнение результатов использования рассмотренных алгоритмов, при этом ошибка обучения задана на уровне 0,05 для всех алгоритмов. Кроме алгоритма Rprop, преимущество алгоритмов сопряженных градиентов перед алгоритмами градиентного спуска очевидно. Сложностью использования Rprop является выбор коэффициентов $\Delta 1$ и $\Delta 2$.

При решении данной задачи возникает вопрос, как правильно обучить НС для того, чтобы она хорошо распознавала эти объекты в реальных условиях, т.е. при наличии помех. Приведем обучение НС с добавленными множествами различного уровня шума. В табл. 2 показаны результаты вероятности правильного распознавания 5 воздушных объектов. В результате использовался алгоритм Повел - Беал для обучения НС, количество эпох для каждого процесса обучения - 1000.

По результатам вычислений в табл. 2 можно сделать вывод о том, что лучшим является обучение НС с добавленными множествами уровня сигнал-шум до 5 дБ. Обучение НС до уровня сигнал-шум 30 дБ не приводит к лучшим результатам распознавания, потому что до этого уровня НС ещё не обеспечивает условие обобщения.

Таблица 2

Уровень сигнал-шум при обучении	Отношение сигнал-шум, дБ				
	5	10	20	30	40
	Вероятность правильного распознавания				
30 дБ	0.595	0.794	0.903	0.911	0.913
20 дБ	0.725	0.855	0.899	0.903	0.904
10 дБ	0.845	0.893	0.907	0.909	0.909
5 дБ	0.873	0.906	0.915	0.916	0.916
0 дБ	0.857	0.882	0.891	0.893	0.893

Выводы. Вычислительные эксперименты показывают, что: обучение многослойных нейронных сетей является процедурой, которая требует большого числа экспериментов для рационального выбора параметров и алгоритмов обучения. Для задачи распознавания воздушных объектов скорость обучения алгоритмов градиентного спуска следует выбирать при значении $\mu \leq 0.1$. Алгоритмы метода сопряженных градиентов работают в десятки раз быстрее, чем алгоритмы градиентного спуска. Для работоспособности сети в реальных условиях обучение проводится до уровня отношения сигнал-шум не менее 5 дБ.

Библиографический список

1. *Хайкин Саймон*. Нейронные сети: полный курс, 2-е издание: Пер. с англ. М.: Издательский дом «Вильямс», 2006. 1104 с.

2. *Медведев В. С. и Потемкин М. Г.* Нейронные сети Матав-6 – М.: Диалог-МИФИ, 2002.-489 с.

3. *Michael A. Arbib*. The Handbook of Brain Theory and Neural Networks. Massachusetts Institute of technology, 2003. 1290 p.

4. *Галушкин А. И.* Решение задач в нейросетевом логическом базисе// Нейрокомпьютеры. -2006. № 2. С. 49-70.

5. *Shirman Y. D., Gorshkov S. A., Leshchenko S. P., Orlenko V. M., Sedyshev S. Yu.* “Radar Target Backscattering Simulation - Sofware and User's Manual”. Artech House, 2002. 290 p.

6. *Ширман Я. Д., Горшков С. А., Лещенко С. П., Братченко Г. Д., Орленко* В. М. Методы радиолокационного распознавания и их моделирование // Зарубежная радиоэлектроника. Успехи современной радиоэлектроники. 1996. № 11. С. 3 – 63.