УДК 004.896

СИНТЕТИЧЕСКИЙ АЛГОРИТМ ОПТИМИЗАЦИИ И НАСТРОЙКИ НЕЙРОННЫХ СЕТЕЙ

И.Ф.ЯСИНСКИЙ

(Ивановская государственная текстильная академия)

В современных наукоемких технологиях, включая текстильные, все более широкое распространение находят нейросетевые технологии. Быстрота обучения и реагирования таких систем при этом имеют определяющее значение.

Задача настройки нейронной сети является задачей многоэкстремального поиска. При этом одним из самых распространенных методов настройки является метод обратного распространения ошибки, полученный в результате применения метода градиента для минимизации целевой функции ошибки распознавания. При использовании этого метода часто возникают ситуации, когда процесс обучения, достигнув некоторого уровня, останавливается. Происходит это из-за того, что поисковая точка попадает в один из локальных минимумов целевой функции, из которого выйти уже не может.

Для решения этой проблемы нами разработан синтетический алгоритм, включающий следующие составляющие его алгоритмы: усовершенствованный случайный поиск, случайный поиск с памятью, метод инерционной минимизации и две разновидности генетического поиска.

Численные эксперименты позволили убедиться в существенном увеличении быстродействия синтетического алгоритма по сравнению с традиционным подходом при поиске глобального минимума на множестве локальных, что принципиально важно для решения задачи обучения нейронной сети.

Рассмотрим подробнее предложенные нами модификации составляющих алгоритмов поиска, включенных в синтетический.

Усовершенствованный случайный поиск. Множество весов нейронной сети можно рассматривать как вектор $\vec{W}$ в N-мерном пространстве. Очевидно, целевая функция Q ошибки распознавания зависит от значений весов $\vec{W}$. Пусть $\vec{W}^0 = (W_1^0, ..., W_N^0)$ – исходная точка в пространстве поиска. Последовательность $(W_1^k, ..., W_N^k)$; к = 0,1,…,M назовем последовательностью удачных точек, если в каждой следующей точке целевая функция $Q(\vec{W})$ меньше, чем в предыдущей.

Последовательность удачных точек получается так. Около очередной удачной точки строится окрестность в виде N-мерного прямоугольного "ящика" с полуразмерами DW^k_m. Из этой окрестности выбирается пробная случайная точка $\vec{W}$ с координатами

$$W_m = W_m^k + \zeta DW_m^k;\ m = 1,2,\ldots,N, \quad (1)$$

где ζ – случайные числа, равномерно распределенные в интервале (-1, +1).

В этой точке вычисляется значение целевой функции $Q(\vec{W})$, которое сравнивается с ее значением в последней удачной точке $Q(\vec{W}^k)$.

Если $Q(\vec{W}) < Q(\vec{W}^k)$, то такая точка объявляется новой удачной

$$Q^{k+1} = Q(\vec{W}),\ W_m^{k+1} = W_m;\ m=1,2,\ldots,N, \quad (2)$$

и она становится центром новой окрестности с полуразмерами DW^{k+1}_m.

Следующая пробная точка будет взята из этой новой окрестности. Последовательность удачных точек $\vec{W}^k$, очевидно, сходится к точке минимума целевой функции $Q(\vec{W})$.

Случайный поиск эффективен, если удачно выбраны размеры окрестности DW^k_m, из которой берутся случайные пробы, и они оптимально изменяются по мере приближения этой окрестности к точке минимума.

Мы использовали следующий способ выбора начальных размеров окрестности поиска:

$$DW_m^0 = 0{,}28 h_m / \sqrt{\varepsilon^2 + \left|Q_m^+ - 2Q^0 + Q_m^-\right|} / N, \quad (3)$$

где $Q^0 = Q(W_1^0,\ldots,W_N^0)$; $Q_m^+ = Q(W_1^0,\ldots,W_m + h_m,\ldots,W_N^0)$; $Q_m^- = Q(W_1^0,\ldots,W_m - h_m,\ldots,W_N^0)$; h_m – малые пробные приращения аргументов W_m^0; ε – малое положительное число (защита от нуля в знаменателе).

В процессе поиска размеры окрестности DW^k_m мы изменяли следующим образом. Выполнена серия из M_S проб. Из них ν проб оказались удачными. Вычисляем эмпирически полученный множитель:

$$\mu = 1 + 2{,}8(\nu / M_S - 1/4). \quad (4)$$

По окончании указанной серии проб все размеры окрестности DW^k_m умножаем на множитель μ. Согласно такому алгоритму пробная окрестность по мере приближения к точке минимума сжимается. Поиск прекращается, когда для всех весов начинает выполняться условие $DW^k_m < E$, где E – допустимая погрешность в определении координат минимума.

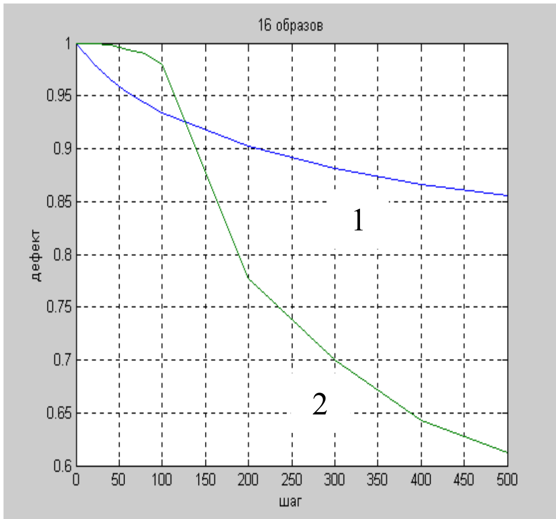

Рис. 1

На рис.1 представлены процессы обучения нейронной сети распознаванию образов по методу обратного распространения ошибки (кривая 1) и при помощи усовершенствованного случайного поиска (кривая 2). На вертикальной оси отложено отношение текущего значения суммы модулей невязок к ее начальному значению в зависимости от длины обучающей последовательности.

Преимущество усовершенствованного метода случайного поиска очевидно.

Случайный поиск с памятью содержит блок, вычисляющий значения "стохастического градиента" от целевой функции при случайных пробах (удачных и неудачных). Окрестность, из которой берется следующая случайная проба, смещается в направлении стохастического антиградиента:

$$\vec{G} = -\sum_{i=1}^{n} \frac{Q\left(\vec{W}_i\right) - Q\left(\vec{W}^k\right)}{\left|\delta\vec{W}_i\right|^2} \delta\vec{W}_i, \qquad (5)$$

где $\delta\vec{W}_i = \vec{W}_i - \vec{W}^k$ – случайный сдвиг пробной точки относительно к-й удачной.

В численном эксперименте нами найдены значения коэффициентов α_i, которые учитывают успешность предыдущих шагов. Эти коэффициенты используются для предварительного сдвига окрестности поиска, из которой берутся случайные пробы:

$$\vec{W}^k_{new} = \vec{W}^k + \tau(\alpha_1\vec{G}^k + \alpha_2\vec{G}^{k-1} + \alpha_3\vec{G}^{k-2}), \quad (6)$$

$$\alpha_1 + \alpha_2 + \alpha_3 = 1, \qquad (7)$$

где $\vec{W}^k_{new}$ – новые координаты центра окрестности, из которой берутся случайные пробы; τ – шаг по времени

Алгоритм инерционной минимизации. Численные эксперименты показали, что эффективной является также следующая методика настройки весовых коэффициентов нейронной сети. Подчиним движение поисковой точки в пространстве весовых коэффициентов следующим дифференциальным уравнениям:

$$\left.\begin{aligned} \frac{dV_s}{dt} &= G_S - \mu V_S, \\ \frac{dW_s}{dt} &= V_S;\ S = 1,2,...,N, \end{aligned}\right\}, \qquad (8)$$

где V_S – составляющие скорости поисковой точки; μV_S – составляющие силы трения, которая будет гасить колебания поисковой точки; μ – "вязкость среды".

Согласно (8) поисковая точка обладает инерцией, которая позволит ей выходить из локальных минимумов и находить глобальный минимум. Если через F(x) обозначить передаточную функцию нейрона и взять ее в виде F(x) = 1 / (1 + exp(–x)), то, воспользовавшись неявной разностной схемой, можно получить следующие итерационные выражения:

$$V^{n+1}_{ij} = \left(V^n_{ij} - \tau E_j Y_j Y_i \left(1 - Y_i\right)\right) \Big/ \left(1 + \tau\mu\right), \quad (9)$$

$$W^{n+1}_{ij} = W^n_{ij} + \tau V^{n+1}_{ij}, \qquad (10)$$

где $E_i = Y_i - Y_i^*$, Y_i, Y_i^* – действительная и требуемая реакция нейронной сети; W_{ij} – весовые коэффициенты; τ – шаг по времени.

Генетический алгоритм брался в следующих двух видах: традиционном и модифицированном. При традиционном подходе было сформировано множество наборов хромосом, представляющих веса нейронной сети. Текущий массив весов W_m и случайно выбранный набор весов $W^{/}_m$ составляют "родительскую" пару.

В массив весов-"потомков" входит первая половина весовых коэффициентов текущего массива и вторая половина от выбранного набора. Далее, после предъявления сети обучающего образа и получения на выходе ответа, производится сравнение: если целевая функция ошибки после генетических операций уменьшилась $Q'(\vec{W})<Q(\vec{W})$, то тестируемый массив весов становится текущим (действующим). В противном случае, то есть при $Q'(\vec{W})>Q(\vec{W})$, тестируемый массив стирается.

Был использован также следующий модифицированный вариант генетического алгоритма. В пространстве $\vec{W}$ выбирается некоторое количество исходных точек $W^f_r\ (f = 1,2,...)$. Они были получены с помощью усовершенствованного случайного поиска, повторением поиска из различных начальных позиций. Это "родители", образующие пары.

Пары строились следующим образом. Целевые функции для "родителей" должны быть близки, а расстояние между ними максимально:

$$\left.\begin{aligned} &\left|Q^f - Q^l\right| < \varepsilon_Q \to \min, \\ &\sum_{r=1}^{R} (W_r^f - W_r^l)^2 = R^2 \to \max. \end{aligned}\right\} \qquad (11)$$

Пара дает "потомство". Для этого координаты точек "родителей" соединяются прямой согласно выражению:

$$W_i = W_i^f + \eta(W_i^l - W_i^f), \qquad (12)$$

где η – одномерная координата, по которой затем осуществляется глобальный поиск. Для этого применялись методы интерполяции и Р.Г. Стронгина. Таким образом, многомерный поиск заменяется одномерным, что ускоряет вычисления.

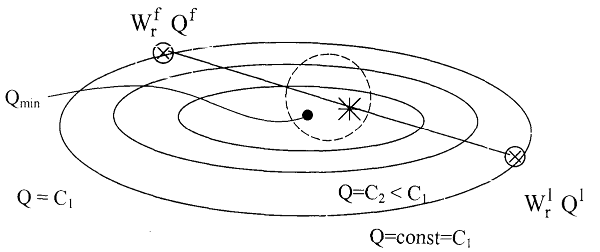

Рис. 2

Рис.2 поясняет идею такого поиска. $\otimes$ – "родительские" точки; $*$ – точки-"потомки"; Q = const – линии равного значения целевой функции; Q_{min} – значение целевой функции в минимуме; W_r^f, W_r^l, Q^f, Q^l – координаты "родительских" точек и значение в них целевой функции.

Точки-"потомки" находятся ближе к искомому минимуму, чем "родительские" точки. "Потомков" должно быть не меньше, чем "родителей" с тем, чтобы этот процесс можно было продолжить неограниченно.

В проведенных нами численных экспериментах последовательно чередовалось обучение нейронной сети по методу случайного поиска и минимизация ошибки по генетическому алгоритму. Данный вариант показал большую эффективность по сравнению со случайным поиском и генетическим алгоритмом в отдельности.

Следует обратить внимание, что как случайный поиск, так и генетический алгоритм легко поддаются распараллеливанию и поэтому удобны для реализации на многопроцессорных вычислительных машинах.

Сравнение комбинированного алгоритма генетического и усовершенствованного случайного поиска со случайным алгоритмом на основе распознавания образов нейронной сетью представлено на графике рис.3 – обучение нейронной сети распознаванию 8 образов.

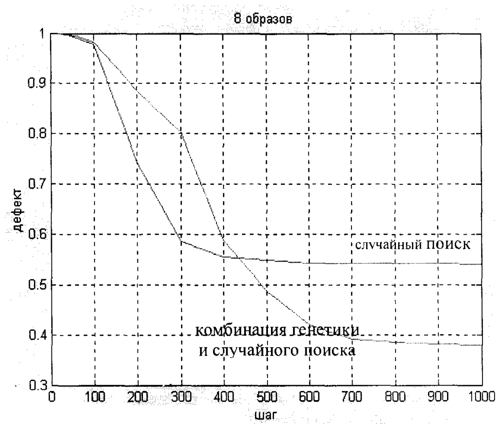

Рис. 3

Нами выполнялись численные эксперименты по определению наилучшей последовательности указанных выше составляющих алгоритмов в *синтетическом алгоритме.* Для оценки эффективности синтетического алгоритма, помимо настройки нейронных сетей, изучалась также минимизация эталонных целевых функций (функция Розенброка и специальные тестовые функции с множеством локальных минимумов).

Пусть С – усовершенствованный случайный поиск; П – случайный поиск с памятью; И – инерционный поиск; Г – генетический алгоритм. В ходе экспериментов было отмечено, что последовательность, позволяющая найти глобальный минимум с наименьшими затратами машинного времени, чаще всего имеет следующий вид:

ПГСИ ПГСИ ПГСИ ……

ВЫВОДЫ

1. Предложен синтетический алгоритм оптимизации и настройки нейронных сетей, включающий модифицированные варианты усовершенствованного случайного поиска, случайного поиска с памятью, инерционного и генетического поисков.

2. Данный алгоритм позволяет находить глобальный минимум среди множества локальных, ускоряет обучение нейронных сетей и улучшает качество распознавания образов.

Рекомендована кафедрой прикладной математики и информационных технологий. Поступила 11.12.06.