УДК 519.8

А.В. Скороходов, А.В. Тунгусова

Сравнительный анализ градиентных методов минимизации в задаче обучения многослойного персептрона

Приводится сравнительный анализ градиентных методов минимизации целевой функции ошибки в задаче обучения нейронной сети. Представлены описание и характеристики этих методов. Результаты анализа и численных экспериментов показали, что наиболее эффективным для обучения нейронной сети является метод сопряженных градиентов.

Ключевые слова: нейронная сеть, градиентные методы.

Процедура классификации облачности и подстилающей поверхности (ПП) по типам необходима для решения различных научных и практических задач. С развитием космических систем дистанционного зондирования Земли стало возможным проведение широкомасштабного мониторинга атмосферы, земной и водной поверхности. Использование информации о текстуре является одним из возможных вариантов описания объектов на спутниковых снимках. Для настройки нейронной сети (НС) необходимы эталонные образцы текстур облачности и ПП. Задача обучения НС заключается в коррекции весовых коэффициентов нейронов таким образом, чтобы при предъявлении НС схожих по текстуре образцов она относила их к определенному типу облачности и ПП.

Алгоритм обратного распространения ошибки является самым распространенным методом обучения многослойного персептрона и основан на вычислении целевой функции ошибки, которую необходимо минимизировать. Для этого применяются градиентные методы (ГМ) минимизации. В классическом алгоритме [1] используется метод наискорейшего спуска. Однако можно применять и другие ГМ, которые рассматриваются в данной статье. При этом ГМ имеют ряд ограничений и особенностей при их использовании для обучения НС. Целью данной работы является сравнительный анализ ГМ минимизации в задаче обучения НС для классификации облачности и ПП по типам на основе спутниковой информации.

Описание нейронной сети

Для классификации облачности и ПП использовался трехслойный персептрон, архитектура которого показана на рис. 1. Сеть состоит из двух скрытых и выходного слоев. Первый скрытый слой содержит F нейронов, второй − S нейронов, а число нейронов выходного слоя L равно числу классифицируемых типов облачности и ПП.

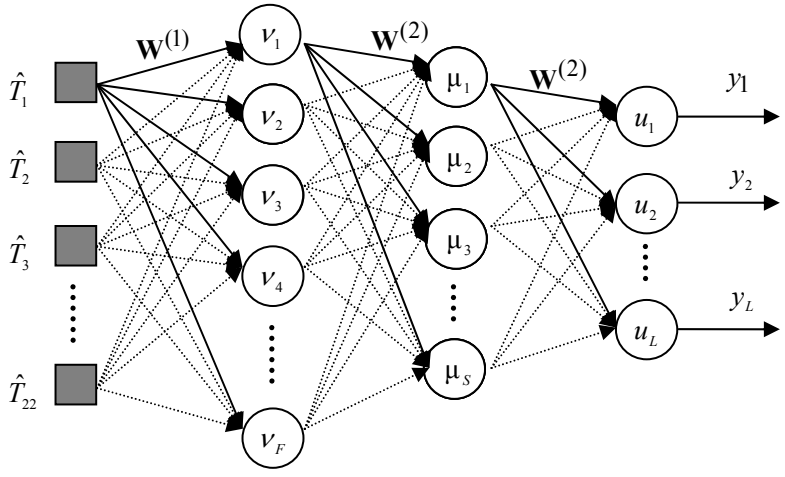


Рис. 1. Трехслойный персептрон

Элементами матрицы $\mathbf{W}^{(i)}$ являются весовые коэффициенты нейронов i-го слоя; ν_j, μ_k, u_t − суммарные сигналы, приходящие на вход нейронов в первом, втором и выходном слоях соответственно. Активационная функция нейронов имеет вид гиперболического тангенса.

На вход НС подается вектор $\mathbf{X}=(\hat{T}_1,\hat{T}_2,\hat{T}_3,...,\hat{T}_{22})$, где $\hat{T}_1,\hat{T}_2,\hat{T}_3,...,\hat{T}_{22}$ – масштабированные значения текстурных признаков, вычисленные для эталонных фрагментов изображения облачности и ПП. Для описания текстуры снимков используется статистический подход [2], в основе которого лежит вычисление матриц смежности яркости пикселей для нескольких угловых направлений. Энергия, энтропия, максимальная вероятность, однородность и контраст вычисляются по матрицам для четырех угловых направлений. Первый начальный момент и вариация рассчитываются по яркости пикселей эталонного фрагмента изображения.

В процессе обучения сети на ее вход последовательно предъявляются векторы признаков обучающих образцов. Вычисляются отклики нейронов всех слоев. Затем рассчитывается функция ошибки, которая для последовательного режима обучения равна

$$\mathbf{E}(\mathbf{W})=\frac{1}{2}\sum_{k=1}^{L}(y_k-d_k)^2 ,$$

где $\mathbf{y}=[y_1,y_2,....,y_L]^T$ – текущий отклик сети на входной сигнал, $\mathbf{d}=[d_1,d_2,....,d_L]^T$ – вектор ожидаемых выходных сигналов сети.

После предъявления каждого образца происходит подстройка весовых коэффициентов нейронов каждого слоя по формуле

$$\mathbf{W}(l+1)=\mathbf{W}(l)+\Delta\mathbf{W}(l),$$

где $\Delta\mathbf{W}(l)=\eta(l)\mathbf{p}(\mathbf{W})$ – величина коррекции весовых коэффициентов; $\eta(l)$ – коэффициент обучения на l-м шаге, а $\mathbf{p}(\mathbf{W})$ – направление поиска минимума $\mathbf{E}(\mathbf{W})$ в многомерном пространстве $\mathbf{W}$. В классическом алгоритме обратного распространения ошибки традиционно используется метод наискорейшего спуска, в котором направление поиска минимума определяется выражением

$$\mathbf{p}(\mathbf{W})=-\nabla\mathbf{E}(\mathbf{W}). \quad (1)$$

Метод наискорейшего спуска находит тот минимум, который расположен ближе всего к начальной точке. При этом найденный минимум может быть не глобальным, а локальным. Вторым серьезным недостатком метода наискорейшего спуска является его чувствительность к форме окрестности минимума. По сравнению с другими алгоритмами метод наискорейшего спуска требует меньших затрат памяти ЭВМ и довольно быстро достигает приемлемого уровня ошибки с заданной погрешностью, хотя к точному минимуму функции $\mathbf{E}(\mathbf{W})$ может сходиться довольно медленно.

Ниже будут рассмотрены три ГМ минимизации целевой функции ошибки, используемые для обучения НС: метод сопряженных градиентов, переменной метрики и Левенберга–Марквардта.

Градиентные методы минимизации

В методе **переменной метрики** направление поиска определяется выражением

$$\mathbf{p}(\mathbf{W})=-[\mathbf{H}(\mathbf{W})]^{-1}\nabla\mathbf{E}(\mathbf{W}),$$

где $\mathbf{H}(\mathbf{W})$ − матрица Гессе, которая должна быть положительно определенной на каждом шаге, что практически неосуществимо. Поэтому вместо точной матрицы $\mathbf{H}(\mathbf{W})$ используют ее приближение $\mathbf{G}(\mathbf{W})$ [3]. Если изменение $\mathbf{W}$ и $\nabla\mathbf{E}(\mathbf{W})$ на двух последовательных шагах итерации обозначить, как $\mathbf{s}$ и $\mathbf{r}$ соответственно, а матрицу, обратную приближению матрицы Гессе $\mathbf{V}=[\mathbf{G}(\mathbf{W})]^{-1}$, то в соответствии с формулой Бройдена−Флетчера−Гольдфабра−Шенно (BFGS) матрица $\mathbf{V}(l)$ на l-м шаге определяется выражением [4]

$$\mathbf{V}(l)=\mathbf{V}(l-1)-\frac{1}{\mathbf{s}^T\mathbf{V}(l-1)\mathbf{s}}\mathbf{V}(l-1)\mathbf{s}\mathbf{s}^T\mathbf{V}(l-1)+\frac{1}{\mathbf{r}^T\mathbf{s}}\mathbf{r}\mathbf{r}^T .$$

Данный метод характеризуется быстрой сходимостью, однако обладает рядом недостатков: менее устойчив, чем метод сопряженных градиентов, имеет тенденцию «застревать» в локальных минимумах и требует затрат памяти ЭВМ, пропорциональной квадрату числа весов в сети. Областью применения этого метода являются сети небольшого размера (несколько сотен связей).

В методе **сопряженных градиентов** выбор направления поиска определяется вектором

$$\mathbf{p}(\mathbf{W}(l))=-\nabla\mathbf{E}(\mathbf{W}(l))+\boldsymbol{\beta}(l-1)\mathbf{p}(\mathbf{W}(l-1)), \quad (2)$$

где $\boldsymbol{\beta}(l-1)$ − коэффициент сопряжения, который накапливает информацию о предыдущих направлениях поиска минимума целевой функции. Существуют несколько методов расчета коэффициента сопряжения: Полака−Рибьера, Флетчера−Ривса и три метода, основанных на взаимной ортогональности градиентов [4].

Метод сопряженных градиентов рекомендуется использовать для сетей с большим числом весов (более двух-трех сотен) и с несколькими выходными элементами [5]. Затраты памяти ЭВМ пропорциональны числу весов, а не их квадрату, а время обучения чуть больше или сравнимо с временными затратами метода BFGS.

В методе **Левенберга−Марквардта** приближенная матрица Гессе рассчитывается по формуле

$$\mathbf{G}(\mathbf{W}) = \left[\mathbf{J}(\mathbf{W})\right]^T \mathbf{J}(\mathbf{W}) + \mathbf{R}(\mathbf{W}),$$

где $\mathbf{J}(\mathbf{W})$ – матрица Якоби, а $\mathbf{R}(\mathbf{W})$ – компоненты $\mathbf{H}(\mathbf{W})$, содержащие высшие производные относительно $\mathbf{W}$. В основе подхода лежит аппроксимация $\mathbf{R}(\mathbf{W})$ с помощью регуляризационного фактора $\gamma \times \mathbf{I}$, где γ – скалярная величина (параметр Левенберга−Марквардта), изменяющаяся в процессе минимизации, $\mathbf{I}$ – единичная матрица. Это самый быстрый и надежный алгоритм минимизации. Однако его применение связано с определенными ограничениями: он используется только для сетей с одним выходным элементом, а также требует затрат памяти ЭВМ, пропорционально квадрату числа весов в сети. Фактически это ограничение не позволяет обучать сети большого размера (порядка тысячи и более весов) [3].

Методика и результаты анализа

Для проведения численного эксперимента по обучению НС было выбрано два метода: наискорейшего спуска и сопряженных градиентов. Использование метода переменной метрики является неэффективным, так как задача классификации облачности по типам предполагает построение НС, имеющей несколько тысяч связей. Метод Левенберга−Марквардта, согласно [3] невозможно использовать для обучения сети, предназначенной для классификации нескольких типов облачности и ПП.

Применение метода сопряженных градиентов требует пакетной обработки данных, при которой вычисляется усредненный градиент ошибки по всей обучающей выборке, и веса корректируются один раз в конце каждой эпохи. Под эпохой понимается предъявление на вход НС всей обучающей выборки. Метод наискорейшего спуска предполагает последовательную обработку данных, когда веса всех нейронов сети пересчитываются после предъявления каждого образца.

Начальное направление в методе сопряженных градиентов определяется вектором (1). Далее расчеты ведутся по формуле (2). Если ошибка возрастает, т.е. выполняется условие $\mathbf{E}(\mathbf{W}(l+1)) > \mathbf{E}(\mathbf{W}(l))$, то направление поиска выбирается в соответствии с (1) из последней найденной точки. Также смена направления $\mathbf{p}(\mathbf{W})$ производится через каждые Q (суммарное число весов сети) шагов, так как после этого исчерпываются возможности сопряжения [5].

Ключевыми факторами, влияющими на скорость и качество обучения НС, являются параметры методов минимизации: $\eta(l)$ – для метода наискорейшего спуска и $\boldsymbol{\beta}(l-1)$ – для сопряженных градиентов. Рассмотрены следующие способы задания $\eta(l)$:

– $\eta_f(l)$ подбирается эмпирически и остается фиксированным на протяжении всего периода обучения для всех слоев;

– в процессе обучения уменьшается от заданного начального значения $\eta_0(l)$ до $\eta_m(l)$ с фиксированным шагом $\Delta\eta(l)$ для всех слоев;

– подбирается эмпирически и остается фиксированным на протяжении всего периода обучения для каждого слоя $\eta_f^{(1)}(l), \eta_f^{(2)}(l), \eta_f^{(3)}(l)$ соответственно;

– в процессе обучения уменьшается от заданных значений $\eta_0^{(1)}(l), \eta_0^{(2)}(l), \eta_0^{(3)}(l)$ до $\eta_m^{(1)}(l), \eta_m^{(2)}(l), \eta_m^{(3)}(l)$ с фиксированным шагом $\Delta\eta^{(1)}(l), \Delta\eta^{(2)}(l), \Delta\eta^{(3)}(l)$ для каждого слоя соответственно;

– для каждого нейрона выбирается обратно пропорционально квадратному корню из суммы его синаптических весов.

В качестве примера рассмотрим трехслойную НС с числом нейронов в слоях $F = 22$, $S = 11$ и $L = 2$, которая обучалась двум типам текстур по 20 образцам для каждой. Сеть обучалась повторно 20 раз для каждого способа задания коэффициента $\eta(l)$. Оценки времени обучения приведены в табл. 1. Наиболее точные результаты классификации получаются при 1, 3 и 4-м способах задания коэффициента обучения. В дальнейшем был выбран 4-й способ, т.к. время обучения и количество эпох при его использовании было наименьшим.

Таблица 1

Время обучения и количество эпох при различных способах задания $\eta(l)$

№	Способ задания $\eta(l)$	Среднее время обучения, с	Среднее количество эпох
1	$\eta_f(l)=0{,}01$	430	11949
2	$\eta_0(l)=0{,}5$; $\eta_m(l)=0{,}01$; $\Delta\eta(l)=0{,}001$	13	218
3	$\eta_f^{(1)}(l)=0{,}03$; $\eta_f^{(2)}(l)=0{,}02$; $\eta_f^{(1)}(l)=0{,}01$	219	5867
4	$\eta_0^{(1)}(l)=0{,}3$; $\eta_m^{(1)}(l)=0{,}03$; $\Delta\eta^{(1)}(l)=0{,}001$; $\eta_0^{(2)}(l)=0{,}2$; $\eta_m^{(2)}(l)=0{,}02$; $\Delta\eta^{(2)}(l)=0{,}001$; $\eta_0^{(3)}(l)=0{,}1$; $\eta_m^{(3)}(l)=0{,}01$; $\Delta\eta^{(3)}(l)=0{,}001$	32	545
5	$\eta_f^{(1)}(l)=1/\sqrt{22}$; $\eta_f^{(2)}(l)=1/\sqrt{11}$; $\eta_f^{(3)}(l)=1/\sqrt{2}$	17	253

Способы задания коэффициента сопряжения, перечисленные выше, анализировались на примерах функций Вуда и Розенброка [4]. По результатам анализа были выбраны методы, для которых поиск минимума осуществлялся за наименьшее число итераций:

$$\boldsymbol{\beta}^1(l-1)=-\frac{\left[\nabla\mathbf{E}\left(\mathbf{W}(l)\right)-\nabla\mathbf{E}\left(\mathbf{W}(l-1)\right)\right]^T\nabla\mathbf{E}\left(\mathbf{W}(l)\right)}{\left[\left[\nabla\mathbf{E}\left(\mathbf{W}(l)\right)-\nabla\mathbf{E}\left(\mathbf{W}(l-1)\right)\right]\right]^T\mathbf{p}\left(\mathbf{W}(l-1)\right)},\quad \boldsymbol{\beta}^2(l-1)=\frac{\left[\nabla\mathbf{E}\left(\mathbf{W}(l)\right)-\nabla\mathbf{E}\left(\mathbf{W}(l-1)\right)\right]^T\nabla\mathbf{E}\left(\mathbf{W}(l)\right)}{\left\|\nabla\mathbf{E}\left(\mathbf{W}(l-1)\right)\right\|_2^2},$$

где $\|\bullet\|_2$ – евклидова норма.

Далее выбранные способы использовались для обучения сети из рассмотренного выше примера. Скорость обучения зависит от начального выбора направления поиска, которое задается случайно [5]. В табл. 2 приведены оценки скорости обучения. Коэффициент $\boldsymbol{\beta}^1$ был выбран, исходя из наименьшего среднего времени обучения сети.

Таблица 2

Сравнение времени обучения и количества эпох для разных коэффициентов сопряжения

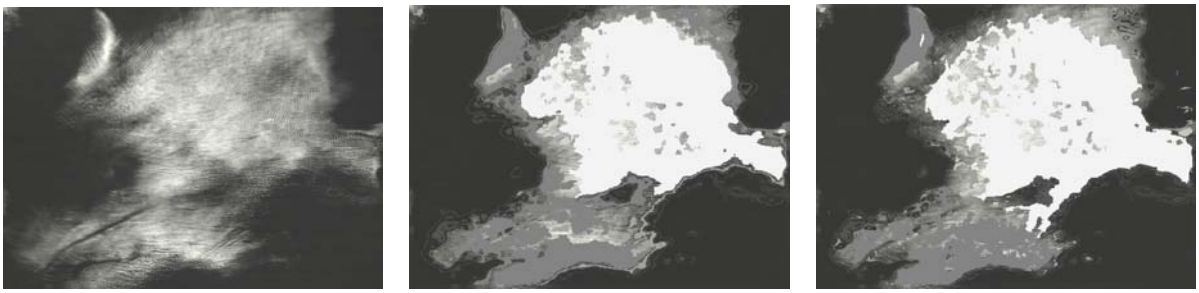
Коэффициент сопряжения	Среднее время обучения, с	Среднее количество эпох
$\boldsymbol{\beta}^1$	31	778
$\boldsymbol{\beta}^2$	63	1618

После выбора способов вычисления параметров $\eta(l)$ и $\boldsymbol{\beta}(l-1)$ проводилось сравнение методов наискорейшего спуска и сопряженных градиентов для сети с $F=22$, $S=11$ и $L=3$ или $L=4$. При этом сеть обучалась по 20 образцам для каждого типа текстуры. При $L=3$ анализировался снимок MODIS территории Аргентины с разрешением 250 м (рис. 2, *а*). Приведенные результаты классификации на рис. 2, *б* и *в* совпадают с данными наземных метеостанций. Светло-серым цветом выделена кучевая облачность, темно-серым – слоистая и черным – поверхность океана. Сходство и однородность выделенных областей облачности и ПП с незначительным числом разрывов свидетельствуют об эффективности исследуемых ГМ. Выбор наиболее эффективного метода осуществлялся исходя из минимального времени обучения НС, приведенного в табл. 3. При этом более эффективным оказался метод сопряженных градиентов, так как среднее количество эпох и время обучения было меньше, чем при использовании метода наискорейшего спуска.

Таблица 3

Время обучения и количество эпох для методов наискорейшего спуска и сопряженных градиентов

Метод	Среднее время обучения, с		Среднее количество эпох	
	$L=3$	$L=4$	$L=3$	$L=4$
Наискорейшего спуска	612	7428	10382	283009
Сопряженных градиентов	274	2295	7621	98937



а *б* *в*

Рис. 2. Результаты классификации текстур облачности и водной поверхности: исходное изображение (*а*), преобразованное методом наискорейшего спуска (*б*) и методом сопряженных градиентов (*в*)

Заключение

В ходе проведенного анализа были определены параметры методов $\eta(l)$ и $\boldsymbol{\beta}(l-1)$, при которых обучение занимает наименьшее время при лучших результатах классификации. Установлено, что время обучения НС двум классам практически совпадает для методов сопряженных градиентов и наискорейшего спуска. При обучении сети трем классам метод сопряженных градиентов более чем в 2 раза быстрее метода наискорейшего спуска, а при четырех классах – уже в 3 раза (см. табл. 3). Таким образом, можно сделать вывод о том, что метод сопряженных градиентов является наиболее эффективным при обучении НС для классификации облачности.

Авторы благодарны В.Г. Астафурову за полезные обсуждения и консультации.

Работа выполнена при частичной финансовой поддержке Минобрнауки РФ (госконтракт № 02.740.11.0674).

Литература

1. Rumelhart D.E. Learning representations of back-propagation errors / D.E. Rumelhart, G.E. Hinton, R.J. Williams // Nature. – 1986. – Vol. 323. – P. 533–536.

2. Харалик Р.М. Статистический и структурный подходы к описанию текстур // ТИИЭР. – 1979. − Т. 67, № 5. − С. 98–120.

3. Осовский С. Нейронные сети для обработки информации / пер. с польск. И.Д. Рудинского. − М.: Финансы и статистика, 2002. − 344 с.

4. Гилл Ф. Практическая оптимизация / Ф. Гилл, У. Мюррей, М. Райт. − М.: Мир, 1985. – 509 с.

5. Спуск по сопряженным градиентам [Электронный ресурс]. – Режим доступа: http://www.statsoft.ru/home/portal/applications/neuralnetworksadvisor/adv-new/ConjugateGradientDescent.htm, свободный (дата обращения: 24.07.2011).

Скороходов Алексей Викторович
Аспирант Института оптики атмосферы им. В.Е. Зуева СО РАН
Тел.: 8 (382-2) 49-22-56
Эл. почта: vazime@yandex.ru

Тунгусова Анна Владимировна
Студентка каф. автоматизированных систем управления ТУСУРа
Тел.: 8 (382-2) 49-22-56
Эл. почта: snusmumrik@t-sk.ru

Skorokhodov A.V., Tungusova A.V.

Comparative analysis of gradient minimization methods in the task of multilayer perceptron learning

The article presents the comparative analysis of gradient minimization methods of target error function in the task of neural network training. The description and characteristics of these methods are given. The results of the analysis and numerical experiments demonstrate that the conjugate gradient method is most efficient for neural network training.

Keywords: neural network, gradient methods.