УДК 004.91:004.8

Е. В. Бодянский[1], Н. В. Рябова[2] , О. В. Золотухин[3]
[1] ХНУРЭ, г. Харьков, Украина bodya@kture.kherkov.ua
[2] ХНУРЭ, г. Харьков, Украина ryabova.nv@gmail.com
[3] ХНУРЭ, г. Харьков, Украина zolotukhin.ov@gmail.com

КЛАССИФИКАЦИЯ ТЕКСТОВЫХ ДОКУМЕНТОВ С ПОМОЩЬЮ НЕЙРОННОЙ СЕТИ ВСТРЕЧНОГО РАСПРОСТРАНЕНИЯ С КОНТРОЛИРУЕМЫМ ОБУЧЕНИЕМ

Проведен анализ алгоритмов классификации политематических текстовых документов на основе нейронных сетей. Предложена архитектура нейронной сети встречного распространения для задач классификации текстовых документов, а также алгоритм контролируемого обучения этой сети.

КЛАССИФИКАЦИЯ, НЕЙРОННАЯ СЕТЬ, АЛГОРИТМ ОБУЧЕНИЯ, ФУНКЦИЯ АКТИВАЦИИ

Введение

Задачи интеллектуальной обработки текстовых документов в рамках Text Mining и Web Mining в настоящее время привлекают все большее внимание, а среди таких задач в качестве одной из наиболее актуальных можно отметить задачу классификации, которую удобно рассматривать с позиций теории распознавания образов. Для решения такого типа задач в качестве весьма эффективного аппарата в настоящее время широко используются искусственные нейронные сети, благодаря своим универсальным аппроксимирующим свойствам и способности к обучению [1, 2]. Так, в [3] задачу классификации текстовых документов было предложено решать с помощью нейронной сети обучаемого векторного квантования (LVQ), а в [4] для решения этой же задачи в условиях пересекающихся классов было предложено использовать нечеткий вариант LVQ. В [5] для решения задачи была использована вероятностная нейронная сеть (PNN), а в [6-8] были введены различные модификации PNN, включая нечеткую. В [9] для обработки текстов была предложена иерархическая нейронная радиально-базисная сеть (RBFN), обучаемая на основе персептронного критерия.

И хотя с помощью этих нейронных сетей были получены вполне приемлемые результаты, все они не лишены некоторых недостатков, ограничивающих их применимость. Так, алгоритмы обучаемого векторного квантования, являясь по сути процедурами стохастической аппроксимации [10], характеризуются низкой скоростью сходимости, что требует больших объемов обучающих выборок. Вероятностные и радиально-базисные нейронные сети подвержены «проклятию размерности» в случае высокой размерности входных векторов-образов, что существенно затрудняет их использование в режиме реального времени.

Таким образом, представляется целесообразным разработка быстродействующей нейронной сети, предназначенной для решения задач классификации текстовых документов в режиме реального времени при последовательном поступлении данных на вход системы.

1. Архитектура нейронной сети встречного распространения для классификации текстовых документов

Гетерогенные нейронные сети встречного распространения были предложены Р. Хехт-Нильсеном [11-13] в качестве альтернативы классическим многослойным сетям с прямой передачей информации и сокращают время обучения, как минимум, на порядок по сравнению с многослойными персептронами, хотя несколько проигрывают им по точности аппроксимации. Исходно эти сети были разработаны для аппроксимации по экспериментальным данным некоторого отображения $y = F(x)$ и нахождения обратного оператора $x = F^{-1}(y)$, хотя впоследствии с различными модификациями, включая нечеткий вариант [14-16], применялись в основном в задаче распознавания образов (классификации), компрессии данных, ассоциативной памяти.

Наиболее известная сеть встречного распространения (CPN) [2, 15] представляет собой гибрид самоорганизующейся карты Т. Кохонена (первый скрытый слой) и набора звезд С. Гроссберга (выходной слой) и соответственно сочетает в себе конкурентное самообучение с контролируемым обучением с учителем. При этом можно отметить, что узлами этой сети, как правило, являются адаптивные линейные ассоциаторы, чей вход линейно зависит от настраиваемых синаптических весов, что определяет высокую скорость их настройки.

Специфика задачи классификации текстовых документов требует существенной модификации как архитектуры CPN, так и алгоритмов ее обучения. Во-первых, поскольку в режиме обучения на вход сети подаются классифицированные образы, целесообразно в первом скрытом слое использовать не традиционную самоорганизующуюся карту (SOM), обучаемую без учителя, а нейронную сеть векторного квантования LVQ [17, 18], обучаемую с учителем, что позволяет повысить быстродействие, а, кроме того, если SOM может оценить только центроиды классов, то LVQ – и центроиды, и границы этих классов.

Во-вторых, в выходном слое вместо звезд Гроссберга целесообразно использовать элементарные персептроны Розенблатта с нелинейной функцией активации (релейной), принимающей только два значения: 1, если предъявляемый образ относится к данному конкретному классу, и 0 – в противном случае.

Архитектура предлагаемой сети встречного распространения приведена на рис. 1.

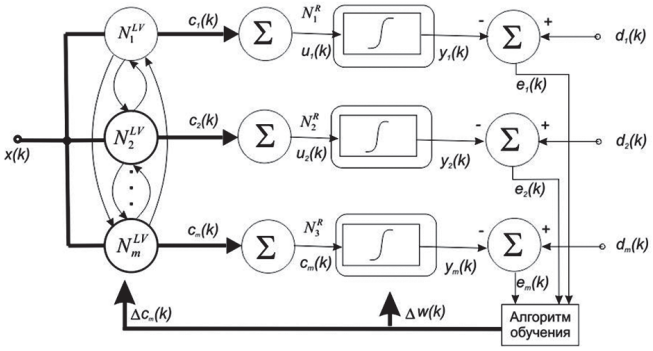

Рис. 1. Модифицированная нейронная сеть встречного распространения

Исходной информацией для обучения является последовательность векторов-образов

$$x(1), x(2),..., x(k), x(k) = (x_1(k), x_2(k),..., x_n(k))^T \in R$$

с известной классификацией, которая в реальном времени, что характерно для задач Web Mining, подается на нулевой (рецепторый) слой сети. Нейроны первого скрытого слоя N_j^{LV} $(j = 1,2,...,m)$, m – количество возможных классов, задаваемые априорно, являясь по сути адаптивными линейными ассоциаторами [2], предназначены для нахождения центроидов и границ классов, при этом $(n \times 1)$ -векторы, описывающие эти центроиды $c_j(k) = (c_{j1}(k), c_{j2}(k),..., c_{jn}(k))^T$, являются синаптическими настраиваемыми весами каждого из нейронов N_j^{LV}. При этом все входные сигналы перед подачей на нулевой слой нормируются так, что $\|x(k)\| = 1$.

Выходной слой сети образован m-элементарными персептронами Розенблатта [2] N_j^R с сигмоидальной функцией активации, при этом

$$y_j(k) = \psi(\gamma u_j(k)) = \psi(\sum_{i=1}^{n} \gamma w_{ji} c_{ji}(k)) = \\ = \psi(\gamma w_j^T c_j(k)) = \frac{1}{1 + e^{-\gamma u_j(k)}} = \frac{1}{1 + e^{-\gamma w_j^T c_j(k)}}, \quad (1)$$

где γ – параметр крутизны активационной функции, $w_j = (w_{j1}, w_{j2},..., w_{jn})^T$ – вектор синаптических весов N_j^R.

На рис. 2 показан вид сигмоидальной активационной функции $\psi(\gamma u_j)$ в зависимости от параметра крутизны γ. При этом, чем больше значение γ, тем ближе $\psi(\gamma u_j)$ к релейной функции

$$\psi(\gamma u_j) = \begin{cases} 1, \text{ при } u_j \geq 0, \\ 0, \text{ при } u_j < 0, \end{cases} \quad (2)$$

обычно используемой в задачах распознавания образов.

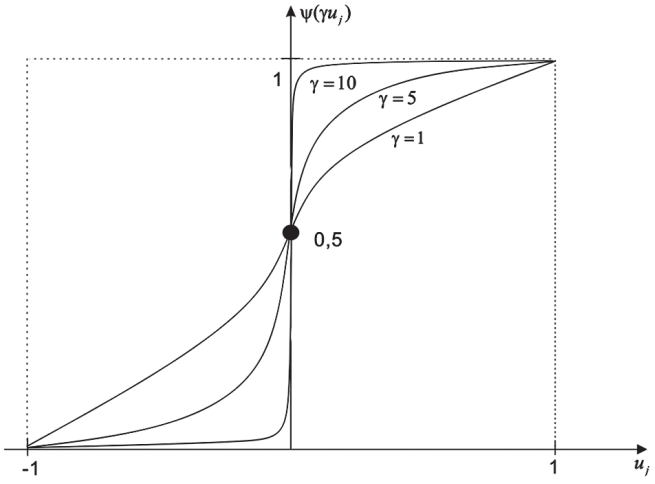

Рис. 2. Зависимость сигмоидальной функции от параметра крутизны

Понятно, что при $\gamma \to \infty$, $\psi(\gamma u_j)$ совпадает с (2), не претерпевая при этом разрыва производной в точке $u_j = 0$.

Здесь интересно также заметить, что векторы $c_j(k)$, являясь синаптическими весами N_j^{LV}, подаются в качестве входных сигналов на выходные нейроны N_j^R.

Выходные сигналы сети $y_j(k)$ принимают два значения

$$y_j(k) \approx \begin{cases} 1, \text{если } x(k) \text{ относится к } j\text{-ому классу}, \\ 0, \text{в противном случае}, \end{cases}$$

при этом точные значения 1 или 0 никогда не достигаются.

2. Обучение нейронной сети встречного распространения для классификации текстовых документов

При подаче на вход нейронной сети вектора-образа $E(k)$ $(\|x(k)\| = 1)$ в процессе конкуренции, реализуемой по латеральным (поперечным) связям первого скрытого слоя между нейронами N_j^{LV}, определяется нейрон-победитель j^*, вектор синаптических весов которого $c_{j^*}(k-1)$ в смысле принятой метрики (как правило, евклидовой) наиболее близок ко входному сигналу:

$$j^* = \arg\min_j D(x(k), c_j(k-1)) = \\ = \arg\min_j \|x(k) - c_j(k-1)\|^2 = \arg\max_j c_j^T(k-1)x(k) = \\ = \arg\max_j \cos(c_j(k-1), x(k)),$$

при этом $-1 \leq \cos(c_j(k-1), x(k)) = c_j^T(k-1)x(k) \leq 1$, а $0 \leq \|x(k) - c_j(k-1)\|^2 \leq 4$.

Поскольку обучение в первом скрытом слое является контролируемым (в отличие от традиционной CPN), принадлежность каждого вектора $x(k)$ к конкретному классу известна, что позволяет рассмотреть две возможные ситуации, возникающие в обучаемом векторном квантовании:

– входной вектор $x(k)$ и нейрон-победитель $N_{j^*}^{LV}$ принадлежат одному классу;

– входной вектор $x(k)$ и нейрон-победитель $N_{j^*}^{LV}$ принадлежат разным классам.

Тогда LVQ-правило обучения может быть записано в виде:

$$c_j(k)=\begin{cases}\dfrac{c_{j^*}(k-1)+\eta(k)(x(k)-c_{j^*}(k-1))}{\left\|c_{j^*}(k-1)+\eta(k)(x(k)-c_{j^*}(k-1))\right\|},\\ \text{если } x(k) \text{ и } c_{j^*}(k-1)\in \text{одному классу},\\ \dfrac{c_{j^*}(k-1)-\eta(k)(x(k)-c_{j^*}(k-1))}{\left\|c_{j^*}(k-1)-\eta(k)(x(k)-c_{j^*}(k-1))\right\|},\\ \text{если } x(k) \text{ и } c_{j^*}(k-1)\in \text{разным классам},\\ A_j(k-1), \text{если } j\text{-ый нейрон не победил},\end{cases} \tag{3}$$

где $0<\eta(k)\le 1$ – параметр шага обучения.

Правило контролируемого обучения (3) имеет ясный физический смысл: если нейрон-победитель и предъявляемый образ относится к одному классу, то центроид $c_{j^*}(k-1)$ притягивается к $x(k)$; если же $c_{j^*}(k-1)$ и $x(k)$ относятся к разным классам, то $c_{j^*}(k-1)$ отталкивается от $x(k)$, минимизируя или максимизируя расстояние

$$D(x(k),c_{j^*}(k-1))=\left\|x(k)-c_{j^*}(k-1)\right\|^2,$$

при этом автоматически производится нормирование $\left\|c_j(k)\right\|=1$.

Что касается выбора шага обучения $\eta(k)$, то общая рекомендация сводится к тому, что он должен монотонно уменьшаться в процессе настройки согласно правилам стохастической аппроксимации. Выбор $\eta(k)=k^{-1}$ соответствует рекуррентной оценке среднего арифметического с помощью первого соотношения системы (3). При $\eta(k)=1$ вместо оценок $\mathrm{c}_j(k)$ на выходной слой классификации подаются сами входные образы $x(k)$. В [4] для обучения LVQ была использована рекуррентная оценка $\eta(k)$, обеспечивающая процессу обучения дополнительные фильтрующие свойства.

Для обучения синаптических весов выходного слоя, образованного элементарными персептронами Розенблатта, будем использовать стандартный квадратичный критерий вида

$$E_j(k)=\frac{1}{2}c_j^2(k)=\frac{1}{2}(d_j(k)-y_j(k))^2=$$

$$=\frac{1}{2}(d_j(k)-\psi(\gamma u_j(k))^2=\frac{1}{2}(d_j(k)-\psi(\sum_{i=1}^{n}\gamma w_{ji}c_{ji}(k)))^2,$$

где $d_j(k)=\begin{cases}1, \text{если } x(k) \text{ относится к } j\text{-ому классу},\\ 0, \text{в противном случае}.\end{cases}$ (4)

Здесь следует отметить, что хотя в задаче распознавания образов обычно используется так называемый персептронный критерий [19], применение квадратичного критерия (4) с большим значением γ позволяет оптимизировать процесс обучения по быстродействию.

Минимизация критерия обучения (4) может быть обеспечена с помощью рекуррентной процедуры (δ-правила обучения [2]) вида:

$$\begin{aligned}w_{ji}(k)&=w_{ji}(k-1)-\eta^R(k)\frac{\partial E_j(k)}{\partial e_j(k)}\bullet\frac{\partial e_j(k)}{\partial w_{ji}}=\\ &=w_{ji}(k-1)-\eta^R(k)e_j(k)\frac{\partial e_j(k)}{\partial w_{ji}}=\\ &=w_{ji}(k)-\eta^R(k)\mathrm{e}_j(k)\frac{\partial e_j(k)}{\partial u_j(k)}\bullet\frac{\partial u_j(k)}{\partial w_{ji}}=\\ &=w_{ji}(k)+\eta^R(k)e_j(k)\psi'(\gamma u_j(k))c_{ji}(k)=\\ &=w_{ji}(k)+\eta^R(k)\delta_j(k)\mathrm{c}_{ji}(k),\end{aligned} \tag{5}$$

где $\delta_j(k)=e_j(k)\psi'(\gamma u_j(k))=-\dfrac{\partial E_j(k)}{\partial u_j}$ – локальная ошибка.

В векторной форме алгоритм (5) можно представить как:

$$w_j(k)=w_j(k-1)+\eta^R(k)\delta_j(k)\mathrm{c}_j(k), \tag{6}$$

а с учетом (1):

$$\begin{aligned}w_j(k)&=w_j(k-1)+\eta^R(k)\gamma e_j(k)y_j(k)(1-y_j(k))c_j(k)=\\ &=w_j(k-1)+\eta^R(k)\gamma(d_j(k)-\\ &-w_j^T(k-1)c_j(k))y_j(k)(1-y_j(k))c_j(k)=\\ &=w_j(k-1)+\eta^R(k)e_j(k)J_j(k),\end{aligned} \tag{7}$$

где $J_j(k)=\gamma y_j(k)(1-y_j(k))\mathrm{c}_j(k)$.

Повысить быстродействие процесса обучения выходного слоя можно, переходя от градиентных процедур типа (7) к псевдоньютоновским алгоритмам, среди которых можно отметить популярный в теории и практике нейронных сетей алгоритм Левенберга-Марквардта [2]. Вводя одношаговую модификацию этого алгоритма

$$\begin{aligned}w_j(k)&=w_j(k-1)+\\ &+(J_j(k)J_J^T(k)+\eta^R I)^{-1}J_j(k)(d_j(k)-\\ &-\psi(\gamma w_j^T(k-1)\mathrm{c}_j(k)))\end{aligned} \tag{8}$$

(здесь $\eta^R>0$ – параметр регуляризации, $I-(n\times n)$ – единичная матрица) и используя формулу Шермана-Моррисона обращения матриц, можно переписать (8) в простом виде [20]

$$\begin{aligned}w_j(k)&=w_j(k-1)+\\ &+\frac{d_j(k)-\psi(\gamma w_j^T(k-1)c_j(k))}{\eta^R+\left\|J_j(k)\right\|^2}J_j(k),\end{aligned} \tag{9}$$

являющимся распространением на нелинейной случай широко используемого оптимального алгоритма обучения нейронных сетей Уидроу-Хоффа [2] и аддитивного алгоритма Качмажа [21], принятого в задачах адаптивной идентификации.

Можно заметить, что алгоритмы обучения CPN (3) и (9) отличаются вычислительной простотой и высоким быстродействием, что позволит использовать их при обработке информации в реальном времени.

Выводы

Рассмотрена задача автоматической классификации текстовых документов, поступающих на обработку в реальном времени. Предложена архитектура специализированной нейронной сети и алгоритм ее контролируемого обучения, являющиеся расширением сетей встречного распространения для рассматриваемой задачи. Предложенная нейронная сеть может найти применение для задач Text- и Web-Mining, распознавания образов, классификации и т.п., а алгоритмы ее обучения характеризуются вычислительной простотой и высоким быстродействием, являясь обобщением оптимальных линейных правил.

Список литературы: 1. *Bilshop C.M.* Neural Networks for Pattern Recognition. – Oxford: Clarendon Press, 1995. – 482 p. **2.** *Haykin S.* Neural Networks. A Comprehensive Foundation – Upper Saddle River, N.J.: Prentice Hall, Inc., 1999. – 842 p. **3.** *Umer M.F.*, Khiyal M.S.H. Classification of textual documents using learning vector quantization// Information Technology Journal. – 2007. – 6(1). – P. 154-159. **4.** *Бодянский Е.В.,* Рябова Н.В., Золотухин О.В. Обработка текстовых документов с помощью адаптивного нечеткого обучаемого векторного квантования // Вісник Національного технічного університету «ХПІ». – 2011. – № 53. – С.109-115. **5.** *Ciarelli P.M.*, Oliveira E. An enhanced probabilistic neural network approach applied to text classification// Lecture Notes on Computer Science. – V. 5856. – Berlin – Heidelberg: Springer-Verlag, 2009. – P. 661-668. **6.** *Бодянский Е.В.,* Рябова Н.В., Золотухин О.В. Классификация текстовых документов с помощью нечеткой вероятностной нейронной сети // Восточно-европейский журнал передовых технологий – 2011. – №6/2 (54). – С. 16-18. **7.** *Bodyanskiy Ye.,* Shubkina O. Semantic annotation of text documents using evolving neural network based on principle "Neurons at Data Points"// Proc. 4th Int. Workshop on Inductive Modelling "IWIM 2011". – Kyiv, 2011. – P. 31-37. **8.** *Bodyanskiy Ye.*, Shubkina O. Semantic annotation of text documents using modified probabilistic neural network// Proc. 6th IEEE Int. Conf. Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications. – 15-17 Sept.2011, Prague, Czech Republic, 2011. – P. 328-331. **9.** *Бодянский Е.В.*, Шубкина О.В. Семантическое аннотирование текстовых документов на основе иерархической радиально-базисной нейронной сети // Восточно-европейский журнал передовых технологий – 2010. – №6/3 (48). – С. 72-77. **10.** *Kosko B.* Neural Networks for Signal Processing. – Englewood Cliffs, N.J.: Prentice Hall, 1992. – 398 p. 11. Hecht-Nielsen R. Counterpropagation networks // Applied Optics. – 1987. – 26. – P. 4989-4984. **12.** Hecht-Nielsen R. Application counterpropagation networks // Neural Networks. – 1988. – 1. – P. 131-139. **13.** Hecht-Nielsen R. Counterpropagation networks // Proc. Int. Conf. on Neural Networks.– San Diego, CA, 1990. – V. 3. – P. 17-20. **14.** *Бодянский Е.В.*, Руденко О.Г. Искусственные нейронные сети: архитектуры, обучение, применение. – Харьков: ТЕЛЕТЕХ, 2004. – 372 с. **15.** *Alavala C.R.* Fuzzy Logic and Neural Networks: Basic Concepts and Applications. – New Delhi: New Age Int. Ltd, 2008. – 276 p. **16.** *Bodyanskiy Ye.*, Gorshkov Y., Otto P., Pliss I. Medical image analysis using neuro-fuzzy network // Proc. 54. Int. Wiss. Kolloquium IWK-2009. – Ilmenau: TU Ilmenau, 2009. – P. 55-62. **17.** *Kohonen T.* Improved version of learning vector quantization// Proc. Int. Joint Conf. on Neural Networks.- San Diego, CA, 1990.-v.1.-P.545-550. **18.** *Kohonen T.* Self-Organizing Maps.-Berlin: Springer, 1995.-362 p. **19.** *Shynk J.J.* Perfomance surfaces of a single-layer perceptron // IEEE Trans. on Neural Networks. – 1990. – 1. – P.268-274. **20.** *Бодянський Є.В.*, Михальов О.І., Плісс І.П. Адаптивне виявлення розглядань в об'єктах керування за допомогою штучних нейронних мереж. – Дніпропетровськ: Системні технології, 2000. – 140 с. **21.** *Райбман Н.С.*, Чадеев В.М. Построение моделей процессов производства. – М.: Энергия, 1975. – 376 с.

Поступила в редколлегию 15.01.2014

УДК 004.91:004.8

Класифікація текстових документів за допомогою нейронної мережі зустрічного поширення з контрольованим навчанням / Є. В. Бодянський, Н. В. Рябова, О. В. Золотухін // Біоніка інтелекту: наук.-техн. журнал. – 2014. – № 1 (82). – С. 3–6.

У статті розглядаються актуальні методи класифікації політематичних текстів. Стверджується, що алгоритми, засновані на нечітких нейронних мережах, користуються значною популярністю, тому що забезпечують високу точність результатів.

Іл. 2. Бібліогр.: 21 найм.

UDC 004.91:004.8

Classification of text documents using counter propagation neural network with controlled training / Ye.V. Bodyanskiy, N.V. Ryabova, O.V. Zolotukhin // Bionics of Intelligence: Sci. Mag. – 2014. – № 1 (82). – P. 3-6.

In this article discusses the current classification methods polythematic texts. it is alleged that the algorithms based on fuzzy neural networks are highly popular because provide highly accurate results.

Fig. 2. Ref.: 21 items.