**Список литературы:** **1.***Свами М.* Графы, сети и алгоритмы / М. Свами, К. Тхуласираман. – М.: Мир, 1984.– 455с. **2**. *Зайцев Д.А.* Синтез моделей Петри телекоммуникационных протоколов / Д.А.Зайцев // Наукові праці ОНАЗ ім. О.С.Попова. – 2005. – №2. – С.36-42. **3**.*Клейнрок Л.* Вычислительные системы с очередями / Клейнрок Л. – М.: Мир, 1979. –600 с. **4.**Фрактальный анализ и процессы в компьютерных сетях : учеб. пособие / Ю.Ю. Громов, Н.А. Земской, О.Г. Иванова, А.В. Лагутин, В.М. Тютюнник. – 2-е изд. Тамбов : Изд-во Тамб. гос. техн. ун-та, 2007. – 108 с.*5.*A Brief Introduction to Neural Networks. David Kriesel. – Режим доступа:http://www.dkriesel.com/en/science/neural_networks **6**.*Петров А.Е.* Тензорный метод Крона, LT метод Бартини-Кузнецова и двойственные сети / А.Е.Петров // Электронное научное издание «Устойчивое инновационное развитие: проектирование и управление». – 2010. – том 6, №4 (9), ст. 2. – С. 13-32. **7**.Лемешко О.В. Теоретичні основи управління мережними ресурсами з використанням тензорних математичних моделей телекомунікаційних систем: автореф. дис. на здобуття наук. ступ. докт. техн.наук / Лемешко Олександр Віталійович. – Харків, 2005. – 37 с. **8**.*Тихонов В.И.* Построение тензорной модели асимметричных цифровых потоков в комплексном пространстве [Электронный ресурс] // Проблеми телекомунікацій. – 2011. – № 2 (4). – С. 42 – 53. – Режим доступа к журн.: http://pt.journal.kh.ua/2011/2/1/112_tikhonov_tensor.pdf **9**.*Тихонов В.И.* Фрактальная топологическая модель открытой телекоммуникационной сети / В.И.Тихонов // Наукові праці ОНАЗ ім. О.С.Попова. – 2010. – №1. – С.49-58. **10**.*Kuratowski K.* Set theory / K. Kuratowski, A. Mostowski. – Amsterdam: North-Holland Publishing Company,1976. – 416 p. **11**.*Дж.Келли.* Общая топология / *Дж.Келли*. – М.:Наука, 1968. – 384 с.

*Поступила в редколлегию 23.11.2011*

**УДК 004.91:004.8**

***Е.В. БОДЯНСКИЙ*****,** докт.техн.наук, проф., ХНУРЭ, Харьков
***Н.В. РЯБОВА*****,** канд.техн.наук, зав. каф., ХНУРЭ, Харьков
***О.В. ЗОЛОТУХИН*****,** асп., ХНУРЭ, Харьков

# ОБРАБОТКА ТЕКСТОВЫХ ДОКУМЕНТОВ С ПОМОЩЬЮ АДАПТИВНОГО НЕЧЕТКОГО ОБУЧАЕМОГО ВЕКТОРНОГО КВАНТОВАНИЯ

Рассматривается проблема интеллектуальной обработки текстов. Представлена архитектура нейро-фаззи системы для классификации текстовых документов и on-line алгоритм обучения сети адаптивного нечеткого векторного квантования.
**Ключевые слова**: текстовый документ, нечеткая классификация, AFLVQ

Розглядається проблема інтелектуальної обробки текстів. Представлена архітектура нейронечіткої системи для класифікації текстових документів та on-line алгоритм навчання мережі адаптивного нечіткого векторного квантування.
**Ключові слова**: текстовий документ, нечітка класифікація, AFLVQ

This article discusses a problem of an intelligent text processing. Architecture of the neuro-fuzzy system is presented for classification of text documents and on-line learning algorithm for fuzzy network adaptive vector quantization.
**Keywords:** text document, fuzzy classification, AFLVQ

## 1.Введение

Обработка текстовых документов, по сути, включает в себя комплекс взаимосвязанных задач, направленных на представление текстов в виде, пригодном для их использования компьютерными программами. Одной из

важных задач в этом комплексе является классификация, т.е. отнесение текстовых документов к заранее определенным классам. На сегодняшний день классификация текстов считается достаточно сложной проблемой, как в научном, так и в прикладном аспектах. Важность классификации набора взаимосвязанных текстовых документов неуклонно приобретает все большее значение, учитывая тот фактор, что большая часть информации в Интернет-пространстве представлена в текстовом виде. При этом Web-документы, подлежащие обработке, зачастую характеризуются разнородностью, широким охватом сразу нескольких тем, т.е. политематичностью. При наличии политематических текстов и большого количества классов задача становится значительно сложнее. Заметим также, что современные текстовые базы данных являются политематическими, с большим количеством категорий, что значительно усложняет задачу классификации.

Учитывая постоянно возрастающие объемы доступной информации в текстовом виде и связанную с этим проблему смыслового поиска, актуальность разработки методов и моделей автоматической классификации текстовых документов различного типа чрезвычайно высока.

В данной работе представлена нейронная сеть адаптивного нечеткого обучаемого векторного квантования для решения задачи классификации политематических текстовых документов. Проведены теоретические исследования для сравнения предлагаемого нами классификатора и других, хорошо известных из литературы, которые были специально разработаны для решения такого рода проблем. В связи с полученными результатами показано, что предлагаемый подход превосходит другие классификаторы и является более быстрым, чем вероятностные нейронные сети.

Классификация текстовых документов рассматривается как один из возможных вариантов решения проблемы использования информационных ресурсов. Коротко она характеризуется следующим образом. К настоящему моменту различными хранилищами знаний накоплены огромные информационные массивы. Однако, отсутствие возможности получить наиболее актуальную и полную информацию по конкретной теме делает бесполезной большую часть накопленных ресурсов. Поскольку исследование конкретной задачи требует все больших трудозатрат на непосредственный поиск и анализ информации по теме, многие решения принимаются на основе неполного представления о проблеме.

Использование классификаторов позволяет сократить трудозатраты на поиск нужной информации, представленной электронными текстами, а использование искусственных нейронных сетей упрощает процедуру построения классификатора.

**1. Интеллектуальная обработка текстов.**

Большинство существующих систем, работающих с документами, представляют собой электронный вариант архива документов со стандартным набором технических средств: удобный ввод-вывод, поддержка большого количества форматов, надёжное хранение, система ограничений прав доступа и т.п. Однако часто это оказывается недостаточным для современных

информационных систем: во-первых, постоянный рост потока документов приводит к тому, что многие документы не доходят до фактического адресата, т.е. до тех, кто заинтересован в получении данного документа; во-вторых, в постоянно разрастающемся архиве становится трудно (практически невозможно) найти нужные документы. Актуальность задачи интеллектуальной обработки документов, в частности, состоит в преодолении этих проблем. Современные классификаторы документов должны решать задачу, связанную с управлением потоком входящих документов в режиме реального времени, – их автоматическую классификацию и последующий поиск в нем документов по содержанию.

С позиций Text Mining – это задача классификации, когда каждый документ может быть отнесен к одному из априори заданных классов, при этом предполагается, что априори задана обучающая выборка с известной классификацией, на основании которой формируются границы между этими классами. Классические методы распознавания образов в этой задаче малоэффективны, поскольку их использование связано с гипотезой компактности и линейной разделимости классов. Для построения нелинейной разделяющей гиперповерхности между разными классами текстовых документов с успехом могут быть использованы искусственные нейронные сети (ИНС) [1-4], при этом предпочтение, естественно, отдается ИНС, обучение которых может производиться в on-line режиме, когда тексты на обработку поступают последовательно одним за одним. Задача существенно усложняется, когда один и тот же документ с различными уровнями принадлежности может одновременно относиться сразу к нескольким классам. В данной ситуации наиболее эффективными представляются методы нечеткой (фаззи) классификации [5], предназначенные для обработки данных, однозначная классификация которых в принципе невозможна.

В настоящей работе предлагается архитектура и on-line алгоритм обучения нейро-фаззи системы, предназначенные для последовательной обработки текстовых документов в условиях перекрывающихся классов.

## 2. Адаптивная нечеткая нейронная сеть обучаемого векторного квантования

В основу предлагаемой системы положена искусственная нейронная сеть обучаемого векторного квантования (LVQ) [6,7], имеющая крайне простую однослойную архитектуру, настройка семантических весов которой производится в режиме обучения с учителем с элементами конкуренции по типу «победитель получает все» (WTA). Основными преимуществами этой ИНС по сравнению с другими нейросистемами является простота архитектуры, незначительное количество входящих в нее нейронов, малый объем обучающей выборки и возможность on-line обучения [1], что крайне важно в задачах обработки текстовых документов. К настоящему времени известно множество вариантов LVQ-нейросетей [8-13], отличающихся выбором параметра шага обучения, используемой метрикой, необходимым объемом обучающей выборки. Эти системы подтвердили свою эффективность во многих приложениях, связанных с четкой классификацией и распознаванием образов.

Для решения задач нечеткой классификации в условиях пересекающихся классов был введен целый ряд модификаций LVQ-систем. Так, в [5] было введено нечеткое обучаемое векторное квантование FLVQ, представляющее собой по сути гибрид метода нечетких С-средних (FCM) и LVQ-сети и предназначенное для работы только в пакетном режиме. В [14] были предложены нечеткие алгоритмы обучаемого векторного квантования (FALVQ), в которых с каждым вектором-прототипом класса связывается та или иная функция принадлежности, определяющая подобие каждого прототипа с предъявляемым вектором-образом. Можно отметить громоздкость этого подхода и субъективизм при выборе конкретной функции принадлежности. В [15] введено нечеткое мягкое векторное квантование (FSLVQ), основанное на использовании мягкой конкуренции, ядерных функций соседства-принадлежности и, опять-таки, пакетной обработки данных. Весьма перспективным представляется подход, предложенный в [16] и представляющий собой гибрид нейронных сетей адаптивного резонанса (ART) и обучаемого векторного квантования. Данная система предназначена для работы в on-line режиме, однако весьма громоздка с вычислительной точки зрения.

Архитектура предлагаемой нами нейро-фаззи системы адаптивного обучаемого векторного квантования (AFLVQ) приведена на рис.

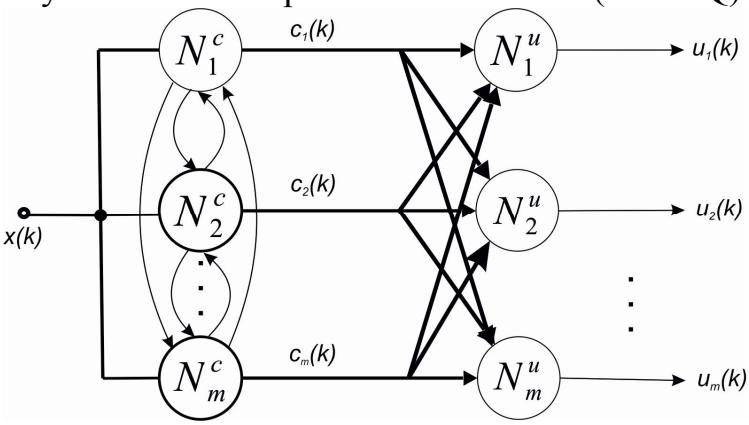


Рис. Нейронная сеть адаптивного нечеткого обучаемого векторного квантования (AFLVQ)

Система содержит два слоя обработки информации, при этом нейроны первого скрытого слоя связаны между собой латеральными связями, с помощью которых реализуются процессы конкуренции.

Исходной информацией для обучения является последовательность векторов-образов $x(1), x(2), \ldots, x(k), \ldots, x(N), \ldots;$ где $x(k) = (x_1(k), x_2(k), \ldots, x_n(k))^T \in R^n$ с известной классификацией, при этом входные сигналы предварительно нормируются так, что $\|x(k)\| = 1$. Нейроны первого скрытого слоя $N_j^c$ $(j = 1,2,\ldots,m; m$ - априори задаваемое количество возможных классов) предназначены для нахождения прототипов (центроидов) классов $c_j(k) = (c_{j1}(k), c_{j2}(k), \ldots c_{ji}(k) \ldots c_{jn}(k))^T$, при этом компоненты $c_{ji}(k)$ являются по сути настраиваемыми синаптическими весами нейрона $N_j^c$. Нейроны выходного слоя $N_j^u$ вычисляют уровни принадлежности $u_j(k)$ предъявленного образа $x(k)$ к $j-$ому классу.

Итак, при подаче на вход системы образа $x(k)$ в процессе конкуренции определяется нейрон-победитель $j^*$, синаптические веса которого $c_{j^*}(k-1)$ в смысле принятой метрики (в нашем случае евклидовой) наиболее близки к входному сигналу, т.е.

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

при этом очевидно, что

$$-1 \le \cos\left(x(k), c_j(k-1)\right) = x^T(k) c_j(k-1) \le 1, \qquad (1)$$

а

$$0 \le \left\|x(k) - c_j(k-1)\right\|^2 \le 4.$$

Поскольку обучение является контролируемым, то принадлежность 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 $N_{j*}^c$ принадлежат одному классу;

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

Тогда стандартное LVQ-правило обучения может быть записано в виде

$$c_j(k) = \begin{cases} c_{j*}(k-1) + \eta(k)\left(x(k) - c_{j*}(k-1)\right) - \text{если } x(k) \text{ и } c_{j*}(k-1) \text{ принадлежат одному классу}, \\ c_{j*}(k-1) - \eta(k)\left(x(k) - c_{j*}(k-1)\right) - \text{если } x(k) \text{ и } c_{j*}(k-1) \text{ принадлежат разным классам}, \\ c_j(k-1) - \text{если } j-\text{ый нейрон не победил.} \end{cases} \qquad (2)$$

Правило обучения (2) имеет ясный физический смысл: если нейрон-победитель и предъявленный образ относятся к одному классу, то прототип $c_{j*}(k-1)$ отталкивается от $x(k)$, увеличивая тем самым расстояние $D\left(x(k)\right), c_{j*}(k-1))$.

Что касается выбора величины шага обучения $\eta(k)$, то общая рекомендация сводится к тому, что он должен монотонно уменьшаться в процессе настройки. В [17] была доказана асимптотическая сходимость процесса обучаемого векторного квантования в предположении, что параметр $\eta(k)$ изменяется в соответствии с условиями стохастической аппроксимации Дворецкого. Понятно, что в этом случае процесс обучения протекает слишком медленно. В [11] для вычисления шага поиска была предложена процедура

$$\eta(k) = r^{-1}(k), \; r(k) = \alpha r(k-1) + \left\|x(k)\right\|^2 = ar(k-1) + 1, \; 0 \le \alpha \le 1,$$

при этом при $\alpha = 1$ параметр шага $\eta(k) = k^{-1}$, т.е. удовлетворяет условиям Дворецкого. Варьируя фактором забывания $\alpha$, несложно обеспечить достаточно широкий интервал изменения шага поиска

$$\frac{1}{k} \le \eta(k) < 1,$$

при этом $\alpha < 1$ обеспечивает LVQ-процедуре следящие свойства, необходимые в случае, если центроиды классов дрейфуют во времени.

Отметим также, что нормирование входящих сигналов $x(k)$ вовсе не гарантирует того, что прототипы классов будут отвечать условию $\left\|c_j(k)\right\| = 1$, а его невыполнение делает невозможным в качестве оценки расстояния использование скалярного произведения (1). Обойти данное затруднение несложно, введя

дополнительное нормирование синаптических весов в процессе обучения. В результате приходим к адаптивной процедуре вида

$$c_j(k)=\begin{cases}\dfrac{c_{j^*}(k-1)+\eta(k)\left(x(k)-c_{j^*}(k-1)\right)}{\left\|c_{j^*}(k-1)+\eta(k)\left(x(k)-c_{j^*}(k-1)\right)\right\|}-\text{если } x(k) \text{ и } c_{j^*}(k-1) \text{ принадлежат одному классу},\\ \dfrac{c_{j^*}(k-1)-\eta(k)\left(x(k)-c_{j^*}(k-1)\right)}{\left\|c_{j^*}(k-1)-\eta(k)\left(x(k)-c_{j^*}(k-1)\right)\right\|}-\text{если } x(k) \text{ и } c_{j^*}(k-1) \text{ принадлежат разным классам},\\ \eta(k)=r^{-1}(k), r(k)=\alpha r(k-1)+1, 0<\alpha\le 1,\\ c_j(k-1)-\text{если } j-\text{ый нейрон не победил}.\end{cases} \quad (3)$$

Рассчитанные с помощью правила обучения (3) прототипы $c_j(k)$ ($c_j(N)$ в случае, если обучающая выборка имеет фиксированный объем) подаются на входной слой, где вычисляются уровни принадлежности

$$u_j(k)=\frac{\left\|x(k)-c_j(N)\right\|^{-2}}{\sum_{l=1}^{m}\left\|x(k)-c_l(N)\right\|^{-2}}, \quad (4)$$

определяемые координатами седловой точки лагранжиана

$$\nabla_{u_j} L(u_j,c_j,\lambda)=\nabla_{u_j}\left(\sum_{k=1}^{N}\sum_{j=1}^{m}u_j^2(k)\left\|x(k)-c_j\right\|^2+\lambda\left(\sum_{j=1}^{m}u_j(k)-1\right)\right)=\vec{0},$$

лежащего в основе широко распространенного метода нечетких С-средних (FCM) вероятностной нечеткой кластеризации [5], (здесь $\lambda$ — неопределенный множитель Лагранжа).

Переписав (4) в виде

$$u_j(k)=\frac{1}{1+\left\|x(k)-c_j(k)\right\|^2\sum\limits_{\substack{l=1\\ l\ne j}}^{m}\left\|x(k)-c_l(k)\right\|^{-2}}=\frac{1}{1+\dfrac{\left\|\left\|x(k)-c_j(k)\right\|^2\right\|}{\sigma_j^2(k)}}, \quad (5)$$

несложно заметить, что выражение (5) задает колоколообразную функцию принадлежности с центром в точке $c_j(k)$ и параметром ширины

$$0\le\sigma_j^2=\left(\sum_{\substack{l=1\\ l\ne j}}^{m}\left\|x(k)-c_l(N)\right\|^{-2}\right)^{-1}\le\frac{4}{m-1},$$

т.е. вопрос о конкретном виде функции принадлежности в отличие от [14,15] здесь решается автоматически.

Таким образом, соотношения (3),(4) задают on-line алгоритм обучения адаптивной нечеткой нейронной сети обучаемого векторного квантования.

**Выводы**

Рассмотрена задача автоматической классификации текстовых документов, поступающих на обработку в реальном времени. Предложена архитектура адаптивной нечеткой нейронной сети обучаемого векторного квантования (AFLVQ) и on-line алгоритм ее обучения, отличающийся вычислительной простотой и высоким быстродействием.

**Список литературы:** **1.**Umer M.F., Khiyal M.S.H. Classification of textual documents using learning vector quantization// Information Technology Journal.–2007.–6(1).–p.154-159. **2**. *Ciarelli P.M., Oliveira E.* An enhanced probobalistic neural network approach applied to text classification// Lecture Notes on Computer Science.-V.5856.–Berlin- Heidelberg: Springer-Verlag, 2009.–p.661-668. **3**. *Bodyanskiy Ye., Shubkina O.* Semantic annotation of text documents using modified probobalistic 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. **4**. *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. **5**. *Bezdek J.C., Keller J.,* Krishnapuram R., Pal N.R. Fuzzy Models and Algorithms for Pattern Recognition and Image Processing.–N.Y.: Springer Science + Business Media, Inc., 2005.–776 p. **6**. *Kohonen T.* Self-Organizing Maps.-Berlin: Springer, 1995.–362 p. **7**.*Kohonen T.* Improved version of learning vector quantization// Proc. Int. Joint Conf. on Neural Networks.- San Diego, CA, 1990.-1.–p.545-550. **8**. *Hammer B., Villmann T.* Generalized relevance learning vector quantization// Neural Networks.-2002.–15.–p.1059-1068. **9**. *Biehl M., Ghosh A., Hammer B.* Learning vector quantization:The dynamics of winner-takes-all algorithm// Neurocomputing.–2006.–69.–p.660-670. **10**. *Sanches J.S., Marques A.I.* An LVQ-based adaptive algorithm for learning from very small codebooks// Neurocomputing.–2006.–69.–p.922-927. **11**.*Бодянский Е.В., Руденко О.Г.* Искусственные нейронные сети: архитектуры, обучение, применения. –Харьков: ТЕЛЕТЕХ, 2004.–372 с. **12**. *Руденко О.Г., Бодянський Є.В.* Штучні нейронні мережі.–Харків: ТОВ «Компанія СМІТ», 2006.–404 с. **13**.*Souza R.M.*C.R. de, Silva Filho T.M. de. Optimized learning vector quantization classifier with an adaptive Euclidean distance// Lecture Notes in Computer Science.–V.5768.–Berlin-Heidelberg: Springer-Verlag, 2009.–p.799-806. **14.** *Karayiannis N.B.*, Pai P.-I. Fuzzy algorithm for learning vector quantization// IEEE Trans. on Neural Network.–1996.–7. –№5.–p.1196-1211. **15**.*Wu K.-L.*, Yang M.-S. A fuzzy-soft learning vector quantization//Neurocomputing.–2003.-55.–p.681-697. **16**. *Kim Y.-S., Kim S.-I.* Fuzzy neural network model using a fuzzy learning vector quantization with the relative distance// Proc.7th Int. Conf. on Hybrid Intelligent System "HIS 2007".–Kaiserlautern, Germany, 2007.–p.90-94. **17**. *Baras J.S. LaVigna A.* Convergence of Kohonen's learning vector quantization// Proc. Int. Joint Conf. on Neural Networks.– San Diego, CA, 1990.–V.3–p.17-20.

*Поступила в редколлегию 16.11.2011*