

УДК 004.853

У статті розглядається архітектура та алгоритм нечіткої ймовірнісної нейронної мережі, яка може виконувати завдання класифікації в режимі реального часу

Ключові слова: нечітка нейронна мережа, класифікація

В статье рассматривается архитектура и алгоритм нечеткой вероятностной нейронной сети, которая может выполнять задачу классификации в режиме реального времени

Ключевые слова: нечеткая нейронная сеть, классификация

This article discusses the architecture and algorithm of fuzzy probabilistic neural network that can perform a task on-line classification

Keywords: fuzzy probabilistic neural network, classification

КЛАССИФИКАЦИЯ ТЕКСТОВЫХ ДОКУМЕНТОВ С ПОМОЩЬЮ НЕЧЕТКОЙ ВЕРОЯТНОСТНОЙ НЕЙРОННОЙ СЕТИ

Е. В. Бодянский

Доктор технических наук, профессор, научный руководитель
Проблемная научно-исследовательская лаборатория
автоматизированных систем управления**

Контактный тел.: (057) 702-18-90

E-mail: bodya@kture.kharkov.ua

Н. В. Рябова

Кандидат технических наук, заведующая кафедрой*

Контактный тел.: (057) 702-13-37

E-mail: ryabova.nv@gmail.com

О. В. Золотухин

Аспирант*

Контактный тел.: 067-317-49-69

E-mail: zolotukhin.ov@gmail.com

Кафедра искусственного интеллекта

*Харьковский национальный университет радиоэлектроники
пр. Ленина, 14, к.255, г. Харьков, 61166

Введение

Достаточно эффективно для решения задачи классификации текстовых данных могут быть использованы вероятностные нейронные сети (PNN), введенные Д.Ф. Шпехтом [1], обучение которых «производится по принципу «нейроны в точках данных», что делает его крайне простым и быстрым. В [2-4] были введены модификации PNN, предназначенные для обработки текстовой информации и отличающиеся наличием элементов конкуренции в процессе обучения и возможностью коррекции рецепторных полей ядерных активационных функций.

Вместе с тем, использование PNN в задачах обработки текстов усложняется в случаях, когда объемы анализируемой информации велики, а векторы признаков (образы) имеют достаточно высокую размерность.

Это затруднение объясняется тем, что как в PNN, так и в других нейронных сетях, обучаемых по принципу «нейроны в точках данных» [5], количество нейронов первого скрытого слоя (слоя образов) определяется числом векторов-образов обучающей выборки N , что, естественно, приводит к снижению быстродействия и требует хранения всех данных, использованных в процессе обучения сети.

Для преодоления этого недостатка в [6] была предложена улучшенная вероятностная нейронная сеть (EPNN), где первый скрытый слой образован не образами, а прототипами классов, вычисленных с помощью обычного K -среднего (НСМ) в пакетном режиме.

Поскольку в задачах классификации число возможных классов m обычно существенно меньше объема обучающей выборки N , EPNN гораздо лучше приспособлена для решения прикладных задач, чем стандартная PNN.

Вместе с тем, можно отметить такие основные недостатки EPNN как возможность обучения только в пакетном режиме, когда обучающая выборка задана заранее, и четкий результат классификации (отнесение предъявляемого образа только к одному классу), в то время как при обработке текстовых документов достаточно часто возникает ситуация, когда анализируемый текст с различными уровнями принадлежности может одновременно относиться сразу к нескольким возможно сразу пересекающимся классам.

В настоящей работе предпринимается попытка синтеза нейро-фаззи сети, обучаемой в on-line режиме и предназначенной для нечеткой классификации текстовых документов, представленных в форме векторов-образов, последовательно поступающих на обработку.

1. Нечеткая вероятностная нейронная сеть

Архитектура предлагаемой нами нейро-фаззи сети представлена на рис. 1.

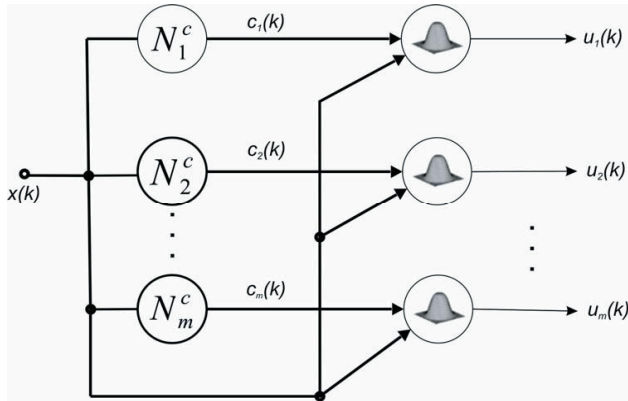


Рис. 1. Архитектура нечеткой вероятностной нейронной сети (FPNN)

Данная нейро-фаззи сеть содержит два слоя обработки информации: первый скрытый слой прототипов (вместо первого скрытого слоя образов в обычной PNN) и выходной слой вычисления уровней принадлежности. Исходной информацией для обучения является классифицированная последовательность векторов-образов

$$x(1), x(2), \dots, x(k), \dots, x(N),$$

$$x(k) = (x_1(k), x_2(k), \dots, x_n(k))^T \in R^n,$$

при этом предполагается, что N может изменяться со временем, количество классов m также может изменяться, прототипы (центроиды) классов описываются векторами $c_j = (c_{j1}, c_{j2}, \dots, c_{jn})^T$, подлежащими определению, обозначение $x(k, j)$ относим образ $x(k)$ к j -му классу, $j = 1, 2, \dots, m$, каждый класс содержащим N_j классифицированных образов, $\sum_{j=1}^m N_j = N$.

Для вычисления прототипов в [6] используется обычная оценка среднего арифметического (НСМ)

$$c_j = \frac{1}{N_j} \sum_{k=1}^{N_j} x(k, j),$$

которую несложно переписать в рекуррентной форме

$$c_j(k) = c_j(k-1) + \frac{1}{k} (x(k, j) - c_j(k-1)), \quad (1)$$

что, как можно заметить, соответствует правилу обучения Т.Кохонена [7] с параметром шага

$$\eta(k) = \frac{1}{k},$$

соответствующим условиям стохастической аппроксимации.

Поскольку в реальных задачах прототипы классов могут дрейфовать во времени, вместо (1) мож-

но использовать либо экспоненциально взвешенное среднее

$$c_j(k) = \alpha c_j(k-1) + (1-\alpha)(x(k, j) - c_j(k-1)), \quad 0 < \alpha < 1,$$

либо адаптивную процедуру [8]

$$\begin{cases} c_j(k) = c_j(k-1) + \eta(k)(x(k, j) - c_j(k-1)) \\ \eta(k) = \gamma^{-1}(k), \quad \gamma(k) = \alpha \gamma(k-1) + \|x(k, j)\|^2, \quad 0 \leq \alpha \leq 1, \end{cases}$$

удовлетворяющую при $\alpha=1$ стохастической аппроксимации А. Дворецкого.

Выходной слой сети оценивает уровень принадлежности классифицированных наблюдений $x(k)$ ($k > N$) к сформированным классам с прототипами $c_j(N)$ с помощью выражения

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

лежащего в основе вероятностной процедуры нечеткой классификации, известный как метод нечетких С-средних (FCM) [9].

Таким образом, в процессе обучения сети одновременно используются четкие и нечеткие процедуры (НСМ и FCM).

Переписав (2) в виде

$$u_j(k) = \frac{1}{1 + \|x(k) - c_j(N)\|^2 \sum_{l=1, l \neq j}^m \|x(k) - c_l(N)\|^{-2}}, \quad (3)$$

можно заметить, что (3) есть не что иное, как колоколообразная (ядерная) функция активации

$$u_j(k) = \frac{1}{1 + \frac{\|x(k) - c_j(N)\|^2}{\sigma_j^2}},$$

с параметром ширины рецепторного поля

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

устанавливаем автоматически в процессе классификации.

Поскольку выражения (2), (3) относятся к вероятностной нечеткой классификации [10], т.е. выполняется условие

$$\sum_{j=1}^m u_j(k) = 1,$$

ситуация, при которой $u_j(k) = m^{-1} \forall j$, означает, что наблюдение $x(k)$ относится либо ко всем классам одинаково, что является маловероятно, либо ни к одному из них.

В этой ситуации можно увеличить число классов до $m+1$, положив $x(k)$ в качестве начального прототипа нового класса. Если же будет обнаружено, что для p классов $p < m$ уровень принадлежности $c_j(k)$ окажется меньше m^{-1} , это означает, что $x(k)$ не может принадлежать этим классам, и уровни принадлежности следует пересчитать с помощью выражения (2), изменив верхний индекс суммирования в знаменателе на $m-p$.

Для исключения возможных p классов, не включающих в себя $x(k)$, может быть также использована процедура, основанная на V -критерии [11] (Vigilance criterion) и проверке условия

$$e^{u_j(k)} \|x(k) - c_j(N)\| \leq \varepsilon,$$

где пороговое значение ε устанавливается эмпирически. Понятно, что при $p = m-1$, получаем четкий результат классификации.

Выводы

Рассмотрена задача on-line классификации текстовых документов, поступающих на обработку последовательно в реальном времени. Введена нечеткая модификация вероятностной нейронной сети, отличающаяся крайне простой численной реализацией и малым объемом необходимой для ее реализации памяти. Рассмотрена задача автоматической классификации текстовых документов.

Литература

1. Specht D.F. Probobalistic neural networks // Neural Networks. – 1990.-3.-p.109-118.
2. Бодянский Е.В., Шубкина О.В. Семантическое аннотирование текстовых документов с использованием модифицированной вероятностной нейронной сети// Системные технологии.- Днепропетровск, 2011. – Вып.4 (75).-с.48-55.
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. Zahirnak D.R., Chapman R., Rogers S.K., Suter B.W., Kabriski M., Pyatti V. Pattern recognition using radial basis function network// Proc. 6th Ann. Aerospace Application of AI Conf., Dayton, OH, 1990.- p.249-260.
6. 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.
7. Kohonen T. Self-Organizing Maps.-Berlin: Springer, 1995.-362 p.
8. Бодянский Е.В., Руденко О.Г. Искусственные нейронные сети: архитектура, обучение, применение.-Харьков: ТЕЛТЕХ, 2004.-372 с.
9. Bezdek J.C., Hathaway R.J., Sabin M.J., Tucker W.T. Convergence theory for fuzzy c-means: Counterexamples and repairs//IEEE Transactions on Systems, Man, and Cybernetics. Vol. SMC-17, no. 5, 1987.- p.873-877.
10. 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.
11. Cannady J., Garcia R.G. The Application of Fuzzy ARTMAP in the Detection of Computer Network Attacks// Proc. Int. Conf. Artificial Neural Networks – ICANN 2001, Vienna, Austria, 2001.- p.225-230.