НЕЙРОІНФОРМАТИКА ТА ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ

НЕЙРОИНФОРМАТИКА И ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

NEUROINFORMATICS AND INTELLIGENT SYSTEMS

УДК 004.91

Е. В. Бодянский, В. В. Волкова, А. С. Егоров

КЛАСТЕРИЗАЦИЯ МАССИВОВ ТЕКСТОВЫХ ДОКУМЕНТОВ НА ОСНОВЕ АДАПТИВНОЙ НЕЧЕТКОЙ САМООРГАНИЗУЮЩЕЙСЯ НЕЙРОННОЙ СЕТИ

Предложена адаптивная нечеткая самоорганизующаяся нейронная сеть, настраиваемая с помощью рекуррентного алгоритма самообучения, являющегося обобщением правила обучения Кохонена, и позволяющая находить в реальном времени не только прототипы (центроиды) формируемых кластеров, но и оценивать уровни принадлежности каждого вновь поступившего образа к конкретному кластеру, что позволяет использовать данную архитектуру для кластеризации текстовых документов в условиях взаимно перекрывающихся классов.

ВВЕДЕНИЕ

В общей проблеме интеллектуального анализа данных – Data Mining, Explorary Data Analysis и, особенно, Web-Mining достаточно часто возникает задача поиска и классификации информации, содержащейся в текстовых документах, количество которых в Internet практически неограниченно и постоянно увеличивается. Фактически речь идет об очень больших и непрерывно растущих в реальном времени базах данных, образованных, как правило, не связанными между собой текстами самого различного содержания и происхождения, поиск в которых также должен производиться в online режиме.

В настоящее время существует достаточно много подходов к решению этой задачи, однако, большинство из них связано с интенсивным использованием человеческого интеллекта и квалифицированного труда, которые весьма дороги. В связи с этим представляется перспективным использование методов искусственного и вычислительного интеллекта для решения этой задачи в автоматическом режиме без участия человека. Среди таких методов высокую эффективность продемонстрировали искусственные нейронные сети и, прежде всего, самоорганизующиеся карты Т. Кохонена (SOM) [1], положенные в основу систем автоматической классификации больших массивов документов WEBSOM [2, 3] и WEBSOM2 [4]. Эффективность карт Кохонена определяется, прежде всего, их вычислительной простотой и возможностью работы в реальном времени путем последовательной обработки информации по мере ее поступ-

ления. Процесс настройки этих нейросетей реализуется в режиме самообучения на основе принципов «победитель получает все» (WTA) или «победитель получает больше» (WTM), при этом априори предполагается, что структура обрабатываемых данных такова, что образуемые ими кластеры взаимно не пересекаются, т. е. в процессе обучения сети теоретически может быть построена разделяющая гиперповерхность, четко разграничивающая разные классы.

Вместе с тем, при обработке реальных данных часто возникает ситуация, когда один образ-документ принадлежит разным классам, а сами эти классы взаимно пересекаются (перекрываются) [4]. В рамках традиционных самоорганизующихся карт это обстоятельство никак не учитывается, однако может быть рассмотрено с позиций нечеткого кластерного анализа, который к настоящему времени также получил достаточное развитие и распространение [5, 6].

Представляется естественным объединить простоту и быстродействие самоорганизующихся карт Кохонена с возможностью работы в условиях взаимно перекрывающихся классов.

Так, в [7, 8] была предложена модификация SOM, в которой нейроны исходной архитектуры, представляющие собой по сути адаптивные линейные ассоциаторы, заменены нечеткими множествами и нечеткими правилами. Данная нейросеть подтвердила эффективность в задачах распознавания образов, однако ее обучение связано с рядом существенных проблем. В [9] была предложена модификация самоорганизующейся карты с нечетким выводом и комбинированным алгоритмом самообучения на основе правил Кохонена и Гроссберга. Недостатком этой сети является наличие свободных параметров алгоритма, неудачный выбор которых может привести к неудовлетворительной кластеризации. В [10] была введена, а в [11] получила развитие, так называемая, нечеткая кластеризующая сеть Кохонена (fuzzy Kohonen clustering network – FKCN), в основе которой лежит алгоритм нечетких С-средних (fuzzy C-means – FCM) Бездека [12]. Особенностью этой нейро-фаззи сети является пакетный режим обучения, при котором весь массив данных, подлежащий обработке, должен быть задан априори. Таким образом FKCN не может работать в реальном времени, обрабатывая информацию по мере ее поступления.

В связи с этим, в настоящей работе предлагается в качестве альтернативы SOM и FKCN адаптивная нечеткая самоорганизующаяся нейронная сеть, настраиваемая с помощью рекуррентного алгоритма самообучения, являющегося обобщением правила обучения Кохонена, и позволяющая находить в реальном времени не только прототипы (центроиды) формируемых кластеров, но и оценивать уровни принадлежности каждого вновь поступившего образа к конкретному кластеру.

АРХИТЕКТУРА АДАПТИВНОЙ НЕЧЕТКОЙ САМООРГАНИЗУЮЩЕЙСЯ НЕЙРОННОЙ СЕТИ

Архитектура рассматриваемой нечеткой нейронной сети приведена на рис. 1 и содержит единственный слой нейронов N_i, $i = 1, 2, \ldots, p$, отличающихся от традиционных адаптивных линейных ассоциаторов, образующих SOM Кохонена.

На рецепторный слой сети последовательно подаются образы, подлежащие кластеризации, в виде $(n \times 1)$-векторов признаков $x(t) = (x_1(t), x_2(t), \ldots, x_n(t))^T$, где $t = 1, 2, \ldots, V$ имеет смысл или номера образа в обучающей выборке, или текущего дискретного времени. При этом сами векторы признаков $x(t)$ формируются на основе усеченных гистограмм частот появления отдельных слов в обрабатываемых текстах [2–4].

Настраиваемые синаптические веса m_{ij}, $i = 1, 2, \ldots, p$; $j = 1, 2, \ldots, n$ определяют координаты центроидов p взаимно перекрывающихся кластеров $m_i(t) = (m_{i1}(t), m_{i2}(t), \ldots, m_{in}(t))^T$, а выходом сети, в отличие от SOM, выходной сигнал которой определяется только нейроном-победителем, является $(p \times 1)$-вектор $u(t) = (u_1(t), u_2(t), \ldots, u_p(t))^T$, определяющий уровень принадлежности образа $x(t)$ к каждому из p формируемых кластеров и вычисляемый нейронами N_i. По латеральным связям нейроны обмениваются координатами $m_i(t)$, необходимыми для вычисления принадлежностей $u_i(t)$.

АДАПТИВНЫЙ АЛГОРИТМ САМООБУЧЕНИЯ

В основе самообучения лежит вероятностный алгоритм кластеризации, основанный на оптимизации целевой функции вида [12]

$$E(u_i, m_i) = \sum_{t=1}^{V} \sum_{i=1}^{p} u_i^{\beta}(t) \| x(t) - m_i \|^2 \qquad (1)$$

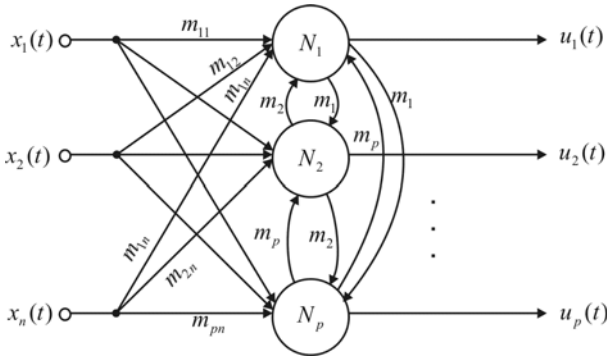

Рисунок 1 – Адаптивная нечеткая самоорганизующаяся нейронная сеть

при ограничениях

$$\sum_{i=1}^{p} u_i(t) = 1, \quad t = 1, 2, \ldots, V, \tag{2}$$

$$0 \le \sum_{t=1}^{V} u_i(t) \le V, \quad i = 1, 2, \ldots, p, \tag{3}$$

где $u_i(t) \in [0, 1]$, β – неотрицательный параметр, именуемый «фаззификатором», определяющий нечеткую границу между классами и влияющий на уровень нечеткости в окончательном разбиении данных по кластерам.

Применение стандартного аппарата нелинейного программирования, основанного на неопределенных множителях Лагранжа и решении системы уравнений Куна – Таккера, ведет к известному результату

$$\begin{cases} m_i^* = \dfrac{\sum\limits_{t=1}^{V} u_i^{\beta}(t) x(t)}{\sum\limits_{t=1}^{V} u_i^{\beta}(t)}, \\ u_i(t) = \dfrac{(\|x(t) - m_i^*\|^2)^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{p} (\|x(t) - m_l^*\|^2)^{\frac{1}{1-\beta}}}, \end{cases} \tag{4}$$

который при $\beta = 2$ совпадает с популярным FCM-алгоритмом Бездека:

$$\begin{cases} m_i^* = \dfrac{\sum\limits_{t=1}^{V} u_i^2(t) x(t)}{\sum\limits_{t=1}^{V} u_i^2(t)}, \\ u_i(t) = \dfrac{\|x(t) - m_i^*\|^{-2}}{\sum\limits_{l=1}^{p} \|x(t) - m_l^*\|^{-2}}. \end{cases} \tag{5}$$

Именно соотношения (5) положены в основу самообучения FKCN, однако при этом количество обрабатываемых образов V полагается фиксированным.

С целью преодоления этого ограничения в [13, 14] на основе процедуры нелинейного программирования Эрроу – Гурвица был введен вероятностный рекуррентный алгоритм нечеткой кластеризации вида

$$\begin{cases} m_i(t+1) = m_i(t) = \alpha(t) u_i^{\beta}(t)(x(t+1) - m_i(t)), \\ \qquad\qquad i = 1, 2, \ldots, p, \\ u_i(t+1) = \dfrac{(\|x(t+1) - m_i(t+1)\|^2)^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{p} (\|x(t+1) - m_l(t+1)\|^2)^{\frac{1}{1-\beta}}}, \end{cases} \tag{6}$$

где $\alpha(t)$ – параметр шага поиска, влияющий на скорость сходимости и выбираемый обычно из эмпирических соображений в соответствии с условиями Дворецкого [15].

Анализируя (6), можно заметить, что рассматривая сомножитель $u_i^{\beta}(t)$ в качестве функции соседства $h_{c(x), i}$, приходим к правилу самообучения Кохонена на основе WTM-принципа

$$m_i^{\mathrm{WTM}}(t+1) = m_i^{\mathrm{WTM}}(t) + h_{c(x), i}(t)(x(t+1) - m_i^{\mathrm{WTM}}(t)), \tag{7}$$

где $c(x) = \arg\min\limits_i \{\|x - m_i\|\}$ определяет координаты нейрона-победителя, $h_{c(x), i(t)}$ – колоколообразная функция соседства, аргументом которой есть расстояние в принятой метрике между центроидом нейрона-победителя и нейрона N_i.

Заметим также, что в пакетной форме рекуррентной формуле (7) соответствует выражение [4]

$$m_i^{*\mathrm{WTM}} = \frac{\sum\limits_{x(t) \in V_i} h_{c(x), i} x(t)}{\sum\limits_{x(t) \in V_i} h_{c(x), i}}, \tag{8}$$

где V_i определяет множество всех образов, прототипом которых является $m_i^{*\mathrm{WTM}}$, $\sum\limits_{i=1}^{P} V_i = V$.

Как видно, формула (8) структурно совпадает с первым выражением в (4), что опять-таки подтверждает близость понятий «принадлежности» и «соседства».

Полагая далее в (6) $\beta = 1$, приходим к алгоритму С-средних (hard C-means – HCM), а $\beta = 0$ соответствует стандартному WTA-правилу Кохонена для нейрона-победителя:

$$m_i^{\mathrm{WTA}}(t+1) = m_i^{\mathrm{WTA}}(t) + \alpha(t)(x(t+1) - m_i^{\mathrm{WTA}}(t)). \tag{9}$$

Несложно заметить также, что рекуррентная процедура (9) минимизирует целевую функцию вида

$$E(m_i) = \sum_{x(t) \in V_i} \|x(t) - m_i^{\mathrm{WTA}}\|^2.$$

Ее прямая оптимизация ведет к обычной оценке среднего арифметического

$$m_i^{*\mathrm{WTA}} = \frac{\sum\limits_{x(t)\in V_i} x(t)}{V_i},$$

запись которой в рекуррентной форме – к соотношению

$$m_i^{\mathrm{WTA}}(t+1) = m_i^{\mathrm{WTA}}(t) + \frac{1}{t+1}(x(t+1) - m_i^{\mathrm{WTA}}(t)).$$

Такой выбор параметра шага $\alpha(t)$ согласуется с требованиями стохастической аппроксимации и придает результатам ясный физический смысл.

Таким образом, в окончательном виде адаптивный алгоритм самообучения нечеткой самоорганизующейся сети может быть записан в простой форме

$$\begin{cases} m_i(t+1) = m_i(t) + \dfrac{u_i^{\beta}(t)}{t+1}(x(t+1) - m_i(t)), \\ \qquad\qquad i = 1, 2, \ldots, p, \\ u_i(t+1) = \dfrac{(\|x(t+1) - m_i(t+1)\|^2)^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m}(\|x(t+1) - m_l(t+1)\|^2)^{\frac{1}{1-\beta}}}, \end{cases}$$

объединяющей в себе вычислительную простоту и последовательную обработку кохоненовского самообучения с возможностями нечеткой кластеризации.

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТАЛЬНЫХ ИССЛЕДОВАНИЙ

В ходе изучения нейросетевых методов кластеризации и разработки адаптивной нечеткой самоорганизующейся нейронной сети были проведены экспериментальные исследования на тестовой выборке, состоящей из 86 текстовых документов, содержащих рефераты статей (abstracts). Документы принадлежат к трем различным категориям: Neural Networks, Semantic Web, Text Mining.

Целью исследования было оценить качество кластеризации традиционной самоорганизующейся нейронной сети Кохонена (SOM) и разработанной в ходе исследования адаптивной нечеткой самоорганизующейся нейронной сети (адаптивная процедура обучения на основе FCM).

Следует отметить, что качество кластеризации сильно зависит от выбранного пространства признаков. Пространство признаков выбиралось по значимости признаков согласно формуле

$$tf*idf = tf_{ij} * \log_2\left(\frac{|D|}{df_i}\right), \qquad (10)$$

где tf_{ij} – частота встречаемости i-го терма в j-м документе, $|D|$ – мощность обучающего множества, df_i – количество документов, в которых встречается i-й терм.

Эксперимент показал, что по мере роста тестового множества нечеткие алгоритмы дают более точные результаты (в среднем 6–8 %) по сравнению с четкой процедурой кластеризации.

Таким образом, было установлено, что в задаче кластеризации документов, принадлежащих нескольким категориям одновременно, нечеткие процедуры дают более точные результаты.

ВЫВОДЫ

Предложен адаптивный алгоритм самообучения нечеткой самоорганизующейся нейронной сети, предназначенной для кластеризации больших массивов текстовых документов, и позволяющий осуществлять в реальном времени нечеткую классификацию данных, последовательно поступающих на обработку. Алгоритм не содержит свободных параметров, прост в реализации и объединяет в себе достоинства самоорганизующихся карт Кохонена и вероятностных процедур нечеткой кластеризации.

ПЕРЕЧЕНЬ ССЫЛОК

1. *Kohonen T.* Self-Organizing Maps / T. Kohonen // Berlin : Springer-Verlag. – 1995. – 362 p.
2. *Kaski S.* WEBSOM – Self-organizing maps of document collections / S. Kaski, T. Honkela, K. Lagus, T. Kohonen // Neurocomputing. – 1998. – 21. – P. 101–117.
3. *Lagus K.* WEBSOM for textual data mining / K. Lagus, T. Honkela, S. Kaski, T. Kohonen // Artificial Intelligence Review. – 1999. – 13. – P. 345–364.
4. *Kohonen T.* Self organization of a massive document collection / T. Kohonen, S. Kaski, K. Lagus, J. Salojärvi, J. Honkela, V. Paatero, A. Saarela // IEEE Trans. on Neural Networks. – 2000. – 11. – P. 574–585.
5. *Höppner F.* Fuzzy-Klusteranalyse. Verfahren für die Bilderkennung, Klassification und Datenanalyse / F. Höppner, F. Klawonn, R. Kruse. – Braunschweig : Vieweg. – 1996. – 280 s.
6. *Höppner F.* Fuzzy Clustering Analysis: Methods for Classification, Data Analysis, and Image Recognition / F. Höppner, F. Klawonn, R. Kruse, T. Runkler. – Chichester : John Willey&Sons. – 1999. – 289 p.
7. *Vuorimaa P.* Fuzzy self-organizing maps / P. Vuorimaa // Fuzzy Sets and Systems. – 1994. – 66. – P. 223–231.
8. *Vuorimaa P.* Use of the fuzzy self-organizing maps in pattern recognition / P. Vuorimaa // Proc. 3-rd IEEE Int.Conf. Fuzzy Systems «FUZZ-IEEE'94». – Orlando, USA, 1994. – P. 798–801.
9. *Bodyanskiy Ye.* Combined learning algorithm for a self-organizing map with fuzzy inference / Ye. Bodyanskiy, Ye. Gorshkov, V. Kolodyazhniy, A. Stephan ; ed. by B. Reusch // Computational Intelligence, Theory and Applications. – Berlin-Heidelberg : Springer, 2005. – P. 641–650.
10. *Tsao E. C.-K.,* Fuzzy Kohonen clus-tering networks / E.C.-K. Tsao, J.C. Bezdek, N. R. Pal // Pattern Recognition. – 1994. – 27. – P. 757–764.

11. *Pascual-Marqui R. D.* Smoothly distributed fuzzy C-means: a new self-organizing map / R. D. Pascual-Marqui, A. D. Pascual-Montano, K. Kochi, J. M. Carazo // Pattern Recognition. – 2001. – 34. – P. 2395–2402.
12. *Bezdek J. C.* Pattern Recognition with Fuzzy Objective Function Algorithms / J. C. Bezdek. // N. Y. : Plenum Press, 1981. – 272 p.
13. *Bodyanskiy Ye.* Recursive fuzzy clustering algorithms / Ye. Bodyanskiy, V. Kolodyazhniy, A. Stephan // Proc. East West Fuzzy Coll, 2002. – Zittau – Görlitz : HS, 2002. – P. 164–172.
14. *Bodyanskiy Ye.* Computational intelligence techniques for data analysis / Ye. Bodyanskiy // Lecture Notes in Informatics. – Bonn : GI, 2005. – P-72. – P. 15–36.
15. *Dvoretzky A.* On stochastic approximation / A. Dvoretzky // Proc. 3-rd Berkley Symp. Math. Statistics and Probability. – 1956. – 1. – P. 39–55.

Надійшла 31.10.2008

Запропоновано нечітку нейронну мережу, що самоорганізується, яка дозволяє знаходити в реальному часі не лише прототипи (центроїди) кластерів, що формуються, але й оцінювати рівні належності кожного образу, що надходить, до конкретного кластеру. Мережа настроюється за допомогою рекурентного алгоритму самонавчання, що є узагальненням правила навчання Кохонена. Запропонована нечітка нейронна мережа, що самоорганізується, може бути використана для кластеризації текстових документів в умовах класів, що взаємно перекриваються.

A self-organizing fuzzy neural network is proposed. It allows both to deter-mine the prototypes (centroids) of forming clusters and estimate attachment level of each image of certain cluster in real time. The network is tuned by recurrent algorithm of self-learning which is a general Kohonen's learning rule. The proposed neural network can be used in clasterization of text documents in overlapping classes conditions.