Journal of Southeast University (English Edition) Vol. 22, No. 3, pp. 439 – 444 Sept. 2006 ISSN 1003—7985

Document classification approach by rough-set-based corner classification neural network

Zhang Weifeng[1,2,3] Xu Baowen[2,3] Cui Zifeng[2,3] Xu Junling[2,3]

([1]College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China)
([2]College of Computer Science and Engineering, Southeast University, Nanjing 210096, China)
([3]Jiangsu Institute of Software Quality, Nanjing 210096, China)

Abstract: A rough set based corner classification neural network, the Rough-CC4, is presented to solve document classification problems such as document representation of different document sizes, document feature selection and document feature encoding. In the Rough-CC4, the documents are described by the equivalent classes of the approximate words. By this method, the dimensions representing the documents can be reduced, which can solve the precision problems caused by the different document sizes and also blur the differences caused by the approximate words. In the Rough-CC4, a binary encoding method is introduced, through which the importance of documents relative to each equivalent class is encoded. By this encoding method, the precision of the Rough-CC4 is improved greatly and the space complexity of the Rough-CC4 is reduced. The Rough-CC4 can be used in automatic classification of documents.

Key words: document classification; neural network; rough set; meta search engine

Information classification plays an important role in all information retrieval systems[1-7]. There are normally two types of information classification: artificial classification and automatic classification. In the earlier period of Yahoo, the artificial classification method was used. When a new file text is stored, it is first categorized by the editors. The method makes use of the human intelligence in the categorizing process. The second type, the automatic classification method, makes use of the computer to realize the classification, which is actually a clustering technique. At present, the commonly-used clustering methods include: the layer-based algorithms such as CHAMELEON[1], CURE[2] and BIRCH[3]), the surface-partition-based algorithms such as k-means[4] and FREM[5], the density-based algorithms such as DENCLUE[6], OPTICS[7] and DBSCAN[8]), the rule-and-model-based algorithm[9], the network-and-subspace-based algorithms such as STING[10], WaveCluster[11] and CLIQUE[12]). But recently, the knowledge-based clustering method is becoming the focus of study[13]. The knowledge-based method conducts machine learning with the existing classification information, stores the human's knowledge of classification with the method which can be understood by the computer, and then makes use of the knowledge when necessary. The artificial neural network imitates the human's brain activity and has the following advantages: the very strong non-linear approach, parallel processing in large scale, self training and studying, self-organizing, error tolerance, and so on. Meanwhile, the text file can be classified quickly and accurately by the neural network. Ref. [14] suggests using the corner classification neural network (CC4) to carry out the text file classification. Based on the analysis of the time complication and the space complication of the Extended-CC4[14], we present a rough-set-based corner classification neural network, the Rough-CC4. It describes the documents through a rough set, which can solve the problems of document concept description caused by the relations of approximate items and the problems of classification accuracy caused by the different sizes of documents. At the same time, it uses the binary coding-based L scattering method to improve the input accuracy of the Rough-CC4 and compress the storage space of the Rough-CC4.

This paper first introduces the basic concepts of the CC4 neural network and the Extended-CC4 nerve network. Then based on the analysis of the Extended-CC4, the rough set based vector representative of documents and the rough set based corner classification

Received 2006-04-06.

Foundation items: The National Natural Science Foundation of China (No. 60503020, 60373066, 60403016, 60425206), the Natural Science Foundation of Jiangsu Higher Education Institutions (No. 04KJB520096), the Doctoral Foundation of Nanjing University of Posts and Telecommunication (No. 0302).

Biographies: Zhang Weifeng (1975—), male, doctor, associate professor, wfzhang@yahoo.com; Xu Baowen (1961—), male, doctor, professor, bwxu@seu.edu.cn.

neural network, the Rough-CC4, are presented. After that the scattering input method in the Rough-CC4 by binary coding is presented.

1 Corner Classification Neural Network CC4 Based Document Classification Method

The CC4 (see Fig. 1) is a three-layered feedforward neural network having three layers: input layer, hidden layer, and output layer.

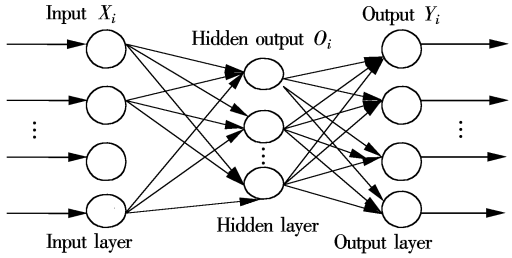

Fig. 1 Corner classsification neural network CC4

In the CC4, the input vector $\boldsymbol{X}$ represents the document vector, the output vector $\boldsymbol{Y}$ represents the class vector. Every neuron is a binary neuron which accepts only 0 or 1. The activation function of each neuron is as follows:

$$y = f\left(\sum x_i\right) = \begin{cases} 1 & \sum x_i > 0 \\ 0 & \sum x_i \leqslant 0 \end{cases} \tag{1}$$

where $x_i = 1$ or 0. The number of input neurons is equal to the length of the input vector plus one, the additional neuron being the bias neuron which has a constant input of 1. The number of hidden neurons is equal to the number of training samples and each hidden neuron represents one training sample.

The training of the CC4 neural network is very simple. Let the weight of the connection from input neuron i to hidden neuron j be $w_{i,j}$, which is decided by whether the j-th training sample includes the item represented by x_i. Let $X_{i,j}$ be the input for the i-th input neuron when the j-th training sample is presented to the network. It can be formalized as

$$w_{i,j} = \begin{cases} 1 & \text{if } X_{i,j} = 1 \\ -1 & \text{if } X_{i,j} = 0 \\ r - s + 1 & \text{if } i = n \end{cases} \tag{2}$$

where $i = 1, 2, \ldots, N$, N is the number of the neurons in the input layer; $j = 1, 2, \ldots, H$, H is the number of the neurons in the hidden layer; r is the user defined radius of generalization and s is the number of 1 in the training vector.

Let $u_{j,k}$ be the weight of connection from the j-th neuron in the hidden layer to the k-th neuron in the output layer. Let $Y_{j,k}$ be the output of the k-th output neuron for the j-th training sample. The formalized $u_{j,k}$ is as

$$u_{j,k} = \begin{cases} 1 & \text{if } Y_{j,k} = 1 \\ -1 & \text{if } Y_{j,k} = 0 \end{cases} \tag{3}$$

where $k = 1, 2, \ldots, M$ and M is the number of the output neurons (the number of the classes).

In meta search engine Anvish[13], the CC4 uses the term frequency (TF) vectors of documents as its input, i. e., if the TF is greater than 1, the binary input will be 1, or the binary input will be 0. In this way, the Anvish search engine can classify web pages at a precision around 80%. The shortcoming of this method is that the sizes of the input documents should be approximately the same, which cannot always be realized in the actual search process. In fact, there may be quite large differences among the sizes of several textual documents that belong to the same class. Therefore, we need to project the high dimension sparse document vectors into the low dimensional vectors, through which the documents with different sizes can be represented by relatively dense vectors. Ref. [14] puts forward a method to select the appropriate items step by step to project the space of n dimensions into the subspace of k dimensions. When selecting the subspace of k dimensions, the relative distances between the documents represented by the subspace of k dimensions should be the same as the distances between the documents represented by the space of n dimensions, shown as follows:

$$\text{Stress} = \sqrt{\frac{\sum_{i,j} (d'_{ij} - d_{ij})^2}{\sum_{i,j} d_{ij}^2}} \tag{4}$$

where d_{ij} is the distance between the original document vectors d_i and d_j, and d'_{ij} is the distance between the projected vectors $P_i = \{x_{i1}, \ldots, x_{ik}\}$ and $P_j = \{x_{j1}, \ldots, x_{jk}\}$ in the subspace of k dimensions of corresponding documents d_i and d_j.

Eq. (4) is used to select which k dimensions as the subspace. It first randomly selects the k dimensions from the n dimensions. The k dimensions having minimal stress are selected. The time complexity of this method is $O\left(\frac{m^2 n^{k+1}}{k!}\right)$, where n is the number of dimensions of original space, $k \leqslant n$, k is the number of dimensions of some subspace, and m is the number of training samples. For example, let the number of training samples m be 1 000, let n be 1 000, and let k be 100. The time is very considerable. If we use the quickest computer which can process 1000 billions

times every second, this process will require about 3×10^{280} years. So this algorithm cannot be applied practically and the relationships between the items are not considered in this method in the selection of the k dimensions. One strategy to solve this problem is to use the genetic algorithm to get the local superior solution, which will be discussed in other papers. Here, we build the new subspace which can show the relations between the items, and decrease the dimensions by projecting the original space into the new subspace. Not only the time complexity of this method is low but the approximate relations between the items are considered in this method.

2 Rough Set Based Document Representing

As the documents are usually written by different writers, one concept is often represented by different items. For example, "calculating machine" and "computer" actually represent the same concept. If the approximate items are placed together to form one new dimension in a subspace and the documents are mapped to these new dimensions, on the one hand, the number of dimensions representing the documents can be decreased, and on the other hand, it can be realized to categorize the documents according to the concepts, for the approximate items representing the same concept is represented by the same dimension. Therefore, the theory rough set[15] is applied, through which the divisions of the approximate items are built and the approximate items are placed in the same equivalent class.

The theory rough set, which was presented by Polish mathematician, Pawlak, in 1982, is the extension of the standard theory of aggregates. It can support the approximate decision-making in procedures of decision-making. It can efficiently analyze and process the imperfect information such as imprecise information, inconsistent information, half-baked information, etc. and discover the hidden knowledge, and reveal the potential rules. Its basic idea is as follows: Based on the equivalent relations in the considering field, the sets in the field are described by a pair of approximate operators(the inferior approximation and the superior approximation) . These operators can be used in the many application systems needing approximate computing. Based on the approximate relations, we construct the rough operators. The size of document representing can be reduced by representing the documents according to the equivalent relations, so the documents of different sizes can be mapped to the relatively small space, through which the imprecision caused by the problem of understanding the approximate relations in the procedure of document classification can be decreased.

Let R be the approximate relations in the dictionary T. An approximate space $\mathrm{apr} = (T, R)$ is constructed by the non-empty field of objects. The division based on the approximate items can be denoted as $T/R = \{C_1, C_2, \ldots, C_m\}$, where C_i is an equivalent class of R, i. e. , a group of approximate items. For the arbitrary subset S of T.

The inferior approximation of S is

$$\mathrm{lower\_apr}\ (S) = \{x \in C_i \mid C_i \subseteq S\}$$

The superior approximation of S is

$$\mathrm{upper\_apr}(S) = \{x \in C_i \mid C_i \cap S \neq \varnothing\}$$

The two kinds of approximation are actually the approximate description of S on the approximate space (T, R).

We can directly map a document to its rough-set-based representation now. To measure the importance of the documents on each equivalent class, we further describe the document by constructing the fuzzy set for each equivalent class.

To enhance the relation between the rough set and the fuzzy theory, we give a rough set based subjecting function, through which the rough set and the fuzzy set can be associated.

Definition 1 The fuzzy set on the discussion field T: If the discussion filed T is given, x is the arbitrary subset of T; R is the relations of approximate items, and $[x]_R$ is the x existing equivalent class of R. The fuzzy set $\tilde{X}$ on T is decided by X and R as

$$\tilde{X}_R(x) = \frac{|X \cap [x]_R|}{|[x]_R|} \tag{5}$$

Definition 2 The fuzzy set on the discussion field T/R: Based on the fuzzy set on the discussion field T, the fuzzy set on the discussion field T/R can be defined as

$$\tilde{X}_{T/R}(C_i) = \max_{x \in C_i}(\tilde{X}_R(x)) \tag{6}$$

In the fuzzy set on the T/R in definition 2, the frequency of items occurring in the document is not considered, and the importance relative to the set of all documents of items is not considered either. But these two aspects are very important in considering the subjecting degree of items existing in some equivalent class. Therefore, for the item $t(t \in T)$, the place in the document $d_j(d_j \in D)$ and the weight of tag $\mathrm{ptw}(t, d_j.\mathrm{place}, d_j.\mathrm{tag})$, we use the weight of item t in docu-

ment d_j is denoted as $\mathrm{stf}_j(t) = \sum_{t \in d_j} \mathrm{ptw}(t.\mathrm{place}, t.\mathrm{tag})$, through which the importance of some document in its equivalent class can be calculated by

$$d_j(C_i) = \sum_{t \in C_i} (\mathrm{stf}_j(t)\,\mathrm{idf}(t)) \tag{7}$$

The normalized importance can be calculated by

$$\tilde{d}_j(C_i) = \frac{d_j(C_i)}{\sum_{k=1}^{m} d_j(C_k)} \tag{8}$$

If the importance of the documents relative to every equivalent class can be computed, the fuzzy concept of the document d relative to the equivalent class of approximate items can be easily defined.

Definition 3 Let the fuzzy set of document d relative to the equivalent class of approximate items be $\{w_1(C_1), w_2(C_2), \ldots, w_m(C_m)\}$ denoted as $\tilde{d}$, where $C_1, C_2, \ldots, C_m \in T/R$ and $w_1, w_2, \ldots, w_m$ are the subjecting degree of the corresponding equivalent class belonging to document d, which are calculated by Eq. (8) and there is apparently $0 \leqslant w_1, w_2, \ldots, w_m \leqslant 1$.

3 Rough Set Based Corner Classification Neural Network

Only the hamming distance between the documents is considered in CC4, where the document is represented by vector TF $\mathrm{tf} = (\mathrm{tf}(t_1), \mathrm{tf}(t_2), \ldots, \mathrm{tf}(t_n))$ and n is the number of items in a dictionary. If $\mathrm{tf}(t_i) > 0$, the i-th input of CC4 is 1, otherwise it is 0. The relative importance of items in the documents is lost in this encoding method. The rough-set-based dimension decreasing of documents can solve the relation of approximate items based classification.

Ref. [14] considers that the CC4 uses the binary as its input, so the method of L dispersing is used. The procedure of L dispersing of $x(x \in [a, b])$ is as follows: First, let the dispersing length be L, then obtain $m = \frac{b-a}{L}$, $k = \frac{x-a}{m}$, and then encode it according to the following mode: let the frontal k elements be 1, and the rest $L-k$ elements be 0. The region is decided by the dispersing length L. The reciprocal of its precision is linear to L (The curve of diamond dots in Fig. 2). This relation causes a small precision to take a large space. Therefore, we present the binary encoding method through which the importance of documents relative to each equivalent class is encoded.

In order to realize the binary encoding, the result space of the original problem should be mapped to the space of the bit sequence $B = \{0, 1\}^L$. First, L is decided by the definition field of variable and its computing precision. For example, the definition of variable x is $[-2, 5]$ and the requirement of its precision is 10^{-6}, the filed $[-2, 5]$ is divided into 7×10^6 equivalent regions and each region is represented by one binary code. As $4\,194\,304 = 2^{22} < 7 \times 10^6 < 2^{23} = 8\,388\,608$, the vector of bit sequence space $(b_{22}, b_{21}, \ldots, b_0)$ corresponds to a point in region $[-2, 5]$. When a vector in the space of a bit sequence is known, it can be decoded by

$$x = -2.0 + x' \frac{7}{2^{23} - 1}$$

where $x' = \sum_{i=0}^{22} b_i \times 2^i$.

In general, suppose x is a real number in the region $[a, b]$. If the representing precision is 10^{-R} where H is a positive integer and it is equivalent to divide the region $[a, b]$ into $(b-a)10^R$ small regions, where each region is represented by a binary code., the length of vector of a bit sequence is decided by

$$L = \begin{cases} \log_2^{(b-a)10^R} & \log_2^{(b-a)10^R} \text{ is positive} \\ \log_2^{(b-a)10^R} + 1 & \text{otherwise} \end{cases} \tag{9}$$

Fig. 2 shows the relationships of the reciprocal of precision to the length of dispersing. It is obvious that the precision of the Rough-CC4 is far higher than the precision of the Extended-CC4 for the same L. Fig. 3 shows that the space requirement of the Rough-CC4 is far less than the space requirement of the Extended-CC4.

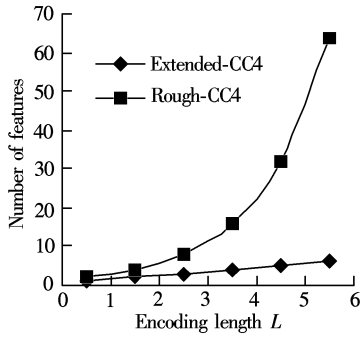

Fig. 2 Curve of number of features related to encoding length L

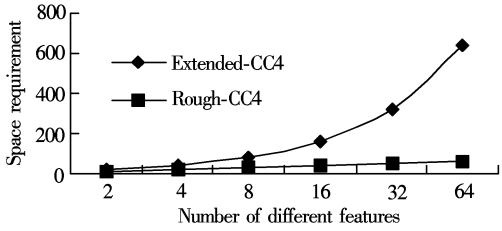

Fig. 3 Curve of relationships between different features and space requirement in L scattering

4 Prototype of Rough Set Based Corner Classification Neural Network

To further specify the document classifying method of rough set based corner classification neural network, we realize the prototype system (see Fig. 4) of rough set based corner classification neural network with agent. This prototype system shows the basic classifying procedure of rough set based corner classification neural network. . Compared to the general meta search engine, the machine learning agent, the result classifying agent and the corner classification neural network are added in Fig. 4. The machine learning agent runs at the side of the meta search engine, which trains the corner classification neural network from the historical classification information by the training algorithm of the Rough-CC4. The query request agent takes charge of distributing the users' query requests to the general search engine. The result classifying agent will map the obtained search results to their corresponding rough set vectors, then weights on every graduation of rough set will be dispersed by L, which will be input to the corner classification neural network, and then the classified search results will be returned to the users. The machine learning agent can train the corner classification neural network in advance, and it will not influence the responding speed of the meta search engine. But the ability of the result classifying agent will influence the speed of obtaining results.

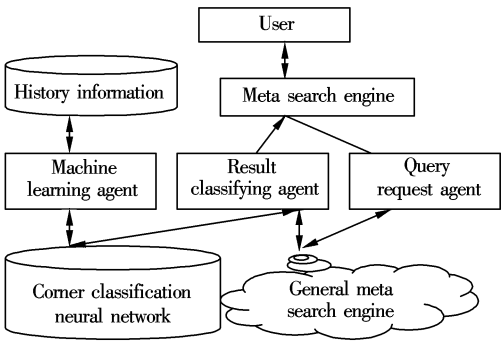

Fig. 4 Prototype of rough set based corner classification neural network

5 Conclusion

There is much research on the application of artificial intelligence in the field of information retrieval. The neural network is widely used for its advantages of parallel structure and parallel processing. Based on the existing training algorithms of corner classification CC4 and the extended corner classification, the Extended-CC4, the corresponding training algorithm of rough-set-based corner classification, the Rough-CC4, is presented in this paper. It uses the relations of approximate items to reduce the size of document representation, which adequately considers the importance of document structure information and equivalent classes related to locality and globality when considering the weight of equivalent classes of approximate items. By comparing the time complexity and space complexity between the Rough-CC4 and the Extended-CC4, we find the Rough-CC4 has the advantages of higher speed and less space requirement than the Extended-CC4. In the future, we will conduct further research on how to use the method of corner classification to realize the hierarchy document classification method.

References

[1] Karypis G, Han E H, Kumar V. CHAMELEON: a hierarchical clustering algorithm using dynamic modeling (No. 99-007) [R]. Department of Computer Science and Engineering of University of Minnesota, 1999.

[2] Guha S, Rastogi R, Shim K. CURE: an efficient clustering algorithm for large databases [A]. In: *Proc of the ACM SIGMOD Int'l Conf on Management of Data*[C]. Seattle, 1998. 73 – 84.

[3] Zhang T, Ramakrishnan R, Livny M. BIRCH: an efficient data clustering method for very large databases [A]. In: *Proc of the ACM SIGMOD Int'l Conf on Management of Data*[C]. Montreal, Canada, 1996. 103 – 114.

[4] Kamber M. *Data mining concepts and techniques* [M]. Translated by Fan M, Meng X F. Beijing: China Machine Press, 2001. (in Chinese).

[5] Ordonez C, Omiecinski E. FREM: fast and robust EM clustering for large data sets [A]. In: *Proc of the ACM CIKM Int'l Conf on Information and Knowledge Management*[C]. McLean, 2002. 590 – 599.

[6] Hinneburg A, Keim D. An efficient approach to clustering in large multimedia databases with noise [A]. In: *Proc of the 4th Int'l Conf on Knowledge Discovery and Data Mining* (*KDD*'98) [C]. New York: AAAI Press, 1998. 58 – 65.

[7] Ankerst M, Breunig M M, Kriegel H P, et al. OPTICS: ordering points to identify the clustering structure [A]. In: *Proc of ACM SIGMOD Int'l Conf on Management of Data* [C]. Philadelphia, 1999. 49 – 60.

[8] Ester M, Kriegel H, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise [A]. In: *Proc of the 2nd Int'l Conf on Knowledge Discovery and Data Mining* (*KDD*'96) [C]. Portland, 1996. 226 – 231.

[9] Song Q B, Shen J Y. A web document clustering algorithm

based on association rule [J]. *Journal of Software*, 2002, **13**(3): 417 - 423. (in Chinese)

[10] Wang W, Yang J, Muntz R R. STING: a statistical information grid approach to spatial data mining [A]. In: *Proc of the 23rd Int'l Conf on Very Large Data Bases* [C]. Athens, 1997. 186 - 195.

[11] Sheikholeslami G, Chatterjee S, Zhang A D. WaveCluster: a multi-resolution clustering approach for very large spatial databases [A]. In: *Proc of the 24th Int'l Conf on Very Large Data Bases* [C]. New York, 1998. 428 - 439.

[12] Rakesh A, Johanners G, Dimitrios G, Prabhakar R. Automatic subspace clustering of high dimensional data for data mining applications [A]. In: *Proc of the ACM SIGMOD Int'l Conf on Management of Data* [C]. Minneapolis, 1994. 94 - 105.

[13] Shu B, Kak S. A neural network-based intelligent meta search engine [J]. *Information Sciences*, 1999, **120**(1): 1 - 11.

[14] Chen Enhong. An extended corner classification neural network based classification approach [J]. *Journal of Software*, 2002, **13**(5): 871 - 878.

[15] Pawlak, Z. Rough sets [J]. *International Journal of Computer and Information Sciences*, 1982, **11**(5): 341 - 356.

一种基于粗糙集角分类神经网络的文档分类方法

张卫丰[1,2,3] 徐宝文[2,3] 崔自峰[2,3] 徐峻岭[2,3]

([1] 南京邮电大学计算机学院,南京 210003)

([2] 东南大学计算机科学与工程学院,南京 210096)

([3] 江苏省软件质量研究所,南京 210096)

摘要:针对文档分类过程中不同大小文档表示、文档特征选择和文档特征编码问题,提出了一种基于粗糙集的角分类神经网络 Rough-CC4. 利用近义词构成等价类,以此表示文档,可以缩小文档表示的维数、解决由于文档不同大小导致的精度问题、模糊近义词之间的差别;利用二进制编码方法对文档特征编码,可以提高 Rough-CC4 的精度,同时减小 Rough-CC4 的空间复杂度. Rough-CC4 可以广泛用于大量文档集合的自动分类.

关键词:文档分类;神经网络;粗糙集;元搜索引擎

中图分类号:TP183