2010 International Conference on Multimedia Information Networking and Security

Improved SOM Algorithm-HDSOM Applied in Text Clustering

SUN Ai-xiang
Management institute
Shandong University of Technology
Zibo, China
e-mail:aixiang12@163.com

***Abstract*—SOM neural network is one of the most commonly used Clustering algorithom in the text clustering. The initial connection weights of SOM neural network will affect the degree of convergence. If the Initial connection weights are not set appropriate, that will cause in a long wandering around the local minimum, accordingly lower the speed of convergence, or even cause local convergence or not convergence. Initializing the connection weights closer to the center of each category can highten the speed of convergence.Because text-data-intensive area may contain category center or close to category center , this paper presents a hierarchical clustering method to detect text-data-intensive areas and use the center of the K detected text-data-intensive areas to initialize the connection weight of SOM neural network, in order to improve the speed of SOM neural network convergence. The experimental results showed that: ensuring the effectiveness of text clustering, the text clustering speed is greatly improved.**

Keywords-text clustering; SOM neural network; weight; convergence

I. INTRODUCTION

Since the 90's of 20th century, With the rapid development of network technology, Information has been expanding in high speed; and Information will be growing fast and fast in the future. Now it is very difficult to estimate the amount of information. Among the many information carriers,The text is the most important one. According to statistics, 80% of the information is in the form of text. Relative to other information carriers, the text has been increasing in even more rapid speed. Text Mining has become the most important branch of data mining. It is a rapidly popularizing area of research[1][2]. Text clustering technology is one kind of the most important text mining technologies. Text Clustering can be applied in many fields;for example：It can be applied to organize the web pages that the search engine searched so as the Surfer can find the the text that he needs rapidly; It can be applied in the information retrieval system to improve the efficiency of information retrieval; It can be applied to help users navigate extra-large text databases; and so on.

However, the speed of text clustering will have a direct impact on its applications in various areas. If the speed f text clustering is low, it will not be helpful; Clustering technology is the core of text clustering technology. It plays an important role on the speed of text clustering. SOM (Self-Organizing Feature Mapping Neural Network) is the most commonly used text clustering methods. It is proposed by Professor Kohonen in 1981 [3]-[4], who worked in Finland Helsinki University. The SOM Neural Network was generated by analoging the transmissing process of neural signal in the cerebral cortex . However, the original SOM learning algorithm Randomly initialize connection weights， that will cause in a long wandering around the local minimum, accordingly lower the speed of convergence, or even cause local convergence or not convergence. Initializing the connection weights closer to the center of each category can highten the speed of convergence.Because text-data-intensive area may contain category center or close to category center , this paper presents a hierarchical clustering method to detect text-data-intensive areas and use the center of the K detected text-data-intensive areas to initialize the connection weight of SOM neural network, with an expect to improve the speed of SOM neural network convergence.

II. TEXT CLUSTRING

Text Clustering refers to divide large quantites of texts into some text clusters without any text category information;the text in a same cluster is similar to each other. As the text is not structured data,we need to transform it to structured data which the computer can directly recognize and process, the Structured form must can fully reflect the characteristics of the text itself, and can highlight the difference with other texts. Vector Space Model (abbr is VSM), is the most widely used text expression model [5] currently. It is proposed by the G. Salton in the last century, 60 years. In the vector space model, each text are expressed as a vector. And VSM is successfully applied to the SMART text retrieval system.Transforming texts into vectors, need to go through series of pre-processing step such as sub-word, stemming, removing stop words, lowering dimension.

III. SOM NEURAL NETWORK

SOM neural network is a kind of competitive neural network without supervisor. The network has a series of excellent features such as topology maintaining, probability distribution, visualization and so on [6]-[7] . Now it is widely used in speech recognition, image processing, classification and clustering, combinatorial optimization (such as the TSP problem), data analysis and forecasting, and many other areas of information processing [8]-[9] .SOM neural network is learning without supervisor when used in clustering.

978-0-7695-4258-4/10 $26.00 © 2010 IEEE
DOI 10.1109/MINES.2010.74

Through the network's self-organizing process the SOM neural network find and extract the intrinsic characteristics of input data, form topological map which reflect certain distribution of the input data in the network's output node weight vector space,automatically achieve the clusters of input data.

SOM network has two layers : input layer and competitive layer. SOM network has not have hidden layers. Input layer neurons and competitive layer neurons are fully connected bi-directional.The difference between SOM neural network and basic competitive learning neural network is that competitive layer of basic competitive learning neural network is one-dimensional while the competitive layer of SOM network can be one-dimensional, can be two-dimensional grid, and also can be three-dimensional. Figure 1 is a kind of SOM neural network which competitive layer is two-dimensional.

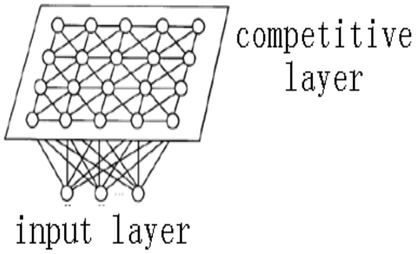

Figure 1. SOM neural network.

For each input vector, only one neuron in competitive layer wins.The most commonly used method to Select the winning neuron is : calculate the Euclidean distance between the input vector and The weight vector of each neuron in output layer;the neuron which takes the minimum Euclidean distance wins. SOM network is based on "winner takes all"principle: the output of the winning neuron is "1"; the output of the remaining neurons is "0".

IV. THE RESEARCH ON HDSOM

A. Initialize the Connection Weights

The initial connection weights have a great influence on the SOM neural network convergence. If the Initial weights are not set appropriate, that will cause in a long wandering around the local minimum, accordingly lower the speed of convergence, or even cause local convergence or not convergence.

The SOM neural network is trainging through adjusting the connection rights, when connection rights change no longer or change very little, the network training is completed, and it has reached convergence state.

B. the Existed Methods of Initializing Connection Weights

1) Randomly initialize connection weights: Generally speaking ,this method is to initialize the network's connection weights with random values in [0,1] . In most cases, input vectors are in the limited direction of the entire space; if the connection weights were randomly initialized, then the connection weight vector will be widely distributed in various random directions, so a large number of connection weight vectors' direction are significantly different to the input vectors' direction, or even opposite to them. When the network training it is very difficult to find the best mapping of input vector, it will take many training times to reach the convergence of the network.This initializing method will cause the network learning time too long or even cause unconvergence of the network.

2) Give all the connection vectors the same initial weight: in the start stage that will reduce the input vector's selection to the connection vectors,then that will increase the winning chance of each connection vector.and then that will reduce the deviation between input vectors and the connection vectors as quickly as possible.This method can highten the speed of convergence; but the convergence speed is still low.

3) According to expert experience, use the probability of a certain word belonging to a certain category to initialize connection weights: Because of high dimensional property of text data, feature selection and feature extraction must be carried out to reduce the dimension of text data before clustering, After feature extraction, a word may be mapped to more than one dimension of input space; this method to determine the initial connection weights becomes very difficult.

*4) Randomly select k input vectors to initialize the connection weights,K is the number of output layer neuron nodes:*Compared to the (1) method to initialize the connection weights, this method to find the best mapping of input vector is relatively easy when the network is training, but because of randomly selecting K input vectors is not necessarily consistent with the category direction, the learning times to achieve network convergence is in large volatility.

C. The new method to determine the initial weights

The ideal distribution of connection weight vectors is that: their directions are consistent with the directions of the categories. But it is unrealistic to do this when initializing.It is the objective that the network training want to achieve. When in network convergence, the directions of connection weight vectors can possibly with the directions of the categories. But when we initialize the connection right, we can try to make the connection weight vectors , direction similar with the categories , so we can try our best to to find out K representative points from the input space. These representative points can represent the various category center;their direction are similar to the categories' direction, at least not too different.

The K selected data points should belong to different categories, and the K data points should be close to the center of the categories.This is the objective we try to achieve when we initialize the connection vector weight. Theory shows that text-data-intensive areas may contain the center of the categories or closer to the center of the categories. This paper presents a hierarchical clustering method to detect text-data-intensive areas, with the detected K data-intensive areas to initialize the connection weights ,with an expect to increase the speed of network convergence.

D. The basic flow of the improved algorithm

*1) Cluster the Nb nearest neighbors of the text (including the text itself) by Hierarchical clustering method using UMPGA so that each text neighborhood formed a cluster tree (Figure 2 below) :*The algorithm select the node of highest score (score = average similarity × number of texts), the node is actually a dense area of documents, and add it to a linked list. In Figure 2 node e will be selected according to the score, which includes the (3,4,5, 6, 7, 8), the intensive area of documents are likely to include the center of the catagory.to the score, which includes the (3,4,5, 6, 7, 8), the intensive area of documents are likely to include the center of the catagory.

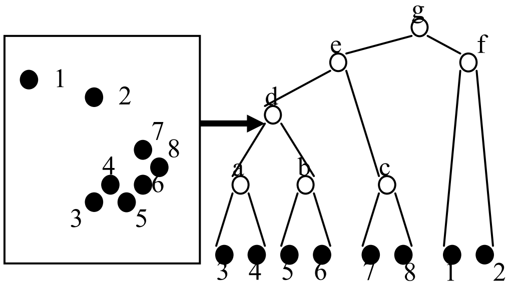

Figure 2. the detection of dense area

2) Sort this list in accordance with these scores.

3) Take the center of these text-densed area: The each dimension weight of the Center vector take the average mapping dimension weight of text datas that in this text-densed area.

4) Assume the number of nodes in the output layer of neuron network is K, take k-data points of the highest score in the lis.

5) Using the K data points to initialize the connection weights of the SOM neural network, while initializing learning rate α_0 , neighborhood radius Nb_0 .

6) Sampling: executing 7) —10) on all the input text data.

7) Determine the winning neuron: Compute the distance between the neuron connection weight vector and the input samples according to the Equation (1),the neuron having the minimum distance wins.

$$d(X_j, Z_i) = \sqrt{\sum_{l=1}^{m} (x_{jl} - \omega_{il})^2} \quad i=1,\ 2,\ \ldots K \qquad (1)$$

Xj—input sample, Zi—neuron in competive layer,Wij—the connection weight between competitive layer neuron Zi and the input layer neuron, m—the number of input neurons, K—the number of output layer neurons.

8) Update: Update connection weights of the winning neuron.at the same time,update connection weights of the neuron in its neighborhood as Equation (2).

$$\omega_i(t+1) = \omega_i(t) + \alpha(t) * (X_j - \omega_i(t)) \qquad (2)$$

i is the winning neuron and its neighborhood neurons.

9) Adjust the learning rate and neighborhood radius: In order to ensure convergence of the algorithm, the initial value of learning rate is set between 0 and 1, and must decreases with the increasing of learning generation，and neighborhood radius is also decreasing with the increasing of learning generation,in the end only the winner neuron is in learning. we can adjust the learning rate and neighborhood radius according to Equation (3) and Equation (4).

$$\alpha(t) = \alpha(0) * (1 - \frac{t}{T}) \qquad (3)$$

$$N(t) = INT(N(0) * (1 - \frac{t}{T}) + 1) \qquad (4)$$

10) Return to Step 6, until the algorithm convergence.

V. Experimental Data And Analysis

This experiment used the Chinese text classification corpus [8] that Tan Song-bo, WANG Yue-fen filed.

For HDSOM algorithm, we assume that when all the network connection weights change less than ξ in average, the neural network is in convergence; this experiment take ξ = 0.002.

The results show that using the method this paper presented to initialize connection weight, the learning generation that the neural network required to reach convergence significantly reduced. And the neural network is not tend to converge to local optimal point.

This ten experiment, the converge generation of original SOM algorithm and HDSOM algorithm are shown in Figure 3, Figure 3 clearly showed a difference between the two algorithms.

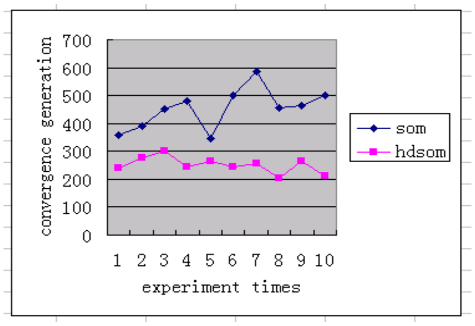

Figure 3. Compare of converge generation.

From Figure 3 we can see, in the traditional SOM algorithm connection weights are initialized randomly, so a large number of connection weights and input vectors vary widely. In order to achieve network convergence, the network must learn many times ,and the times is volatile.

According to that text-data-intensive areas may contain the center of the categories or closer to the center of the categories,the method this paper presented initialized the connection weights with the detected K data-intensive areas ,and then reduced the network convergence times, improved the speed of network convergence.

VI. Conclusion

The paper described an improved SOM Algorithm-HDSOM applied in the text Clustering. The original SOM network randomly initializing connection weights will cause in a long time wandering around the local minimum point, affect the speed of convergence, or even cause local convergence or convergence; HDSOM solved this .It is a relatively excellent text clustering algorithm.

The experiment results show that: ensuring the effectiveness of the text clustering, this method greatly improved the speed of text clustering.

References

[1] Liu Ge Ping, Huang zhi-xing, Li-Xin Li.the study on the Evaluation of e-Learning based on text mining the study [J]. Computer Science, 2005, 32 (5): 170-171.

[2] Jiang Shaohua. Research on text mining applied in Research project Management (PhD Thesis) [D]. Liaoning: Dalian University of Technology, 2006.

[3] KOHONEN T.Self-organized formation of topologically correct feature maps[J].Biological Cybernetics,1982, 4(3):59-69.

[4] Kohonen T.Self-Organization and associative memory[M].3rd ed,Berlin:Springer-Verlag,1989.

[5] Salton G.Automatic Text Processing[M] .Addison-wesley Publishing Company,1988.

[6] KIANG M Y.Extending the kohonen self-organizing map network for clustering analysis[J].Computational Statistics and Data Analysis,2001,38(2):161-180

[7] SIMON H.Neural network principle[M].Beijing:China Machine Press,2004:285-347.

[8] Liang Binmei. improve SOM network applied in stored grain pests Classification [j]. Computer simulation. 2009,26 (10) :202-206.

[9] Zhang Qiyi Kou xue Zhi.the research on intelligent fault diagnosis of internal combustion engine based on SOM neural network[j]. Motorcycle Technology. 2009,38 (4) :49-51.

[10] Tansong Bo, WANG Yue-fen. Chinese text classification corpus - TanCorpV1.0. http://www.searchforum.org.cn/tansongbo/corpus1.php.