The Study of an Improved Text Clustering Algorithm for Self-Organizing Maps

Baolong Zhang\textsuperscript{1,a} and Zemin Hou\textsuperscript{2,b}

\textsuperscript{1}Jiyuan Vocational and Technical College, Jiyuan, China
\textsuperscript{2}Zhengzhou University of Science and Technology, Zhengzhou, China
Email: \textsuperscript{a}115441690@qq.com; \textsuperscript{b}809700541@qq.com

Abstract. The traditional SOM algorithm need to determine the number of clustering categories in advance, which is very subjective. In this paper, an improved k-means initial value selection algorithm is proposed to calculate the number of clustering categories, which is applied to SOM network model. In this algorithm, the Latent Semantic Indexing is applied in the pre-processing stage of clustering, and the improved SOM algorithm is applied in the text clustering stage. Namely, the number of clustering categories obtained by the improved k-means initial value selection algorithm is taken as the number of neurons in the output layer of SOM network.

1. Problem Introduction
In the traditional process of text clustering research, the vector space model VSM is used to represent the text in the text preprocessing stage. Although VSM is simple and efficient, it can only mine the surface match of words and is very sensitive to synonyms and polysemy. For example, “calculator” and “computer” mean the same thing in many semantic contexts, but they don't match in VSM, while the word “virus” has a completely different meaning in many semantic contexts. For example, in the biological domain, virus means bacterial virus in biological sense, while in the computer domain, virus means computer virus program, but in VSM, it is impossible to distinguish them.\textsuperscript{[1]} Therefore, in VSM, the polysemy of the word will cause the search algorithm based on exact match to contain a lot of content that is not what the user is really looking for in the results. One semantic word makes the search algorithm based on exact match leave out a lot of what users are really looking for in the given results.

2. Network Analysis of Potential Semantic Algorithms with (LSI) and Self-Organizing Maps (SOM)

2.1. Latent Semantic Indexing Algorithm
The algorithm makes use of statistical calculation to export the potential semantic relations between words and documents in the text so as to avoid errors in word matching. By bypassing natural language understanding, the potential semantic index finds correlations between different words (phrases, phrases) through statistical analysis of a large sample size, so that the search results are closer to what the user is actually looking for. LSI can better realize dimension reduction through SVD and SDD decomposition of semi-discrete matrix through singular value decomposition, save the time of text clustering, and achieve better clustering effect.
2.2. Self-Organizing Maps Network

In the traditional SOM network model, how to determine the number of neurons in the output layer of the network model is a research problem, because the number of neurons in the output layer is related to the number of patterns in the document set training samples.[2] If not handled properly, the phenomenon of "rough division" and "dead neurons" mentioned earlier in this chapter may easily occur. In the traditional methods, researchers determine the number of clustering categories based on their rich experience and professional knowledge, which is highly subjective and easy to get a wrong number value of clustering categories.

3. The Improved Text Clustering Algorithm Based on Self-Organizing Maps

3.1. Improved Clustering Algorithm K-means

The k-means initial value selection algorithm based on the minimum and maximum principle improved the traditional k-means algorithm to obtain the number of clustering categories, which is then applied to SOM algorithm as the number of neurons in the output layer of SOM network model.

Suppose that the original document should be clustered into k pattern classes, then firstly, two clustering points x1 and x2 with the furthest distance, namely, the smallest cosine similarity, should be selected, and the determination of the remaining clustering points can be deduced by iterative formula.[3] suppose that m poly points have been selected and determined, then the m+1 poly points can be deduced by the following formula:

$$\min\{d(x_{m+1}, x_r), r=1,2,\ldots,m\} = \max\{\min[d(x_j, x_r), r=1,2,\ldots,m]$$,
$$j\neq r\}$$

(1)

In the formula, min and max represent the minimum and maximum distance between the accumulation points, that is, \(\min\{d(x_{m+1}, x_r), r=1,2,\ldots,M\}\) represents the minimum distance of two clustering points, denoted as mindist. According to text clustering rules, small distance between texts within a class means large similarity, while large distance between texts within a class means small similarity. This formula is a recursive iterative formula, which can be used to select the third cluster point. The minimum distance between the third vertex and the first two vertex x1 and x2 is equal to the maximum distance between all the other vertices and the minimum distance from x1 and x2. By analogy, residual points can be determined.

If the number of clustering points is less than the number of real clustering categories, the distance belongs to the distance between texts, so the minimum distance between texts changes smoothly. If the number of clustering points is larger than the number of real clustering categories, the distance belongs to the distance between the texts within the class, so the change of the minimum distance between the texts is relatively stable. However, when the number of clustering points is close to or equal to the number of real clustering categories, there will be a big mutation in the minimum distance change. Based on the above principles, it can be known that the location where the mutation occurs when the minimum distance changes greatly is likely to be the location of the number of real clustering categories.

According to the principles, mechanisms and algorithms mentioned above, an improved algorithm can be obtained. This algorithm mainly calculates the value of the number of clustering categories, and its algorithm flow is as follows:

Input: the number of original documents to be clustered N;
Output: the number value of clustering categories k;

(1) set \(i=1\), L equals the integer part of \(\sqrt{N}\), S0 and S1 as the set of two poly points;
(2) obtain the document set sample similarity matrix, namely, the value of the matrix is the similarity between any two texts. Use the previous formula to calculate the value of similarity between texts;
(3) select the two texts in the similarity matrix with the smallest similarity and the largest distance as the initial clustering points;
(4) \(i=i+1\);
(5) if \(i>L\), then exit;
3.2. The application of improved k-means algorithm in SOM network

The improved k-means initial value selection algorithm can be applied to SOM network model, and the algorithm flow is as follows:

1. The number of clustering categories can be obtained according to the improved k-means initial value selection algorithm, which can be used as the number of neurons in the output layer of SOM network model to construct a SOM network model with two-dimensional planar array structure.

2. To be initialized, the time step is \( n=0,1,2,... \), and the initial value of the weight vector of the output node is \( W_j(0) \); Each element of the vector is randomly evaluated from the interval \((0,1)\). The initial value of learning rate should be larger as far as possible, preferably close to 1; the initial neighborhood radius should also be large, including as many neighboring neurons as possible.

3. For each input vector \( X \) in the training sample set, find the neuron node \( i(X) \) of winning competition.

\[
i(x) = \arg \min_j \| X - W_j \| \quad j = 0,1,...,k
\] (2)

\( K \) is the total number of neurons in the output layer of SOM network;

4. Update the weight vector of the winning neuron node and the node in the neighborhood.

\[
W_j(n+1) = W_j(n) + \eta(n)h_{j,k}(n)(X - W_j(n))
\] (3)

5. Update the learning rate \( \eta(n) \)

\[
\eta(n) = \eta_0 \exp \left( -\frac{n}{\tau_2} \right) \quad n = 0,1,2,...
\] (4)

\( \eta_0 \) is initial value; \( \tau_2 \) is time constant;

6. Update the value of the neighborhood function;

7. Exit when the feature mapping no longer changes significantly or reaches the maximum network training times; Otherwise, let \( n=n+1 \) and turn to step (3).
3.3. The Improved New Som Algorithm Analysis

In the text preprocessing stage, Singular Value Decomposition (SVD) and Semi-Discrete Decomposition (SDD) in latent semantic index are used to decompose and reduce the dimension of word-document matrix, so as to solve the problem of synonym, polysemy and high dimension in text. In the text clustering stage, the improved k-means initial value selection algorithm is used to calculate the number value of clustering categories, which is applied to SOM network model. Therefore, the neuron nodes in the output layer in the new SOM network model are no longer judged by human subjective judgment, but calculated according to the improved k-means initial value selection algorithm.

The new algorithm applies LSI in the pre-processing stage, which can make the search results infinitely close to what the user is really looking for, and achieve better dimensionality reduction. In the clustering stage, k-means initial value selection algorithm is applied to avoid the spatio-temporal overhead caused by multiple iterations in traditional algorithms to determine the number of clustering categories. Therefore, the clustering speed and quality of the new clustering algorithm will be greatly improved, and a good clustering effect will be obtained.

4. The Experiment and Result Analysis of the New Algorithm

Select news stories on 6 topics, each news report containing 50 texts, a total of 300 texts, and then process these news reports into plain text format as the test data set. The six themes are: China-us trade war, college entrance examination reform, environmental governance, food safety, Huawei mobile phones, and the 19th CPC national congress. The details are shown in table1.

| Class number | Subject category                          | Text quantity in the category |
|--------------|------------------------------------------|------------------------------|
| D1           | China-us trade war                       | 50                           |
| D2           | college entrance examination reform      | 50                           |
| D3           | environmental governance,                | 50                           |
| D4           | food safety                              | 50                           |
| D5           | Huawei mobile phones                     | 50                           |
| D6           | the 19th CPC national congress           | 50                           |

F-measure method is adopted to evaluate the effect of clustering algorithm, and the F value in the f-measure method is obtained. The F value and clustering time of the traditional k-means algorithm, SOM algorithm and the new algorithm in this paper are compared, as shown in table2 below.

| The number of categories | The traditional k-means algorithm F value | The traditional SOM algorithm F value | Clustering time (s) | The new SOM algorithm F value | Clustering time (s) |
|--------------------------|------------------------------------------|--------------------------------------|---------------------|-------------------------------|---------------------|
| 3                        | 0.721                                    | 0.741                                | 18.01               | 0.762                         | 10.03               |
| 4                        | 0.754                                    | 0.786                                | 27.04               | 0.827                         | 16.09               |
| 5                        | 0.713                                    | 0.745                                | 37.01               | 0.768                         | 24.08               |
| 6                        | 0.733                                    | 0.761                                | 49.03               | 0.797                         | 31.13               |
It can be clearly seen from table 2 that the F value obtained by using the new SOM clustering algorithm is higher than that obtained by the k-means algorithm and the traditional SOM algorithm. According to the evaluation method of F-measure, the larger the F value is, the better the clustering effect is. Therefore, the new SOM clustering algorithm proposed in this paper is obviously better than the traditional k-means algorithm and SOM algorithm.

5. References

[1] Diao Yufeng, Yang Liang, etc.. Semantic pun detection and pun localization based on potential semantic features [J]. Journal of Chinese Information, 2019, 33(4):12-19.

[2] Rahman M M, Graham P. Responsive and efficient provisioning for multimedia applications[J]. Computers & Electrical Engineering, 2016, 53( C): 458-468.

[3] Zhang Lei, Zhang Yi. Infinite deep neural network method for big data analysis [J]. Computer Research and Development, 2016, 53(1): 68-79