A clustering algorithm based on maximum entropy principle

Yang Zhao[1,2] **and Fangai Liu**[1,2,3]

[1]College of Information Science and Engineering, Shandong Normal University, Jinan Shandong 250358, China;
[2]Shandong Provincial Key Laboratory for Distributed Computer Software Novel Technology, Jinan Shandong 250358, China;
[3]Corresponding author: E- mail: is.zhaoyang@hotmail.com

Abstract. Aiming at the shortcomings of clustering performance of many traditional text clustering methods, a clustering algorithm based on maximum entropy principle is proposed. The algorithm uses the cosine similarity measure cited in the traditional text clustering algorithm SP-Kmeans, and then introduces the maximal entropy theory to construct the maximal entropy objective function suitable for text clustering. The maximum entropy principle is introduced into the spherical K-mean text clustering Algorithm. The experimental results show that compared with DA-VMFS and SP-Kmeans algorithms, in addressing the large number of text clustering problem. The performance of CAMEP clustering algorithm is greatly improved, and has a good overall performance.

1. Introduction

With the growth of the World Wide Web and various text resources, people's desire for rapid, accurate and comprehensive access to information is increasing. Text clustering technology has received more and more attention and research as unsupervised clustering technology. In present, text clustering technology has become the key technology of automatic text categorization [1].

In a certain vector space model, the text can be expressed as a vector of high dimensional space by appropriate preprocessing, which has sparsity and Non-negative [1].In recent years, research shows that the text data also has the direction [2].This feature allows the text vector data to be normalized before clustering, and then the clustering analysis is performed. The SP-Kmeans [3] algorithm uses the cosine similarity to measure the correlation of the text vectors.

In recent years the great entropy principle has also been widely used in natural language processing [7] and text classification [8]. Surian D [9] in pointed out that the text clustering algorithm based on mixed vMF density model movMF in the text clustering process hidden variable entropy changes with self-annealing characteristics, Shi Zhong [10, 11] The deterministic annealing technique is used to improve the clustering performance of the movMF [14] algorithm, which provides the basis for introducing the maximum entropy principle in the traditional text clustering algorithm.

2. Algorithm for Maximal Entropy Clustering Algorithm and Spherical K-means Clustering Algorithm

Using the statistical physical degradation process, Yasuda proposed a deterministic annealing technique [4], which is an important branch of natural law. It is based on the annealing process, the optimal solution of the optimization problem into a series of temperature changes with the physical system of free energy function is minimal. Karayianni [5] Introduced deterministic annealing techniques into clustering. In this algorithm, a very large entropy clustering algorithm is proposed, and its essence is to use the deterministic annealing technique to find the objective function of clustering minimum. In a variety of versions of the maximum entropy clustering algorithm MEC [15], although the description is different, but only the formal differences. The MEC of the maximal entropy clustering algorithm is introduced only in the literature [6].

For the dataset $X=\{x_1, \cdots, x_N\}\subset R^d$, $V=\{v_1, \cdots, v_N\}$ is the K clustering center, $v_i \in R^d, 2\le K<N, U=\{u_{i,j}\}_{k>n}$, $k>n$ is a membership matrix, $u_{i,j}$ for each sample belongs to the center of the probability.

$$\begin{gathered} u_{i,j} \in [0,1], 1 \le j \le n \\ \sum_{i=1}^{K} u_{i,j} = 1 \end{gathered} \tag{1}$$

x_i is divided into K clusters $G_i \in (i=1,2,\cdots, \mathrm{K})$,and the clustering of each cluster is obtained by the maximum entropy fuzzy clustering algorithm MEC Center, the following objective function is minimized.

$$J_T(U,I)=\sum_{i=1}^{K}\sum_{j=1}^{N} u_{i,j} \| x_j - v_i \|^2 + T\sum_{i=1}^{K}\sum_{j=1}^{N} u_{i,j} \ln u_{i,j} \tag{2}$$

Where $\| x_j - v_i \|^2 = (x_j - v_i)^T \times (x_j - v_i)$,T is the Lagrange multiplier. The above equation can also be expressed as:

$$J_T = J_c(U,V) - TH(u) \tag{3}$$

Where $J_C(U,V)=\sum_{i=1}^{K}\sum_{j=1}^{N} u_{i,j} \| x_j - v_i \|^2$.

For a large T, the main attempt is to maximize the entropy H(u), the system is maintained at a higher temperature, with the decrease of T, entropy for the reduction of distortion, when T tends to zero, the minimum Jc (U, V) Directly obtain a non-random (hard) solution. Thus the Lagrangian multiplier there is equivalent to the temperature coefficient of the deterministic annealing technique, also known as the annealing coefficient.

The basic steps of the maximum entropy clustering algorithm MEC are as follows:

Initialization: Given the initial clustering center $V^0=\{v_1^0, v_2^0, v_3^0, \cdots, v_{\mathrm{K}}^0\}$, and the fuzzy partitioning matrix $U=\{u_{i,j}\}$, l is the number of iterations, the maximum number of iterations M, set the annealing coefficient T, the minimum annealing coefficient MinT threshold ε .

Update $u_{i,j}^{i+1}$ with the following formula:

$$u_{i,j} = \frac{\exp(-\frac{\| x_j - v_i \|^2}{T})}{\sum_{h=1}^{K} \exp(-\frac{\| x_j - v_h \|^2}{T})} \tag{4}$$

Update v_i^{i+1} with the following formula:

$$v_i = \frac{\sum_{j=1}^{N} u_{i,j} x_j}{\sum_{j=1}^{N} u_{i,j}} \tag{5}$$

If T is minimized, stop; otherwise adjust the annealing factor T = T-! T to (2).

MEC algorithm can avoid the local minimum and get the global minimum, which has been widely used. However, one of the defects of the MEC algorithm is to use the European metric. For the high-dimensional vector data, the direction feature of the text vector is more important than the size feature, so the MEC is not suitable for clustering the text data.

3. The Clustering Algorithm Based on Maximum Entropy Principle

In this paper, the maximum entropy principle is introduced into the spherical mean clustering, and the clustering algorithm based on maximum entropy principle CAMEP is deduced for text clustering. $V=\{v_1, v_2, \cdots, v_K\}$ is the K clustering center for the sample set $X=\{x_1, \cdots, x_N\}\subset R^d$, $x_i^T x_i = 1(1\le i\le N), V=\{v_1, v_2, \cdots, v_{\mathrm{K}}\}$ and $K\times N$ is a membership matrix. $u_{i,j}$ is the degree of membership of the sample x_j belonging to the K center, the value of which is different from the hard division of the spherical K mean, but the fuzzy division between 0 and 1 truly reflects the data points and the center of the class Practical relationship, and meet $\sum_{i=1}^{K} u_{i,j} = 1$.At this point the global maximum cost function can be considered:

$$J = \sum_{i=1}^{K} \sum_{j=1}^{N} u_{i,j} x_j^T v_i \tag{6}$$

In order to obtain the maximum value of the equation (6), we can get the maximum entropy principle by avoiding the local minimum and get the global minimum. In this case, we can define the minimized objective function:

$$J_T(U,V) = -\sum_{i=1}^{K} \sum_{j=1}^{N} u_{i,j} x_j^T v_i + \frac{1}{T} \sum_{i=1}^{K} \sum_{j=1}^{N} u_{i,j} \ln u_{i,j} \tag{7}$$

Note that the form of Eq. (7) is similar to Equation (2), and entropy terms are introduced. (2) Using the European measure, and (7) uses a cosine similarity measure. Equation (7) can also be expressed as:

$$J_r = J_c(U,V) - \frac{1}{T} H(u) \tag{8}$$

Where $J_c(U,V) = -\sum_{i=1}^{K} \sum_{j=1}^{N} u_{i,j} x_j v_i$,T is a Lagrange multiplier, which can be taken according to the need, and its value has some influence on the final clustering result. $H(u)$ is the entropy of the membership matrix. When the value of $\frac{1}{T}$ is large, minimizing $J_T(U,V)$ actually needs to maximize the entrop $H(u)$ .As the $\frac{1}{T}$ value decreases, the minimized $J_T(U,V)$ turns to the minimized $J_c(U,V)$, thus achieving global minima. The peak of the objective function under the condition of $\sum_{i=1}^{K} u_{i,j} = 1$, the minimum value of $J_T(U,V)$ is obtained. And define the following Lagrangian objective function $L(u,v,\lambda,\gamma)$:

$$L(u,v,\lambda,\gamma) = J_T(U,V) + \lambda \sum_{i=1}^{K} (v_i^T v_i - 1) + \gamma \sum_{j=1}^{N} (\sum_{i=1}^{N} u_{ij} - 1) \tag{9}$$

There is a partial derivative of each center vector v_i in $L(u,v,\lambda,\gamma)$:

$$\frac{\partial L(u,v,\lambda,\gamma)}{\partial v_i} = -\sum_{j=1}^{N} u_{i,j} x_j^T + 2\lambda v_i \tag{10}$$

Let equation (10) be equal to 0, then there is:

$$v_i = \frac{\sum_{j=1}^{N} u_{i,j} x_j}{2\lambda} \tag{11}$$

Since $v_i^T v_i = 1$, by (11) can be further introduced:

$$v_i = \frac{\sum_{j=1}^{N} u_{i,j} x_j}{\sqrt{(\sum_{j=1}^{N} u_{i,j} x_j)^T (\sum_{j=1}^{N} u_{i,j} x_j)}} \tag{12}$$

For $L(u,v,\lambda,\gamma)$ in each $u_{i,j}$ partial guide:

$$\frac{\partial L(u,v,\lambda,\gamma)}{\partial u_{i,j}} = -x_j^T v_i + \frac{1}{T} (\ln u_{i,j} + 1) + r \tag{13}$$

Let equation (13) be equal to 0, then there is:

$$\ln u_{i,j} = Tx_j^T v_i - (T_\gamma + 1) \tag{14}$$

Due to $\sum_i^N u_{i,j} = 1$, by (14) can be further introduced:

$$u_{i,j} = \frac{e^{T^T_{x_j v_i}}}{\sum_{i=1}^{K} e^{T^T_{x_j v_i}}} \tag{15}$$

The minimum process (9) is the clustering algorithm based on maximum entropy principle (CAMEP). It is noted that the Lagrange multiplier T is equivalent to the inverted annealing coefficient, When the T value is small, the system is maintained at a higher temperature, and the process of T increases is the process of system annealing, and the minimum point of the objective function is obtained by a series of changes with temperature T.

A complete description of the CAMEP algorithm is given below.

Step1:Giving $v_0 = \{v_1^0, v_2^0, \cdots, v_K^0\}$, and the fuzzy partitioning matrix $U = \{u_{i,j}\}_{k>n} = 0$,the maximum number of iterations is M, the set annealing coefficient is T, the maximum annealing coefficient is MaxT, the threshold is ε , the number of iterations is r = 0;

Step 2: Find the update $u_{i,j}^{(l+1)}$ according to the formula (15);

Step 3: Update the center $v_i^{(l+1)}$ according to formula (12);

Step 4: If T is equal to the minimum, then stop; otherwise, if $\max_i \| v_i^{(l+1)} - v_i^{(l)} \| \le \varepsilon$ or l> M, adjust the annealing coefficient $T = T - \Delta T$, go to step 2.

4. Experimental Results and Analysis

First, the evaluation criteria of evaluating the performance of text clustering are described, then the various data sets and experimental setups of the experiment are described. Finally, the experimental results of each data set are given and analyzed.

4.1 Algorithm Performance Evaluation Criteria

The performance evaluation criteria of the algorithm based on the objective function are the internal evaluation standard and the external evaluation standard. If vi is the center of the normalized class Ki, the ACS is defined as follows:

$$ACS = \frac{1}{N}\sum_{i=1}^{K} \sum_{x_j \in K_i} x_j^T v_i \tag{16}$$

Where N is the number of samples, and the larger the ACS value, the higher the total tightness of the data and the center vectors. In addition, for text clustering experiments, the text of the class is often known, so the external evaluation criteria of the general use of mutual information (NMI).The NMI value is defined as follows: Assuming that X represents a known text class random variable, and Y represents the class random variable of the clustering result, then:

$$NMI = \frac{I(X:Y)}{\sqrt{H(X)\cdot H(Y)}} \tag{17}$$

Where X (Y) is the mutual information of variables X and Y, H (X) and H (Y) are the entropy of variables X and Y.Because clustering often does not know the number of clusters in advance, the NMI value can be used to evaluate the performance of the algorithm when the number of different clusters is better. The higher the NMI value, the more accurate the clustering result is. The NMI value is 1, marked exactly the same.

4.2 Description of The Experimental Datasets

The experiment uses 20 - Newsgroups data sets and some of the eight datasets from the CLUTO [12] text clustering toolbox. The data set contains the number of samples ranging from 690 to 19949, the smallest data dimension is 8 261 dimension, the largest is 43586 dimension, the actual number of clusters is 3, the largest is 20. From the above characteristics we can see that these data sets reflect the characteristics of the text datasets. Where the NG20 data is averaged from 20 different newsgroups, and the Bow toolkit [13] prepares the 20-

Newsgroups text with 19949 vector text data. NG17-19 is a subset of NG20 data, the actual number of categories for the three categories, each category includes nearly 1 000 from the political news of the text, and according to the characteristics of these news is divided into three categories, the previous clustering algorithm on the The clustering results of the data set show that the clustering of the data set is more difficult because of the overlap between classes and classes. The other data comes from the CLUTO toolbox [12], which has been pre-processed as vector text data. A detailed description of the data set is shown in Table 1.

It should be noted that the balance in Table 1 (n_d is the total number of document n_w is the total number of terms, k is the number of classes) is the balance of the data, that is, the ratio of the number of classes containing the minimum number of texts to the number of texts in the class containing the maximum number of texts, which reflects the balance between the class and the class in the data set. The NG20, NG17-19, and sports data sets used in the experiment are more balanced, ie, the number of samples is similar in each class, and the balance of other data sets is poor.

4.3 Experimental Results and Analysis

In order to test the maximum entropy sphere K-means algorithm proposed by the author, the clustering results of the above data sets are compared with each other when the number of clusters and the number of clusters are different. In the experiment, the algorithm is run 20 times for each case, and the NMI average of its clustering results is taken as the final evaluation value. At the same time, the specific NMI mean and the deviation table of the experimental results and the average cosine similarity of the clustering results degree. Experiments show that the maximum entropy spherical K-means algorithm has achieved satisfactory results for these data sets.

*4.3.1. Comparison of clustering effects of each algorithm when fixed clustering number.*From the clustering results NMI of the different data sets in Table 2 and Table 3, it can be seen that the SP-Kmeans algorithm has the lowest clustering NMI value for each data set, and its clustering effect is obviously lower than the other two clustering algorithms. The clustering effect of using CAMEP is better than that of DA-SPKM.

Table 1. Summary of text datasets

data	Source	n_d	n_w	K	balance
NG20	20 Newsgroups	19949	43586	20	0.991
NG17-19	3 overlapping subgroups from NG20	2998	15810	3	0.998
Reviews	San Jose Mercury (TREC)	4069	18483	5	0.998
Sports	San Jose Mercury (TREC)	8580	14870	7	0.636
Tr54	TREC	690	8261	10	0.088
La1	LA Times(TREC)	3204	31472	6	0.290
La12	LA Times(TREC)	6279	31472	6	0.282
La2	LA Times(TREC)	3705	31472	6	0.274

Table 2. NMI results on NG20, NG17-19, reviews, sports, and tr54 datasets

	NG20	NG17-19	reviews	sports	tr45
K	20	3	5	7	10
SP-KMeans	0.550 ± 0.050	0.340 ± 0.100	0.530 ± 0.100	0.560 ± 0.130	0.600 ± 0.090
DA-VMFS	0.570 ± 0.030	0.460 ± 0.010	0.560 ± 0.090	0.620 ± 0.050	0.680 ± 0.050
CAMEP	0.590 ± 0.010	0.470 ± 0.060	0.620 ± 0.020	0.640 ± 0.004	0.690 ± 0.030

In addition, the authors find that the NMI deviation of the CAMEP clustering algorithm is much smaller than that of the SPK-Means and DA-VMF algorithms in most cases, which means that the algorithm of maximal entropy spherical clustering overcomes the sensitivity to initialization. In the clustering of the difficult data set NG17-19, the author finds that the NMI value of the algorithm CAMEP can reach 0.53, but the NMI value is very large, and its inner reason needs further study.

Table 3. NMI results on classic la1, la12, and la2 datasets

	classic	la1	la12	la2
K	4	6	6	6
SP-KMeans	0.540 ± 0.040	0.480 ± 0.100	0.480 ± 0.100	0.480 ± 0.070
DA-VMFS	0.510 ± 0.010	0.530 ± 0.030	0.520 ± 0.020	0.520 ± 0.040
CAMEP	0.560 ± 0.001	0.560 ± 0.006	0.560 ± 0.001	0.550 ± 0.006

Table 4. Average cosine similarity result on NG20, NG17-19, reviews, sports, tr45 datasets

	NG20	NG17-19	reviews	sports	tr45
K	20	3	5	7	10
SP-KMeans	0.1587	0.1531	0.2131	0.2450	0.3784
DA-VMFS	0.1583	0.1534	0.2140	0.2492	0.3798
CAMEP	0.1602	0.1535	0.2145	0.2476	0.3847

Table 5. Average cosine similarity results on classic la1, la12, and la2 datasets

	classic	la1	la12	la2
K	4	6	6	6
SP-KMeans	0.1505	0.1738	0.1747	0.1803
DA-VMFS	0.1523	0.1730	0.1755	0.1828
CAMEP	0.1527	0.1758	0.1762	0.1868

Table 4 and Table 5 show the average cosine similarity (ACS) of the clustering results of different clustering algorithms for different clustering algorithms. It can be seen from the table that the ACS values of CAMEP and DA-SPKM are greater than those of SP-Kmeans value.

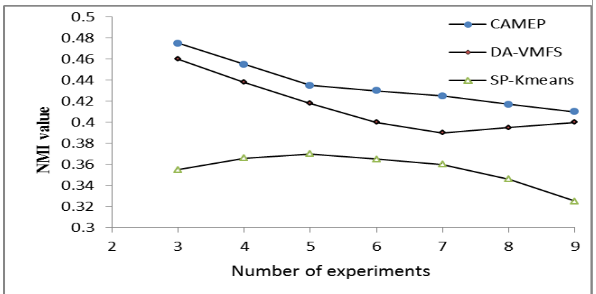

Figure 1. Results on NG17-19 dataset

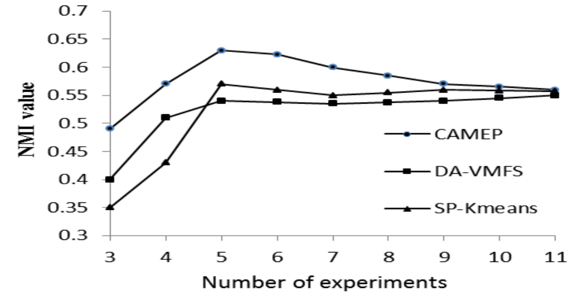

Figure 2. Results on reviews dataset

4.3.2. Comparison of clustering results of different algorithms for different clusters.The clustering algorithm often does not know the actual number of clusters in advance. Therefore, the authors compare the clustering performance of each algorithm in different clustering categories. In order to ensure the accuracy of the experiment, a clustering class running algorithm 20 times, and finally 20 times the average of NMI as the class of NMI value. Figure 1 and Figure 2 are the algorithm for some data sets in a variety of clusters under the NMI value comparison chart, we can see from the figure CAMEP due to the use of the maximum entropy strategy to avoid the local minimum point, so in different poly Class clustering performance is better than SP-Kmeans.

Table 6. Runtime results

	NG20	NG17-19	reviews	sports	classic
K	20	3			
SP-KMeans	84.5	6.4	12.2	25.8	5.1
DA-VMFS	1686.7	50.5	220.1	335.0	96.5
CAMEP	3177.9	55.2	326.3	449.2	125.5

*4.3.3. Comparison of clustering time of each algorithm.*Table 6 shows the clustering time comparison of the partial data sets in the actual clustering number. The clustering time is much smaller than the other two algorithms, Based on the analysis of the above sections, the author thinks that different clustering algorithms can be used for different data clustering tasks.

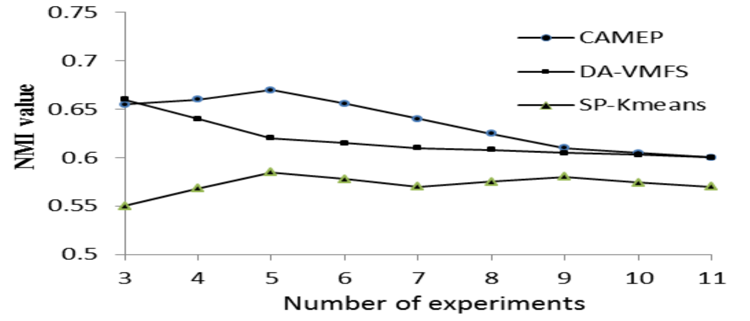

Figure 3. Result on sports dataset

5. Conclusion

In this paper, the maximum entropy principle is applied to the objective function of the spherical K-means algorithm, and the clustering algorithm based on maximum entropy principle is proposed. A large number of experiments show that the algorithm can effectively the clustering performance of text data set is better than that of traditional clustering algorithms. In addition, the author also found some problems: how to improve the clustering effect of CAMEP in the case of high difficulty data cluster, and how to further improve the CAMEP clustering effect reduce its clustering time. The above question is also the author's next research goal.

6. References

[1] Hashimi H, Hafez A, Mathkour H. *Selection criteria for text mining approaches* [M]. Elsevier Science Publishers B. V. 2015.

[2] Jammalamadaka S R, Sengupta A. *Directional statistics* / [J]. 2017.

[3] Dhillon I S, Fan J,Guan Yuqiang. *Efficient clustering of very large document collections*[C]//Data mining for Scientific and Engineering Applications Norwell. MA: Kluwer, 2011.

[4] Yasuda M. *Entropy Maximization and Deterministic Annealing Approach to Fuzzy C-means Clustering* [J]. Scis, 2012, 2010:1515-1520.

[5] Rose K, Gurewitiz E, Fox G A. *Deterministic annealing approach to clustering* [J]. Patter Recognition Letters, 1990, 11:589- 594.

[6] Javed K, Gouriveau R, Zerhouni N. *A New Multivariate Approach for Prognostics Based on Extreme Learning Machine and Fuzzy Clustering* [J]. IEEE Transactions on Cybernetics, 2015, 45(12):2626-2639.

[7] Li R, Tao X, Tang L, et al. *Using Maximum Entropy Model for Chinese Text Categorization*[C]// Asia-Pacific Web Conference. Springer Berlin Heidelberg, 2004:578-587.

[8] Li Ronglu, Wang Jianhui, Chen Xiaoyun, et al. *Using maximum entropy model for Chinese text categorization*[J].Journal of Computer Research and Development, 2005,42(1):94- 101.

[9] Surian D, Chawla S. *Mining Outlier Participants: Insights Using Directional Distributions in Latent Models*[C]// Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, 2013:337-352.

[10] Shi Zhong, Joydeep G J. *A unified framework for model- based clustering* [J]. Journal of Machine Learning Research, 2004, 4(6):1001- 1037.

[11] Shi Zhong, Ghosh J. *Generative model- based document clustering: a comparative study* [J]. Knowledge and Information Systems, 2005, 8(3):374- 384.

[12] ftp://www.cs.umn.edu/~karypis/CLUTO/flies/datasets.tar.gz.

[13] Mow: a toolkit for statistical language modeling, text retrieval, classification and clustering[EB/OL].http://www.cs.cmu.edu/mccallum/bow.

[14] Hornik K, Wu W W, Grün B, et al. *movMF: An R Package for Fitting Mixtures of von Mises-Fisher Distributions [J]*. Journal of Statistical Software, 2014, 058(10):1-31.

[15] Karayiannis N B. *MECA: maximum entropy clustering algorithm[C]//* Fuzzy Systems, 1994. IEEE World Congress on Computational Intelligence. Proceedings of the Third IEEE Conference on. IEEE, 2002:630-635 vol.1.

[16] ZHAO Yang, born in 1992, M. S. candidate. His research interests include data mining, personalized recommendation, etc.

[17] LIU Fangai, born in 1962, Ph. D., professor. His research interests include data mining, personalized recommendation, distributed computing, etc.

[18] This research was financially supported by the National Natural Science Foundation of China (No.61572301, No.90612003), the Natural Science Foundation of Shandong Province (No.ZR2013FM008, No.ZR2016FP07), and the Open Research Fund from Shandong provincial Key Laboratory of Computer Network (No. SDKLCN-2016-01).