A K-means Text Clustering Algorithm Based on Subject Feature Vector

Ji Duo[1,*], Peng Zhang[2] and Liu Hao[1]

[1]*Criminal Investigation Police University of China, China*
[2]*Institute of Information Engineering, Chinese Academy of Sciences, China*
E-mail: jiduo_1@163.com
[*]*Corresponding Author*

Received 18 June 2021; Accepted 20 July 2021;
Publication 14 October 2021

Abstract

As one of the most popular clustering algorithms, k-means is easily influenced by initial points and the number of clusters, besides, the iterative class center calculated by the mean of all points in a cluster is one of the reasons influencing clustering performance. Representational initial points are selected in this paper according to the decision graph composed by local density and distance of each point. Then we propose an improved k-means text clustering algorithm, the iterative class center of the improved algorithm is composed by subject feature vector which can avoid the influence caused by noises. Experiments show that the initial points are selected successfully and the clustering results improve 3%, 5%, 2% and 7% respectively than traditional k-means clustering algorithm on four experimental corpuses of Fudan and Sougou.

Keywords: k-means, initial points, decision graph, iterative class center, subject feature vector.

Journal of Web Engineering, Vol. 20_6, 1935–1946.
doi: 10.13052/jwe1540-9589.20612

1 Introduction

As one of the important study directions in data mining document clustering has been widely used in various fields. For example, documents are clustered in information retrieval systems to enhance performance [1]; genes are clustered in biology to find the relationship between each species and discover new species [2]. In view of the widely use of clustering, many new clustering algorithms are put forward every year to improve the performance of clustering. However, a recent research shows that k-means put forward over half a century ago is still one of the top ten clustering algorithms because of its simplicity [3]. Although k-means is widely used it also has some major limitations [4]. For example, its clustering performance is easily influenced by the clusters' number and the initial points, and the iterative class center during iteration is calculated by the mean of points in each cluster which can affect the finial clustering performance either [5].

For the problem of choosing initial points, the general methods are based on random selection, top K nodes selection, experience, density and so on [6]. Zheng Wei proposes an initial points selection method based on minimum similarity which calculates the similarity between two documents in the dataset firstly and the two documents with minimum similarity are choose to be the first two initial points, then the similarities of the rest documents to the initial points are calculated, the document with minimum similarity is added to the initial points set [7]. Alex Rodriguez proposes a method based on density peaks which selects initial points by using decision graph composed by the local density and distance of each point [8]. It receives good performance in the data set in a certain distribution but fails to find good initial points in data set with random distribution.

For the iterative class center which makes a big impact on clustering performance, the most classic improvement is that Kaufman proposes a k-medoids clustering method which selects the point with minimum distance to others in the cluster as iterative class center [9]. The method which reduces the effect brought by noises is suitable for small data set generally because of its high computational complexity. Inspired by the selection method of the initial points, Pratap proposes a class center selection method based on density, however, the computation is large because that each point needs to calculate its density [10]. Wentian Ji selects the iterative centers by creating a semantic net of web object ontology which improves the clustering performance [11].

For the problem of initial points in this paper, we use local density and cosine similarity instead of distance to select initial points based on the method proposed by Alex Rodriguez since cosine similarity can measure the similarity of two documents much more accurately. Besides, for the iterative class center problem we use subject feature vector as class center to reduce the effect of noises. Experiments show that the initial point selection method operates effectively and the clustering performance improves effectively by using the subject feature vector instead of the mean value as iterative center.

2 Clustering Algorithm Based on Topic Word Vector

Figure 1 illustrates an overview of the clustering algorithm. Initially, preprocessing is applied to the clustered corpus then K initial points are chosen by giving the number of clusters K. The iterative class centers are composed by subject feature vector in this paper compared with the traditional k-means clustering algorithm.

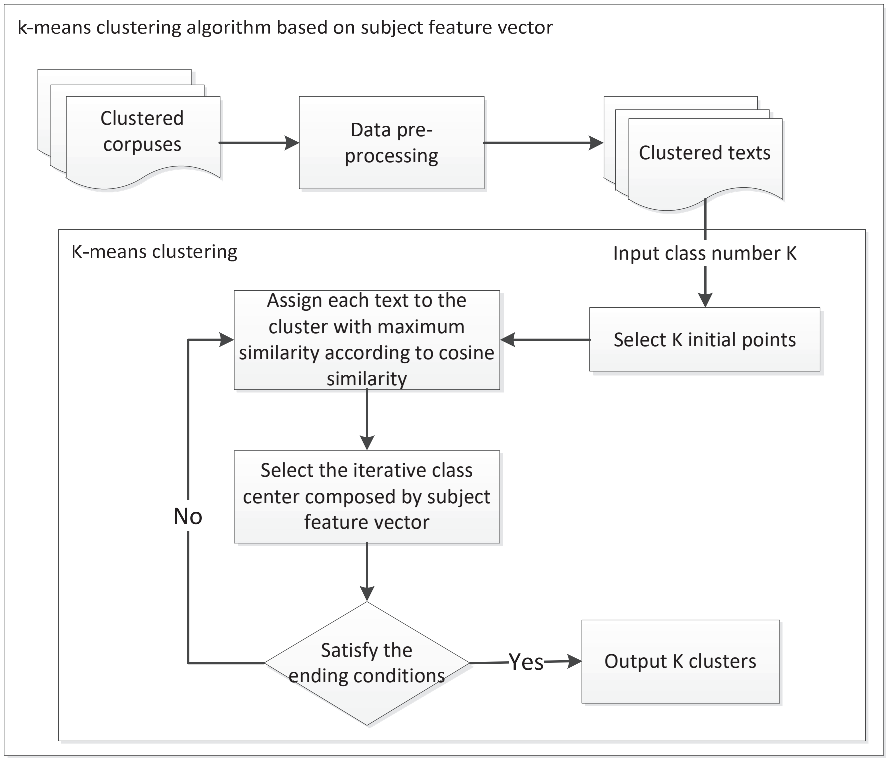

k-means clustering algorithm based on subject feature vector
Clustered corpuses
Data pre-processing
Clustered texts
K-means clustering
Input class number K
Assign each text to the cluster with maximum similarity according to cosine similarity
Select K initial points
Select the iterative class center composed by subject feature vector
No
Satisfy the ending conditions
Yes
Output K clusters

Figure 1 k-means clustering algorithm based on subject feature vector.

2.1 Initial Points' Selection

The selection of initial points makes a great influence on the clustering result of k-means clustering algorithm. It will reduce the performance of clustering if the initial points are noises or crowded so we try to select better initial points based on the method proposed by Alex Rodriguez.

Alex Rodriguez presumes that cluster centers are surrounded by neighbors with lower local density and that they are at a relatively large distance from any points with higher local density [8]. For each point i we need to compute two quantities: the local density ρ_i and the distance δ_i from any point of higher density. Both the two quantities depend only on the distance d_{ij}. The local density is defined as follows.

$$\rho i = \sum_j \chi(dij - dc) \quad \begin{Bmatrix} \chi(x) = 1, & x \leq 0 \\ \chi(x) = 0, & others \end{Bmatrix} \tag{1}$$

d_{ij} is the cutoff of distance, that is to say, ρ_i is equal the number of points that are closer than d_c to point i. As a role of thumb, we choose d_c so that the average number of neighbors is around 1 to 2% of the total number of points in the data set [8].

The distance is measured by computing the minimum distance between the point i to any point with higher density. And δ_i is defined as follows.

$$\delta_i = \min_{j:\rho j > \rho i}(dij) \tag{2}$$

What's more, for the point with highest density, δ_i which different from other points is defined as $\delta_i = \min(dij)_{j:pj > pi}$. Thus, clusters are recognized as points for which the δ_i is large and the ρ_i is in a reasonable range. Example is shown in Figure 2.

Figure 2 shows the decision graph, the abscissa means the local density and ordinate means the distance. We choose the points with large distance and reasonable local distance as initial points generally according to the decision graph.

In Figure 2, the red rectangle points with large distance and reasonable local density are selected as initial points, the green triangle points are recognized as noises because their local density are low and their distance are large.

We select the initial cluster centers accurately according to the decision graph. However, in practice we choose cosine similarity instead of distance to select initial centers because the cosine similarity can measure the similarity

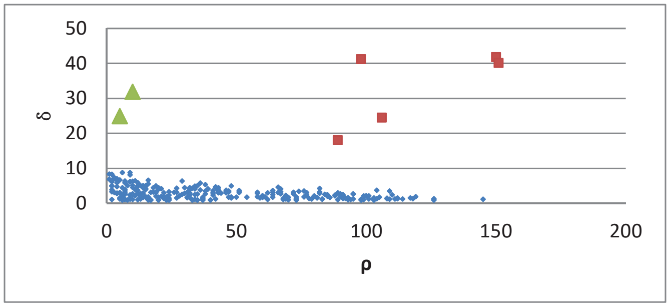


Figure 2 Decision graph.

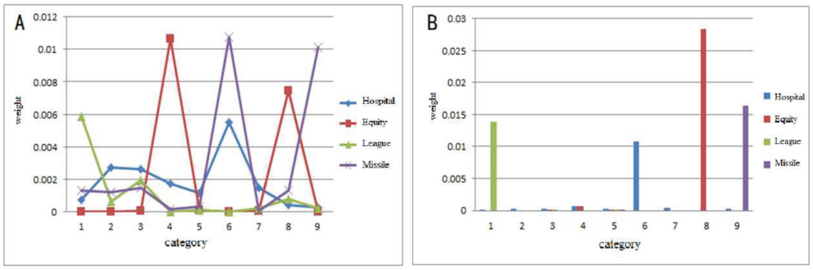


Figure 3 Distribution and weight.

of two documents much more accurately. Thus the minimum distance corresponds to the maximum similarity and the maximum distance means the minimum similarity and in decision graph the ordinate is the reciprocal of similarity.

2.2 Subject Feature Vector Selection

The iterative class centers of traditional k-means clustering algorithm are calculated by the mean of all the points in clusters which can be influenced by the noises and noise features for the reason that every point in the data set and every feature will participate in the calculation. We choose four key features belonging to different categories, and they are “Hospital”, “Equity”, “League” and “Missile”. Their distribution and weight in the start and end of iteration are shown in Figure 3.

Figure 3A shows the distribution and weight of four key features at the start of iteration. From the graph we know that the four key features are distributed in all nine categories with low weight. Figure 3B shows the two quantities at the end of iteration. Each feature is assigned to one category with large weight. Therefore, the words with large distinguishability should be retained during clustering.

In order to obtain key features with large distinguishability and remove the noise features we use subject feature vector instead of mean as iterative center which can not only reduce the length of center greatly but also avoid the influence of noise points and noise features.

2.2.1 Candidate feature set selection

The words in the candidate feature set have the opportunity to form the subject feature vector. Therefore, it's very important to select the candidate feature set. In order to select the candidate key feature set qualitatively we need to filtrate the word set of the clustered corpus. Four feature selection methods are used to remove the meaningless words form all the documents [5].

(1) All single-character words were deleted since they are virtually used as prepositions.
(2) All adjectives and pronouns were removed based on the POS information.
(3) A Zipfs law-based [12] eliminator was used to remove terms that appear less than t_1 times and that occur in over t_2 of the documents in the corpus as they are used in too many topics to effectively discriminate between topics.
(4) The words which term frequency in a document less than t_3 are also removed. Because all such words are useless for document clustering and cannot be employed to express the subject of the documents.

As a rule of thumb, t_1 equals to 30, t_2 equals to 30% and t_3 equals to 2.

2.2.2 Subject feature vector selection

We need to select the subject feature vector during iteration after finishing selecting the candidate keywords set. We use the ATF-PDF proposed by Yang Jie which considers the average frequency of words in the whole corpus as well as the document frequency of words to extracts subject word [13]. The ATF-PDF is defined as follows.

$$w_i = \frac{\sum_{j=1}^{N} |tf_{ji}|}{N} e^{\frac{ni}{N}} \tag{3}$$

$$|tf_{ji}| = \frac{tf_{ji}}{\sqrt{\sum_{i=1}^{n} tf^2 ji}} \tag{4}$$

N is the quantity of documents in each cluster, n_i is the amount of document which includes word w_i in each cluster, n is the length of all the words in document j and tf_i is the frequency of word i in document j of the cluster.

The words with large weight and belonged to the candidate feature set are selected to form the subject key feature vector according to the weight of each word in cluster C_k. The length of the subject key feature vector is chosen to be 10∼20% of the length of the candidate key feature set generally. Then keep iterating with the subject word vector as iterative center until reaching convergence condition or iteration times.

3 Experimental Results and Analysis

3.1 Experimental Corpuses

Two kinds of corpus are selected as experimental corpuses and they are Fudan and Sougou classification corpus which involves various fields of entertainment, military, politics, education, sport etc. The two corpuses are divided into four corpuses, including Fudan corpus is divided into corpus a with the number of documents is 200 each category and corpus b with the number ranges from 200 to 505, and Sougou corpus is divided into corpus c with the number of documents is 300 each category and corpus d with the number ranges from 180 to 360. The details are shown in Chart 1.

Chart 1 Fudan and Sougou corpus

Fudan Corpus				Sougou corpus			
Corpus a		Corpus b		Corpus c		Corpus d	
Categories	Amount	Categories	Amount	Categories	Amount	Categories	Amount
Environment	200	Environment	200	Military	300	Military	180
IT	200	IT	200	IT	300	IT	237
Traffic	200	Traffic	214	Healthy	300	Healthy	342
Education	200	Education	220	Education	300	Education	281
Economics	200	Economics	325	Sport	300	Sport	270
Military	200	Military	249	Recruitment	300	Recruitment	315
Sport	200	Sport	450	Culture	300	Culture	360
Medicine	200	Medicine	204	Traveling	300	Traveling	325
Art	200	Art	248	Finance	300	Finance	290
Politics	200	Politics	505				

3.2 Experiment Introduction

3.2.1 Initial points selection experiment

Marking the initial points of four corpuses selected based on the method recommended above with the right classes to evaluate the effect of initial point selection. The quantity is category coverage rate r which equals to the ratio between the number of categories of initial points and the total number of categories.

3.2.2 Clustering results comparison experiment

Comparing the clustering results brought by the method proposed in this paper and the traditional k-means algorithm with the same initial points selected in Section 3.2.1. The evaluation quantity is F-value defined as follows.

$$F = \sum_{i}^{m} \frac{mi}{m} \max\{F(i,j)\} \tag{5}$$

$$p(i,j) = \frac{nij}{nj} \quad R(i,j) = \frac{nij}{ni} \quad F(i,j) = \frac{2 \times P(i,j) \times R(i,j)}{P(i,j) + R(i,j)} \tag{6}$$

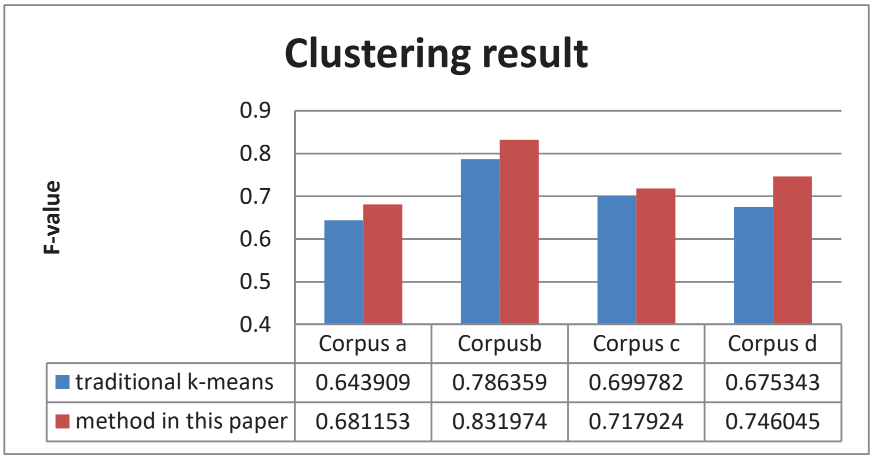
From Equation (6) we know that the F-value of a clustering result C is the maximum F-value of all the categories in the standard clusters got in the clustering result C. $F(i,j)$ is defined as Equation (7), $P(i,j)$ presents the accuracy of the elements in the right category L_i got in the clustered category C_j, $R(i,j)$ is the recall of the elements in L_i got in C_j. n_{ij} is the number of elements divided into C_j form L_i, that is $L_i \cap C_j = n_{ij}$. n_j is the number of the elements of cluster j in clustered results, that is, $C_j = n_j$, n_i is the amount of the elements of cluster i in standard clusters, $L_i = n_i$.

3.3 Experiment Results

The clustering results are shown in Figure 4.

It is shown form Figure 4 that the subject feature vector based k-means clustering algorithm improves 3%, 5%, 2% and 7% respectively on four experimental corpuses than traditional k-means algorithm.

The reason of getting better cluster performance is that the iterative centers composed by subject feature vector can avoid the offset effect caused by noises and noise features. Besides, the length of the center composed by subject feature vector is 10%∼20% of the length of candidate subject feature set which is much less than the length of class center calculated by the mean



	Corpus a	Corpusb	Corpus c	Corpus d
🟦 traditional k-means	0.643909	0.786359	0.699782	0.675343
🟥 method in this paper	0.681153	0.831974	0.717924	0.746045

Figure 4 Clustering results comparison graph.

of all points in cluster. The running speed of the method in this paper is much faster than traditional k-means, and the average running time of the subject feature vector based k-means clustering algorithm is 37.65s compared with 47.2333s of traditional k-means on four experimental corpuses.

4 Conclusions

The quality of initial points has great influence on clustering effect, we selected the initial points successfully based on the method proposed in reference [8] which will lead to good clustering results. Besides, we propose a subject feature vector based k-means clustering algorithm while the tradition class centers calculated by the mean of all points in clusters are easily influenced by noises and noise features. The class center in our method is composed by subject feature vector which will avoid the influence caused by noises and the length of it is 10$\sim$20% of the length of candidate subject feature set leading to faster running speed. Experiments show that the clustering performance of our method gets great improvement compared with traditional k-means as well as running speed.

The subject words used in this paper can be effectively applied to text data, but this method can also be used as a reference for other types of data. As long as the features of the data are discrete and have obvious classification discrimination, the method in this paper can optimize the k-means clustering effect.

References

[1] Sahami M. Using machine learning to improve information access[D]. stanford university, 1998.

[2] Baldi P, Hatfield G W. DNA microarrays and gene expression: from experiments to data analysis and modeling[M]. Cambridge University Press, 2002.

[3] Rao G N, Madhavi D. An Efficient Document Clustering Mechanism in N-dimensional space[J]. 2012.

[4] Lloyd S. Least squares quantization in PCM[J]. Information Theory, IEEE Transactions on, 1982, 28(2): 129–137.

[5] Chang H C, Chiun-Chieh H S U. Using topic keyword clusters for automatic document clustering[J]. IEICE Transactions on Information and Systems, 2005, 88(8): 1852–1860.

[6] Sun Jigui, Liu Jie, Zhao Lianyu. Research on clustering algorithm [J]. Journal of software, 2008, 19(1): 48–61.

[7] Zheng Wei. Research on text clustering technology based on latent semantic index [D]. Shenyang Institute of Aeronautical Technology, 2009.

[8] Rodriguez A, Laio A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492–1496.

[9] Kaufman L, Rousseeuw P. Clustering by means of medoids[M]. North-Holland, 1987.

[10] Pratap R, Devi J R, Vani K S, et al. An Efficient Density based Improved K-Medoids Clustering algorithm[J]. IJACSA) International Journal of Advanced Computer Science and Applications, 2011, 2(6).

[11] Ji W, Guo Q, Zhong S, et al. Improved K-medoids Clustering Algorithm under Semantic Web[C]//Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering. Atlantis Press, 2013.

[12] Zipf G K. Human behavior and the principle of least effort[J]. 1949.

[13] Yang Jie, Ji Duo, Cai Dongfeng. Multi document keyword extraction technology based on joint weight [J]. Journal of Chinese information, 2008, 22(6): 75–79.

Biographies

Ji Duo received his M.En degree from Northeast University. He is an associate professor in department of cyber crime investigation, Criminal Investigation Police University of China. His research direction mainly includes text mining, machine translation, network public opinion analysis, etc. Relevant research results have been published in more than 20 academic journals and conferences at home and abroad, and won the first prize of Liaoning science and technology progress award, and First Prize of Aviation Science and Technology Progress Award of China Aviation Society.

Peng Zhang received his PhD degree from Institute of Computing Technology, Chinese Academy of Sciences. He is an associate professor in Institute of Information Engineering, Chinese Academy of Sciences. His research direction mainly includes social computing and data mining, etc. Relevant research results have been published in more than 60 academic journals and conferences at home and abroad, and is the Member of Youth Innovation Promotion Association of Chinese Academy of Sciences.

Liu Hao received his M.En degree from Criminal Investigation Police University of China. He is a senior experimentalist in department of network information center, Criminal Investigation Police University of China. His research direction mainly includes smart campus construction, campus information, etc.