K-means Clustering Algorithm with improved Initial Center

Zhang Chen
School of Computer Science and Technology
China University of Mining and Technology
Xuzhou ,China
e-mail: zc@cumt.edu.cn

Xia Shixiong
School of Computer Science and Technology
China University of Mining and Technology
Xuzhou ,China
e-mail: xiasx@cumt.edu.cn

***Abstract*—In this paper we present a new clustering method based on k-means that have avoided alternative randomness of initial center. This paper focused on K-means algorithm to the initial value of the dependence of k selected from the aspects of the algorithm is improved. First,the initial clustering number is $\sqrt{N}$ .Second, through the application of the sub-merger strategy the categories were combined.The algorithm does not require the user is given in advance the number of cluster.Experiments on synthetic datasets are presented to have shown significant improvements in clustering accuracy in comparison with the random k-means.**

Keywords- data clustering; k-means; initial center

I. Introduction

Data mining is the analysis of data and the use of software techniques for finding patterns and regularities in sets of data. It is an emerging cross-disciplinary from the collection of machine learning, pattern recognition, databases, statistics, artificial intelligence and other areas of research.Clustering has been one of the most widely studied topics in data mining and k-means clustering has been one of the popular clustering algorithms for partitioning data. Unfortunately, K-means suffers from the well-known problem of locally optimal solutions. Furthermore, the final partition is dependent upon the initial configuration,making the choice of starting partitions all the more important.

In the rest of the paper, we will first present a brief overview of K-Means Limitations, then we will describe the proposed algorithm, and finally we will illustrate implementation details and experimental results.

II. K-Means Limitations

The k-means algorithm is an algorithm to cluster n objects based on attributes into k partitions, k < n. It is similar to the expectation-maximization algorithm for mixtures of Gaussians in that they both attempt to find the centers of natural clusters in the data. It assumes that the object attributes form a vector space. The objective it tries to achieve is to minimize total intra-cluster variance, or, the squared error function

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2$$

where there are k clusters S_i , i = 1, 2, ..., k, and μ_i is the centroid or mean point of all the points $S_j \in S_i$.

The k-means clustering was invented in 1956[1]. The most common form of the algorithm uses an iterative refinement heuristic known as Lloyd's algorithm[2]. Lloyd's algorithm starts by partitioning the input points into k initial sets, either at random or using some heuristic data. It then calculates the mean point, or centroid, of each set. It constructs a new partition by associating each point with the closest centroid. Then the centroids are recalculated for the new clusters, and algorithm repeated by alternate application of these two steps until convergence, which is obtained when the points no longer switch clusters. Approximate k-means algorithms have been designed that make use of coresets: small subsets of the original data.Other variations exist[4], but Lloyd's algorithm has remained popular because it converges extremely quickly in practice. In fact, many have observed that the number of iterations is typically much less than the number of points. However, David Arthur and Sergei Vassilvitskii showed that there exist certain point sets on which k-means takes superpolynomial time[5].

In terms of performance the algorithm is not guaranteed to return a global optimum. The quality of the final solution depends largely on the initial set of clusters, and may, in practice, be much poorer than the global optimum.[citation needed] Since the algorithm is extremely fast, a common method is to run the algorithm several times and return the best clustering found.

A drawback of the k-means algorithm is that the number of clusters k is an input parameter. An inappropriate choice of k may yield poor results. The algorithm also assumes that the variance is an appropriate measure of cluster scatter. In 2006 ,a new way of choosing the initial centers was proposed [3], dubbed "k-means++". The idea is to select centers in a way that they are already initially close to large quantities of points. The authors use L2 norm in selecting the centers, but general Ln may be used to tune the aggressiveness of the seeding.

III. Algorithm Description

This section describes k-means clustering algorithm that have been improved initial center. Initially, we describe the main ideas behind the algorithm. Then,we give some formal definitions,present and explain the pseudo-code and explain some of the choices we have made in our current implementation.

978-0-7695-3543-2/09 $25.00 © 2009 IEEE
DOI 10.1109/WKDD.2009.210

A. *Main ideas*

K for the initial selection to improve the lot of scholars and experts have conducted an in-depth study. The literature [6] use of the ant colony algorithm to deal with strong local maximum ability to dynamically determine the number of cluster and center. The literature [7] through reduced, based on the density and distribution center of the initial search, and other methods to improve K-means algorithm. The literature [8] proved that the author of several cluster on the border of the maximum, In this paper, which first set up the initial cluster number n , $n = \sqrt{N}$. That is, the initial divided into categories to find samples of the space center of mass, calculated Xindao quality of all samples from the point of maximum maxdis, in order to quality for the heart center, with a radius of maxdis to make a round, all samples will be surrounded points in the circle . Will be divided into n , at which point a sample, the sample can be attributed to such points. All samples are divided into their own category, a total of n categories, each category in order to derive the center of mass, such as the initial cluster centers, a total of n initial cluster centers.

Merger: The above polymerization, can be a category in which to continue to air after iteration of the number of categories, we might set up as well, and then through the merger of a number of criteria for the current class of a merger in order to get a final category. That is[9]: if the first class of n that exist in any sample, the sample made with the first i-class center of mass is less than the distance between the first i-class center of mass of all samples with the point of maximum euclidean distance, so that In the first category and a category i can be combined.

B. *Formal definitions*

Input: N object contains a data set.

Output:k clusteres, making the least square error criterion function.

Steps:

(1) Normalized process

①Create a new set of samples instan (used to store normalized after the sample set)

②To be followed by each of the samples in each of the attributes of value, and its normalized

③Each sample will be the first normalized after the property value as the abscissa, each of the other samples normalized after the property's value and as ordinate these values into an array of X and Y

④Will be every sample of each of the normalized values into the treatment of instan

(2) find the essence of the heart

①All samples of each attribute value to find all the mass

②Create a sample, for the initialization of the attributes of each of the above center of mass

③For each attribute the center of mass for normalized

④Will be the first property as a center of mass abscissa and the rest of the property's center of mass, and as ordinate

⑤All samples of the center of mass (abscissa and ordinate)

(3) find samples of the largest radius

The distance between two points using the formula, based on an array of X and Y to find the greatest value from the

(4) All samples assigned to their respective range

①The creation of Boolean variable acc1 and acc2, respectively, on behalf of the samples in a straight line between the two

②Sample round will be divided into sub-000 m interval, followed by all samples to determine

③All samples will be assigned to the host range

(5) The establishment of the initial cluster centers

(6) Repeat

```
Repeat
    do $i = 1$
    do $\forall X_i, i \in \{1,2,\cdots,N\}$
    do $X_i \rightarrow W_j, W_j \in C_j, j \in \{1,2,\cdots, N\}$
    do $i = i++$
    Until $i = N + 1$
    End
    Repeat
        do $j = 1$
        do update $C$
        do $j = j++$
    Until $j = k + 1$
    End
Until $E = \sum_{j=1}^{k} \sum_{p \in C_j} |p - m_j|^2$
End
```

(7) Class merge

If the first class of n that exist in any sample, the sample made with the first i-class center of mass is less than the distance between the first i-class center of mass of all samples with the point of maximum euclidean distance, so that In the first category and a category i can be combined.

IV. Implementation and Experiments

We have implemented the improved K-means clustering algorithm in Java and conducted experiments with synthetic data. In these experiments, we compared improved K-means with two other k-means algorithms with different initial cluster center selection methods: random k-means using the simple initial cluster center selection method, and improved k-means algorithm.

To test the robustness of the algorithm,We download the UCI benchmark data sets Iris.Fig.1, 2 show the clustering results from the random k-means and the improved K-means respectively. We can observe from the figures that the initial cluster centers selected by the NK-means are very close to the inherent cluster centers in the data. Some of the initial cluster centers selected by the random k-means, the refinement k-means were located outside of some inherent clusters. For example, no initial cluster centers were selected from the two middle inherent clusters in Fig. 1. Because of this, the two inherent clusters were clustered into one cluster by the random k-means.

Four initial cluster centers were selected from the large inherent cluster on the upright corner of Figure 1. This cluster was clustered into 6 clusters. Therefore, the random k-means could not recover the eight inherent clusters because of the bad selection of the initial cluster centers. Figure 2, we can see that all eight inherent clusters were completely recovered by improved K-means, due to the good selection of the initial cluster centers.

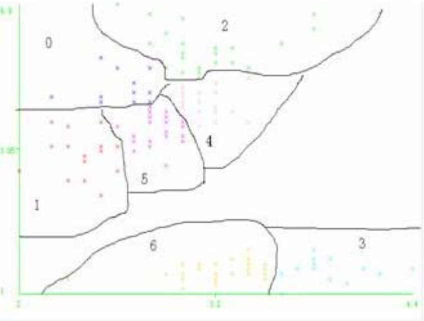

Figure 1. Results of random k-means

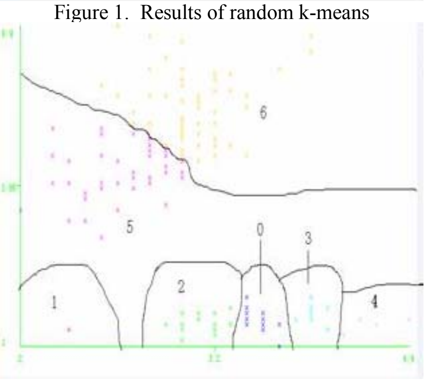

Figure 2. Results of Improved K-means

V. Conclusion

In this paper,we have proposed a new neighborhood density method for selecting initial cluster centers for k-means clustering. We have presented the improved K-means algorithm that select initial cluster centers and use the centers as input to the k-means clustering algorithm to improve the clustering performance of k-means. We have shown the experiments on synthetic data sets to demonstrate that improved K-means was superior to the random k-means.

However, this paper to improve the K-means algorithm for large-scale data set of applications is still a lot of time consumption, including further work to continue to improve for large-scale data sets of serial algorithms for parallel processing In order to reduce the K value generated in the process of traversing the number of data sets to increase data processing speed.

References

[1] H. Steinhaus. Sur la division des corp materiels en parties. Bull. Acad. Polon. Sci., C1. III vol IV:801– 804, 1956.

[2] S. Lloyd, Last square quantization in PCM's. Bell Telephone Laboratories Paper (1957). Published in journal much later: S. P. Lloyd. Least squares quantization in PCM. Special issue on quantization, IEEE Trans. Inform. Theory, 28:129–137, 1982.

[3] D. Arthur, S. Vassilvitskii: "k-means++ The Advantages of Careful Seeding" 2007 Symposium on Discrete Algorithms (SODA).

[4] An efficient k-means clustering algorithm: Analysis and implementation, T. Kanungo, D. M. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Y. Wu, IEEE Trans. Pattern Analysis and Machine Intelligence, 24 (2002), 881-892.

[5] David Arthur & Sergei Vassilvitskii (2006). "How Slow is the k-means Method?". Proceedings of the 2006 Symposium on Computational Geometry (SoCG).

[6] HAND, D.J., and KRZANOWSKI, W.J. (2005), "Optimising k-means Clustering Results with Standard Software Packages", Computational Statisticsand DataAnalysis,49,969–973.

[7] HANSEN,P.,NGAI,E.,CHEUNG,B.K.,and MLADENOVIC,N.(2005),"Analysis of GlobalK-means, An Incremental Heuristic for Minimum Sum-of-squares Clustering", Journal of Classification,22,287–310.

[8] Malay K Pakhiraa, Sangham itra Bandyopadhyayb, UjjwalMaulikc. Validity index for crisp and fuzzy clusters[J].Pattern Recognition, 2004, 37: 487−501.

[9] Yunming Ye.Neighborhood Density Method for Selecting Initial Cluster Centers in K-Mean Clustering In: Proceedings of PAKDD'06. (2006) pp. 189–198.