A Modified K-means Algorithms - Bi-Level K-Means Algorithm

Shyr-Shen Yu
Professor
Department of Computer Science and Engineering,
National Chung Hsing University
Taiwan, R.O.C
pyu@nchu.edu.tw

Shao-Wei Chu
Phd Student
Department of Computer Science and Engineering,
National Chung Hsing University
Taiwan, R.O.C
multi.summer@gmail.com

Ching-Lin Wang
Associate Professor
Department of Information Management,
National Chin-Yi University of Technology
Taiwan, R.O.C
clwang@ncut.edu.tw

Yung-Kuan Chan*
Professor
Department of Management Information Systems,
National Chung Hsing University
Taiwan, R.O.C
ykchan@nchu.edu.tw

Chia -Yi Chuang
Phd Student
Department of Computer Science and Engineering,
National Chung Hsing University
Taiwan, R.O.C
s9356056@cs.nchu.edu.tw

***Abstract*—In this paper, a modified K-means algorithm is proposed to categorize a set of data into smaller clusters. K-means algorithm is a simple and easy clustering method which can efficiently separate a huge number of continuous numerical data with high-dimensions. Moreover, the data in each cluster are similar to one another. However, it is vulnerable to outliers and noisy data, and it spends much executive time in partitioning data too. Noisy data, outliers, and the data with quite different values in one cluster may reduce the accuracy rate of data clustering since the cluster center cannot precisely describe the data in the cluster. In this paper, a bi-level K-means algorithm is hence provided to solve the problems mentioned above. The bi-level K-means algorithm can give an expressive experimental results.**

Keywords- Clustering; K-means algorithm; Classification; Genetic algorithm

I. INTRODUCTION

Data clustering [1] is widely applied in various fields such as pattern recognition [2, 3], image processing [4, 5], data mining [6-8] and data compression [4, 9-13]. The main purpose of data clustering divides a data set into some disjoint subsets (clusters) so that the data within the same cluster are highly similar but the data in different clusters are very distinct. Outliers are the data that deviate significantly from the rest of the data. In this paper, we refer to noisy data and outliers as “abnormal” data and others as “normal” data. The classifying results obtained by a good clustering algorithm should not be intervened by abnormal data. In additional, time complexity and clustering accuracy rate are other evaluating ways, especially in processing a great deal of data.

Hierarchical clustering and partition clustering are two most commonly used types of clustering methods [6]. The data set will be processed by agglomerative and divisive approach to combine the clusters with high similarity into larger ones, or to divide larger clusters into smaller ones which can meet supposed condition. Additionally, tree structure may be used to represent the relationship of clusters. However, it is appropriate to deal with categorical data and highly discrepant data among clusters in a data set; on the contrary, not to process continuous numerical data and high-dimensional data because it needs to spend much execution time and memory space.

Partition clustering type [9, 14, 15] assigns each datum X into one C of non-intersection clusters, where the similarity between X and the cluster center of C is higher than the similarity between C and the cluster centers of other clusters. K-means algorithm is the most commonly used clustering method in data partition. A traditional K-means algorithm [16] can be described as follows:

K-Means Algorithm(*S*, *K*)

Input: S is a data set and K is the numbers of clusters that users desire.

Output: K disjoints clusters.

(1) Randomly select K data to represent the cluster centers of the K clusters from S.

(2) Assign each datum X to the cluster with the minimal distance between X and each cluster center.

(3) Recalculate each cluster center from the data in the same cluster.

(4) If the cluster centers obtained in step 3 are the same as those obtained in previous iteration, then output the clustering results; otherwise go to step 2.

The advantages of K-means algorithm are simple, easy to use, and efficiency to execute. It is suitable to process a huge amount of high dimensional and continuous numerical data. Additionally, the data assigned to the same cluster are highly similar. However, the drawbacks of the K-means algorithm are in the following [17]:

(1) The users need to predetermine the number of clusters in the data set, but it is often difficult to decide the appropriate *K*.

(2) The clustering results are often affected by the cluster centers assigned in initial step.

(3) It is not suitable for dealing with categorical data since describing dimensions by value is difficult.

(4) Clustering results are significantly impacted by abnormal date.

(5) The distances between data and cluster centers may be influenced by the measurement unit of data features. Hence they need to be normalized before clustering.

(6) The weights of data features are considerable because the influences of dimensional features are generally different for computing the distances between each datum and cluster centers.

(7) It is time consuming; the time complexity is $\boldsymbol{O(m \times K \times ITER)}$.

In this paper, the bi-level *K*-means algorithm is proposed to improve above problems. The bi-level *K*-means algorithm can give better clustering results than the traditional *K*-means algorithm. Also, it is indifferent to abnormal data and can accelerate the clustering speed.

Generally, in data clustering, the data with huge discrepancy should be classified to different clusters. Additionally, a cluster containing an enormous number of data is often required to be divided into smaller ones. Hence, in this paper, a bi-level *K*-means algorithm is proposed to cure these problems mentioned above. The bi-level *K*-means algorithm will separate the data with huge discrepancy into different big clusters and also divide the big clusters with a great number of data into smaller ones. This method is also insensitive to abnormal data and can accelerate clustering speed. It can hence classify data more efficiently than the traditional K-means algorithm.

In pattern recognition, a set *S* of historic data with *D* dimensions is collected in advance. Let ($x_{ci1}, x_{ci2}, \ldots, x_{ciD}$) be the feature values of the *i*-th data X_{ci} of the *c*-th cluster *C* in *S*. The average feature values ($c_{c1}, c_{c2}, \ldots, c_{cD}$) are frequently used to describe each datum in *C*, where $c_{cj} = \frac{1}{n_c}\sum_{i=1}^{n_c} x_{cij}$, where n_c is the number of data in the *c*-th cluster. We call ($c_{c1}, c_{c2}, \ldots, c_{cD}$) the respective features or cluster center of *C*. For example, the corrected historic data are divided into two clusters respectively describing hepatic carcinoma patients and healthy people, and assume that their respective features are ($c_{p1}, c_{p2}, \ldots, c_{pD}$) and ($c_{h1}, c_{h2}, \ldots, c_{hD}$). When given the features ($x_{i1}, x_{i2}, \ldots, x_{iD}$) of a person, the recognition system will compare ($x_{i1}, x_{i2}, \ldots, x_{iD}$) with ($c_{p1}, c_{p2}, \ldots, c_{pD}$) and ($c_{h1}, c_{h2}, \ldots, c_{hD}$). If the distance between ($x_{i1}, x_{i2}, \ldots, x_{iD}$) and ($c_{h1}, c_{h2}, \ldots, c_{hD}$) is shorter than the distance between ($x_{i1}, x_{i2}, \ldots, x_{iD}$) and ($c_{n1}, c_{n2}, \ldots, c_{nD}$), the system will consider the person to a healthy person or to a hepatic carcinoma patient.

II. Bi-level k-means Algorithm

In a traditional K-means algorithm, a datum *X* will be assigned to the cluster *C* where the distance between *X* and the cluster center of *C* is minimal, comparing to the distances between *X* and the cluster centers of other clusters. However, the abnormal data may be assigned to most of clusters but normal data are classified into a few clusters. In Fig. 1, the red dots are the abnormal data which are only a few data relative to normal data and are classified to two clusters but the vast majority of normal data are classified to only one cluster. It is usually non-helpful for future analysis.

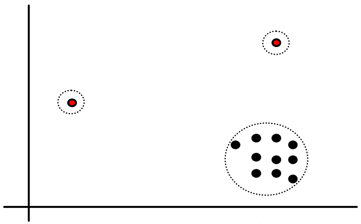

Figure 1. One example of clustering

In data clustering, the significantly different data should be assigned to different clusters and a big cluster containing a big number of data are frequently required to be classified into some small clusters. According to these two requirements, a bi-level K-means algorithm is proposed to improve the traditional K-means algorithm. The bi-level K-means algorithm contains four steps: data normalization, cluster center initialization, first-level clustering, second-level clustering.

A. Data Normalization

In distance-based classification, a small variation in one feature is probably more influencing than a big variation in other feature when computing the distance of two data. It is necessary to normalize every feature value of each feature dimension to a specific range. This step is to transform all features in the data to a specific range. Let $S=\{X_1, X_2, \ldots, X_N\}$ be a data set consisting of *N* data, X_i be the *i*-th data in *S*, and ($x_{i1}, x_{i2}, \ldots, x_{iD}$) be the features of X_i. For each feature value x_{id} is normalized into (1)

$$x'_{ij} = \frac{x_{ij} - \underset{k=1}{\overset{N}{MIN}}(x_{kj})}{\underset{k=1}{\overset{N}{MAX}}(x_{kj}) - \underset{k=1}{\overset{N}{MIN}}(x_{kj})}. \tag{1}$$

B. Initial Cluster Center

In this step, a most discrepant initial cluster center method is proposed to determine the initial cluster centers for K-means algorithm. It uses the biggest discrepant data as the initial cluster centers. Let the distance d_{ik} of two data C_1 and C_2:

$$d_{ik} = \sum_{d=1}^{D} w_d \left| x'_{ij} - x'_{kj} \right|^{r_d} \tag{2}$$

where w_d and r_d are the given constants. The algorithm first decides two data C_1 and C_2, where $C_1=X_i$ and $C_2 = X_k$, and $(i,k) = Arg\left(\underset{i=1}{\overset{N-1}{MAX}} \left(\underset{k=i+1}{\overset{N}{MAX}} d_{ik} \right) \right)$.

After that, it computes the data:

C_3 which is farthest from C_1 and C_2,

C_4 which is farthest from C_1, C_2, and C_2,

⋮

C_K which is farthest from C_1, C_2, …, and C_{K-1},

where C_1, C_2, …, and C_K are in S and are considered to the initial cluster centers of the K clusters.

C. First-Level Clustering

To understand the properties of the data in S, in the first-level clustering step, the bi-level K-means algorithm classifies the data in S into K' big clusters (we call them groups). According the number and the variation of the data in each group, how many numbers of clusters the data in each group should be divided into then can be derived. The bi-level K-means algorithm first partitions the data in S into K' groups by the traditional K-means algorithm where $K' < K$. In the first-level clustering steg, the (2) is also used to measure the distance between the i-th data $X_i=(x'_{i1}, x'_{i1}, \ldots, x'_{iD})$ in S and the cluster center $(c_{g1}, c_{g1}, \ldots, c_{gD})$ of the g-th group.

D. Second-Level Clustering

Let $(x_{gi1}, x_{gi2}, \ldots, x_{giD})$ be the i-th data in the g-th group obtained in the first-level clustering step. The second-level clustering step is to partition the data in the g-th group G_g into K_g clusters according the numbers and the variation of the data in G_g where $K=K_1+K_2+ \ldots +K_{K'}$. Let n_g be the number of data in the g-th group, and Std_g be

$$Std_g = \frac{\sum_{j}^{D} \sqrt{\frac{\sum_{i=1}^{n_g} \left(x_{gij} - u_{gj} \right)}{n_c}}}{D} \tag{3}$$

where $u_{gj} = \frac{\sum_{i=1}^{n_g} x_{gij}}{n_g}$.

Then the bi-level K-means algorithm classifies the data in the g-th group into K_g clusters by the traditional K-means algorithm, where K_g is as follows:

$$K_g = \frac{n_g \times std_g^r}{\sum_{i=1}^{K'} \left(n_i \times std_i^r \right)} \times K \text{ ,} \tag{4}$$

where r is a given constant.

III. Experimental Results

To evaluate the performances of the bi-level K-means algorithm, three well-known datasets **ABALUTE**, **IRIS**, and **WINE** downloaded from UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/) are used as the test data. Table I displays the information about the three data sets. The **ABALUTE** dataset consists of eight physiological features of abalone. These physical measurements are used to predict the age of abalone. There are 28 kinds of ages with 1 year old to 29 years old but no 28 years old. The 28 classes correspond to the 28 kinds of ages. The **IRIS** dataset is widely used to test classification algorithm. Each sample is given by measuring sepal length, sepal width, petal length, and petal width. This dataset has three classes corresponding to iris setosa, iris versicolour, and iris virginica, respectively. Each class has 50 samples and the total samples are 150 data. The wine dataset is chemical analysis of three kinds of wines that are analyzed to get 13 kinds of ingredients to form each data in the same region in Italy. Totally, there are 178 data. Respectively, there are 59, 71 and 48 data in the 1st, 2nd and 3rd categories. The traditional k-means algorithm and the bi-level k-means algorithm are executed on data sets **ABALUTE**, **IRIS**, and **WINE** respectively with K=28, 3, and 3.

TABLE I. UCI Machine Learning Repository

Data set	ABALUTE	IRIS	WINE
Samples	4177	150	178
Features	8	4	13
Classes	28	3	3

F-measure [18] is often used to measure the accuracy rate of the data clustering algorithms. In these experiments, the traditional K-means [16], Fast Global K-means [19], FKCUCD [20], and bi-level K-means algorithms will be used to classify the data sets **ABALUTE**, **IRIS**, and **WINE**, respectively and F-measure will be applied to evaluate their performances. Tables I to IV show the results obtained in these experiments. The experimental results prove that the bi-level k-means algorithm not only provide better performance of running time than others but also better accuracy rate.

TABLE II. Average performance of abalone data set for 30 times.

	K-Means	Global K-means	FKMCUCD	Bi-level
Accuracy Rate	0.1477	0.1424	0.1470	0.1739
Running Time (sec)	23.34	1024.74	44.50	9.16

TABLE III. Average performance of iris data set for 30 times.

	K-Means	Global K-means	FKMCUCD	Bi-level
Accuracy Rate	0.8733	0.8733	0.87333	0.9591
Running Time	0.031	0.12	0.19	0.06

TABLE IV. Average performance of wine data set for 30 times.

	K-Means	Global K-means	FKMCUCD	Bi-level
Accuracy Rate	0.9606	0.9550	0.9607	0.9629
Running Time	0.02	0.15	0.13	0.08

IV. Conclusions

In this paper, a bi-level K-means algorithm is proposed to improve traditional K-means algorithm. It can give a better accuracy rate of data clustering than other *K*-means algorithms. It also can deal with the apt question influenced by noisy data or outliers and accelerate the clustering speed. It improves to avoid high numbers of classifications which causes low speed of convergence of cluster centers too, thus, to divide data points into *K* clusters in two stages rather than in one stage. The experimental results show that the bi-level K-means algorithm is better than other K-means algorithms in accuracy rate and the computation speed.

REFERENCES

[1] M. N. Murty, A. K. Jain, and P. J. Flynn, “Data Clustering: a Review,” ACM Computing. Surveys, Vol. 31, No.3, pp. 264–323, 1999.

[2] M. R. Anderberg, “Cluster Analysis for Applications,” *Academic Press*, New York, NY, 1973.

[3] P. Rai, and S. Singh, “A Survey of Clustering Techniques,” *International Journal of Computer Applcations*, Vol. 7, No. 12, pp. 1-5, Oct. 2010.

[4] A. K. Jain and P. J. Flynn, “Image Segmentation Using Clustering,” In *Advances in Image Understanding: A Festschrift for Azriel Rosenfeld*, N. Ahuja and K. Bowyer, Eds, IEEE Press, Piscataway, NJ, pp. 65-83, 1996.

[5] S. Theodoridis, and K. Koutroumbas, *Pattern Recognition*, 2nd Ed., Academic Press, New York, 2003.

[6] U. M. Fayyad, “Data Mining and Knowledge Discovery: Making sense out of data,” *IEEE Expert* Vol. 11, No. 5, pp. 20–25, Oct. 1996.

[7] O. R. Zaiane, A. Foss, C. H. Lee, and W. Wang, “On Data Clustering Analysis: Scalability, Constraints, and Validation,” *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD '02),* pp. 28-39, 2002.

[8] M. Erisoglu, N. Calis, and S. Sakallioglu, “A New Algorithm for Initial Cluster Centers in K-means Agorithm,” *Pattern Recognition Letters*, Vol. 32, No. 14, pp. 1701-1705, 2011.

[9] M. N. Murty and G. Krishna, “A Computationally Efficient Technique for Data Clustering,” *Pattern Recogn*ition, Vol. 12, pp. 153–158, 1980.

[10] J. Foster, R. M. Gray, and M. O. Dunham, “Finite State Vector Quantization for Waveform Coding, *IEEE Transactions on Information Theory*,Vol. 31, No. 3, pp. 348–359, 1985.

[11] A. Gersho, and R.M.Gray, “Vector Quantization and Signal Compression,” Kluwer Academic Publishers, Boston, MA, 1991.

[12] J. Z. C. Lai, Y. C. Liaw, W. Lo, “Artifact Reduction of JPEG Coded Images Using Mean-Removed Classified Vector Quantization,”*Signal Processing*, Vol. 82, No. 10, 2002.

[13] Y. C. Liaw, J. Z. C. Lai, and W. Lo, “Image Restoration of Compressed Image Using Classified Vector Quantization,”*Pattern Recognition*, Vol. 35, No. 2, pp. 181–192, 2002.

[14] G. Nagy, “State of the art in pattern recognition,” *In: Proceedings of the IEEE*, Vol. 56, No. 5, pp. 836–862, May, 1968.

[15] W. H. E. Day, “Complexity theory: An introduction for practitioners of classification,” In *Clustering and Classification*, P. Arabie and L. Hubert, Eds. World Scientific Publishing Co., Inc., River Edge, NJ, 1992.

[16] J. MacQueen, “Some Methods for Classification and Analysis of Multivariate Observations,” *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, University of California Press, pp. 281–297, 1967.

[17] K. J. Anil, “Data Clustering: 50 Years beyond K-means,” *Pattern Recognition Letters*, Vol. 31, No. 8, pp. 651 – 666, 2010.

[18] C. J. Van Rijsbergen, “Information Retrieval (2nd ed.)” London: Butterworths, 1979.

[19] A. Likas, N. Vlassis, and J. Verbeek, “The Global K-Means Clustering Algorithm,”*Pattern Recognition*, Vol. 36, No. 2, pp. 451-461, 2003.

[20] J. A. Lozano, J. M. Pena, and P. Larranaga, “An Empirical Comparison of Four Initialization Methods for the K-means Algorithm,” Pattern Recognition Letter, Vol. 20, pp. 1027–1040, 1999.