Improved Hierarchical K-means Clustering Algorithm without Iteration Based on Distance Measurement

Wenhua Liu, Yongquan Liang*, Jiancong Fan, Zheng Feng, and Yuhao Cai

College of Information Science and Engineering,
Shandong University of Science and Technology,
Qingdao City, 266590, China
lyq@sdust.edu.cn

Abstract. Hierarchical K-means has got rapid development and wide application because of combining the advantage of high accuracy of hierarchical algorithm and fast convergence of K-means in recent years. Traditional HK clustering algorithm first determines to the initial cluster centers and the number of clusters by agglomerative algorithm, but agglomerative algorithm merges two data objects of minimum distance in dataset every time. Hence, its time complexity can not be acceptable for analyzing huge dataset. In view of the above problem of the traditional HK, this paper proposes a new clustering algorithm iHK. Its basic idea is that the each layer of the N data objects constructs $\left\lceil \frac{N}{2} \right\rceil$ clusters by running K-means algorithm, and the mean vector of each cluster is used as the input of the next layer. iHK algorithm is tested on many different types of dataset and excellent experimental results are got.

Keywords: basic K-means, traditional HK, iHK, Clustering Algorithm.

1 Introduction

Traditional HK algorithm has the advantage of simple and easy to convergence, but also it has some obvious deficiencies. For instance, HK would have a high computational complexity when the k value is uncertain. Agglomerative algorithm only mergers two clusters having minimum distance every time, which leads to have a higher time complexity at high dimension and big data. To overcome the shortcoming of traditional HK, some researchers have done different degrees of improvement for HK algorithm [1-4]. But most of researchers have modified HK algorithm for their specific research fields [5-13].

Now society is rapidly being from information era to age of data, it is significant to accurately grasp the valuable information in dataset. Therefore clustering analysis has become a hot research field chased by researchers. To precisely and quickly analyze the data information carried by the dataset, this paper presents a new clustering algorithm iHK, which is a easily convergent and quite accurate clustering algorithm by integrating with the feature of the K-means and hierarchical algorithm. Moreover, iHK

* Corresponding author.

Z. Shi et al. (Eds.): IIP 2014, IFIP AICT 432, pp. 38–46, 2014.

algorithm is not limit to a particular research field. In essence, iHK clustering algorithm is an improved HK.

Section 1 describes the summary of iHK algorithm. Section 2 briefly introduces some improved HK algorithm in recent years. Section 3 details iHK algorithm proposed by this paper. And Section 4 presents the experimental results . Section 5 makes a conclusion.

2 Related Work

The training set is divided into two parts, a part of the dataset uses hierarchical algorithm to obtain distribution information of data, then runs K-means at another part. This hybrid algorithm was put forward for the first time by Bernard Chen et al.[14] at 2005. Due to its accuracy, simplicity and convergence, the method attracted wide attention after it has been proposed. He Ying et al.[15] presented HK based on PCA, the general idea of algorithm is that on the whole dataset (rather than two parts), it first makes use of PCA technology to reduce dimension of the dataset and then determines the initial cluster center by executing agglomerative algorithm. Finally it gets clustering results by using K-means. Improved HK based on PCA has more accuracy than traditional HK. To overcome the limitation of a binary tree constructed by hierarchical algorithm, a kind of divisive clustering algorithm was come up with by Lamrous S, Taileb M et al.[16]. The algorithm generates an non-binary tree where each node can split more than two branches by employing K-means, where the k value of K-means is determined by Silhouette index. Kohei Arai et al.[2] studied on an integrated HK algorithm to cluster high-dimensional data set. An improved HK algorithm was put forward by Yongxin Liu et al.[5] to solve the problem of document clustering possessing big data and high-dimension. Li Zhang et al.[17] combined divisive algorithm with agglomerative algorithm to address irreversibility of HK. Divisive algorithm gets several clusters by executing K-means at each layer, then utilizes agglomerative algorithm to merger clusters. Bernard Chen et al.[18] added fuzzy theory into traditional HK for boosting the precision of finding protein sequences theme and reducing the time complexity of HK algorithm.

According to the above related work, some ideas about significantly improving efficiency and accuracy are drawn which they can be applied to iHK. Such as, when using agglomerative algorithm no longer relies on binary tree rules and similarity measure between clusters would adopt mean distance rather than minimum distance to avoid serious impact of noise data on precision of algorithm. In iHK algorithm, the number of clusters produced by K-means always varies with the change of layer (data size). So cluster center can present the distribution of data as far as possible.

3 iHK Clustering Algorithm

3.1 Normalization of Data

Paper gives a new algorithm iHK on the basis of learning a large number of improved HK clustering algorithms. iHK firstly standardizes attributes by formula (1), because diverse attributes may adopt different units of measurement in some dataset.

$$Z_{if} = \frac{x_{if} - m_f}{S_f} \tag{1.1}$$

In (1.1), S_f and m_f is given in detail in the following (1.2) and (1.3).

$$S_f = \frac{1}{n}(|x_{1f} - m_f| + |x_{2f} - m_f| + \ldots + |x_{nf} - m_f|) \tag{1.2}$$

where $x_{1f}, x_{2f}, \ldots\ldots, x_{nf}$ is n values of the f attribute and m_f is average of f attribute.

$$m_f = \frac{x_{1f} + x_{2f} + \ldots\ldots + x_{nf}}{n} \tag{1.3}$$

Then data objects obtained by above-mentioned steps are used as clustering data. iHK algorithm greatly promotes execution efficiency by combining the idea of 2-way merge, because running algorithm can reduce half the number of clusters at every time. To improve traditional HK algorithm would use K-means algorithm to clustering data at each layer rather than simply merger two clusters by adopting minimum distance between clusters. Next, traditional Hierarchical clustering algorithm and iHK algorithm are simply introduced.

3.2 Traditional Hierarchical Algorithm

Assuming that dataset is D and the total number of data object is N. At first, treating every data as a single cluster, hence there are N clusters. Similarity measure between cluster C_i and C_j is shown by (2).

$$dist_{\min} = (C_i, C_j) = \min_{p \in C_i, q \in C_j} \{|p - q|\} \tag{2}$$

The number of clusters just reduces one at a time through merging two clusters by (2). This process is being performed continually until meeting the given threshold or all data are in one cluster. The time complexity of hierarchical algorithm is $O(N^2)$, thus it is not suitable for processing huge dataset.

3.3 iHK Clustering Algorithm

iHK overcomes limit of merely aggregating two clusters one time at hierarchical algorithm so that traditional hierarchical algorithm can be better applied to clustering big data. In addition, iHK no longer merges clusters in a minimum distance manner at each layer, but it makes use of K-means algorithm based on mean distance measurement. Supposing that hth layer has L data objects, the process of iHK algorithm based above assumption is that it selects cluster center by employing (3) before L data objects are divided into $\left\lceil \frac{L}{2} \right\rceil$ clusters through K-means algorithm.

$$\begin{cases} 1,3,5,\ldots\ldots i, i+2, \ldots\ldots L, & L \text{ is odd number} \\ 1,3,5,\ldots\ldots i, i+2, \ldots\ldots L-1, & L \text{ is even numbe} \end{cases} \tag{3}$$

where $1,3,\ldots\ldots i$ is a sequence of data objects.

By comparing the distance between the remaining data and cluster centers, data is divided into the closest cluster center. However, attributes may be mixed, therefore it can't adopt simple distance metric value as a criterion. New standard is defined by (4.1).

$$dist\ (x, y) = \frac{\sum_{i=1}^{n} d^{\,i}_{x_i, y_i}}{n} \tag{4.1}$$

where x is a data object with n attributes, namely $x = (x_1, x_2, \ldots\ldots, x_n)$, and y is a cluster center with n attributes, namely $y = (y_1, y_2, \ldots\ldots, y_n)$. If the ith attribute of x is discrete data type, then $d^{i}_{x_i, y_i} = 0$ iff $x_i = y_i$. Otherwise $d^{i}_{x_i, y_i} = 1$. If this attribute is numeric type, then $d^{i}_{x_i y_i}$ is calculated by (4.2).

$$d^{\,i}_{x_i, y_i} = \left| x_i - y_i \right| \tag{4.2}$$

After merging, altogether forms $\left\lceil \frac{L}{2} \right\rceil$ clusters. Mean vector $\overline{x} = (\overline{x_1}, \overline{x_2}, \ldots\ldots, \overline{x_n})$ is used to represent cluster and the calculation formula of mean vector is given by (5).

$$\begin{cases} \overline{x_i} = \dfrac{1}{count\ (C_k)} \sum_{x_{ij} \in C_k} x_{ij}, & x_i \text{ is a numeric type attribute} \\ \overline{x_i} = \max(\, x_{ij} \in C_k \,|\, count\ (x_{ij})), & x_i \text{ is a discrete type attriute} \end{cases} \tag{5}$$

In formula (5), $count(C_k)$ is applied to count the number of data object in C_k cluster and $count(x_{ij})$ is used to compute how many values equal x_{ij}. If x_i is numeric type, $\overline{x_i}$ is a average of x_i in C_k. If x_i is a discrete type, then $\overline{x_i}$ equals to the value of attribute appearing most times in dataset.

$\left\lceil \frac{L}{2} \right\rceil$ clusters are produced after the aforementioned process has been executed, then mean vectors obtained at h th layer are used as the new input of h+1th layer. iHK algorithm is being executed until satisfying given condition or threshold. iHK algorithm performs $\log_2 N$ times K-means in total, in which K-means just compares half of the input data in this layer. Through the above details, the overall description of the iHK algorithm is shown in Figure 1.

In Figure 1, Step1 is to standardize the numeric type data so that between attributes of different measure units can make the correct comparison. Distance calculation between data and cluster center is completed by Step3-4. Mean vectors generated by (5) are used to represent clusters and use them as input of next circulation.

The pseudocode of the algorithm is described by Figure 2.

Input: data set D , the total number of data object n .
Output: Data object are divided into different clusters.
Step1:preprocessing attributes in data set D by (1).
Step2:employing step length 2 to select cluster center by according to the order of the data storage.
Step3:computing distance by (4.2) if attribute is numeric type.
Step4:comparing discrete attribute values whether or not they are the same. Distance is 0,iff their value is the same. Otherwise, the value of distance is 1.
Step5:Performing K-means on the basis of step3-4.
Step6: Treating the results of K-means clustering as new dataset.
Step7:Implementing (5) to process dataset formed by Step6.
Step8:Repeat Step2-7 until meeting the given conditions.

Fig. 1. A complete description of the iHK clustering algorithm

4 Experimental Results

This paper tests the performance of iHK at some typical dataset of diverse type. Test set contain Abalone, Iris, Seeds_dataset, Wine, Credit Approval, Haberman's Survival, Teaching Assistant Evaluation, Ecoli, MONK'S Problem, Balance scale, balloons and so on. They are from UCI database. (`http://archive.ics.uci.edu/ml/`)

In iHK algorithm, K-means based on new distance metric is executed at each layer. Among, the value of k is different, which is half size of each layer input data. Accuracy rate changes with different k values, which is shown in Figure 3.

The accuracy of iHK algorithm on five large dataset is shown at top of the Figure 3 and Accuracy on relatively small dataset is displayed at the bottom.

Input: dataset D , the total number of data object n .
Output: several clusters.
for each numeric attribute f
do
for i←0 to n
standardizing attribute value;
do
for i←0 to n/2
datacenter[i][t]=datacenter[i*2][t];
exdata[i][t]=datacenter[(i*2)+1][t];
end;
dividing exdata[i] into appropriate cluster by defined distance formula;
if (x_i is a numeric type attribute)

$$\text{then} \quad \overline{x_i} = \frac{1}{count(C_k)} \sum_{x_{ij} \in C_k} x_{ij};$$

$$else \quad \overline{x_i} = \max(x_{ij} \in C_k \,|\, count(x_{ij}));$$

forming new dataset by the mean vector of each cluster;
until meeting the given conditions.

Fig. 2. The pseudocode of iHK algorithm

The overall trend of all line chart is that they begin to rise gradually and reach maximum at some value, then decrease. Besides, accuracy obviously varies with the change of k values on some dataset, as Figure 3 expresses.

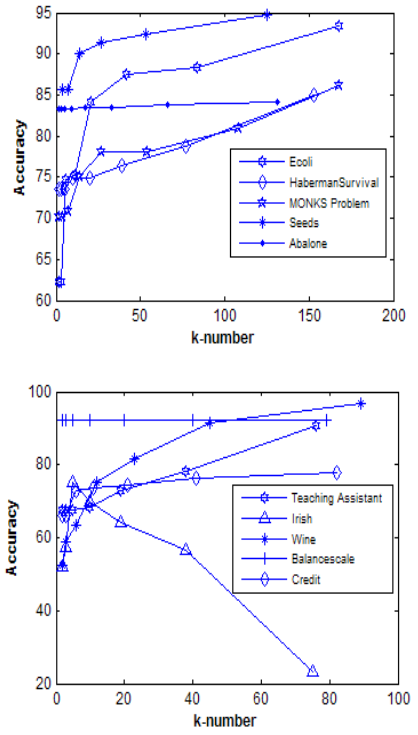

Fig. 3. The line chart of clustering accuracy rate

Table 1. Accuracy of three algorithms

Data Sets	basic k-mean	HK	iHK
Abalone	57.4%	50.8%	83.4%
Wine	68.1%	68.5%	75.3%
Seeds	75.7%	82.8%	85.7%
Iris	70.2%	88.5%	75.3%
Credit Approval	54.7%	54.6%	72.9%
Haberman Survival	73%	73.5%	75.2%
Teaching Assistant Evaluation	65.2%	40%	67.6%
Ecoli	75%	78.3%	76%
MONK Problem	67.6%	66.7%	70.2%
Balance scale	90%	62.2%	92.2%
balloons	65.4%	80%	80%

To make experimental results can show the superiority of iHK clustering algorithm, this paper compares iHK algorithm with basic K-means and traditional HK algorithm at some evaluation indicators of the performance of algorithm. Accuracy rate is frequently used as an important indicator. The accuracy rate of three algorithms at some dataset are shown in Table 1.

Experimental results in Table 1 show that iHK clustering algorithm has higher accuracy than the basic K-means and HK algorithm at most dataset.

Now most of the data come from Web and Web data is mainly big data. When these data are clustered, the time complexity of the algorithm is also considered as a significant indicator of performance of prediction algorithm.

In Table 2, time complexity of HK algorithm, basic K-means, iHK algorithm are compared.

The time complexity of HK algorithm should be the sum of the time complexity of Hierarchical algorithm and K-means', it is greater than $O(N^2)$.

The time complexity of basic K-means is linear, where m is the number of iteration. iHK is similar to K-means. It is also linear, where N is the total number of data objects. By comparison, this conclusion can be drawn, which the time complexity of iHK algorithm is minimum, basic K-means follows and HK algorithm's is maximum .

The different time complexity are generated by three algorithms at experimental data set, which use Figure 4 to show.

The time complexity of HK algorithm is significantly higher than iHK at most dataset. iHK and Basic K-means are about equal at all dataset, which can be clearly observed from Figure 4.

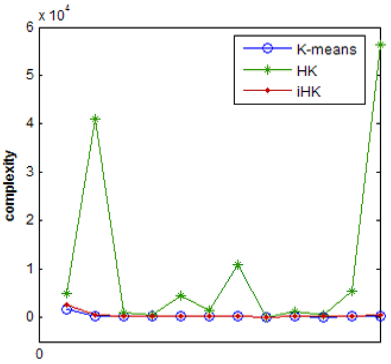

Fig. 4. The time complexity of three algorithms at some data set

Table 2. Comparisons of the time complexity of different algorithms

The time complexity of algorithms	
HK	$O(N^2)$
Basic K-means	$O(mN)$
iHK	$O((1-\frac{1}{2}^{\log_2 N})N$

5 Conclusion

HK algorithm has been widely used, due to its superiority in the clustering analysis. But the HK algorithm also has shortcoming, such as high time complexity. So, it can not be applied to clustering big data. Therefore, some improved HK algorithms have been studied by some researchers for resolving the problem of specific application areas. Some improved HK algorithm have good clustering outcome, but these clustering results are not good in certain application domain. iHK algorithm has good generalization in that it is not limited to a particular field. And it can be used to clustering big data since it has low time complexity.

From accuracy, efficiency and the time complexity acquired by comparing these three algorithms, a conclusion can be drawn that iHK has the advantage of high accuracy and easy to convergence. In addition, its performance is distinctly superior to other algorithms, but time complexity is similar to basic K-means's. The important point is that iHK is not based on any special application areas and easy to integrate to other clustering algorithm. However, iHK algorithm still can not solve irreversibility of HK algorithm. The next mainly task is that iHK algorithm is further improved.

Acknowledgment. This paper is supported by State Key Laboratory of Mining Disaster Prevention and Control Co-founded by Shandong Province and the Ministry of Science and Technology, Shandong University of Science and Technology, Leading talent development program of Shandong University of Science and Technology, National Natural Science Foundation of China under Grant 61203305 and Natural Science Foundation of Shandong Province of China under Grant ZR2012FM003.

References

[1] Wang, Y.C.F., Casasent, D.: Hierarchical k-means clustering using new support vector machines for multi-class classification. In: International Joint Conference on IEEE Neural Networks, IJCNN 2006, pp. 3457–3464 (2006)

[2] Arai, K., Barakbah, A.R.: Hierarchical K-means: an algorithm for centroids initialization for K-means. Reports of the Faculty of Science and Engineering 36(1), 25–31 (2007)

[3] Lu, J.F., Tang, J.B., Tang, Z.M., et al.: Hierarchical initialization approach for K-Means clustering. Pattern Recognition Letters 29(6), 787–795 (2008)

[4] Celebi, M.E., Kingravi, H.A.: Deterministic initialization of the k-means algorithm using hierarchical clustering. International Journal of Pattern Recognition and Artificial Intelligence 26(07) (2012)

[5] Archetti, F., Campanelli, P., Fersini, E., et al.: A Hierarchical document clustering environment bases on the induced bisecting k-means. Flexible Query Answering System, pp. 257–269. Springer, Heidelberg (2006)

[6] Liu, Y., Liu, Z.: An improved hierarchical K-means algorithm for web document clustering. In: International Conference on Computer Science and Information Technology, ICCSIT 2008, pp. 606–610. IEEE (2008)

[7] Murthy, V., Vamsidhar, E., Rao, P.S., et al.: Application of hierarchical and K-means techniques in Content based image retrieval. International Journal of Engineering Science and Technology 2(5), 749–755 (2010)

[8] Chen, T.W., Chien, S.Y.: Flexible hardware architecture of hierarchical K-means clustering for large cluster number. IEEE Transactions on Very Large Scale Integration (VLSI) Systems 19(8), 1336–1345 (2011)

[9] Hu, X., Qi, P., Zhang, B.: Hierarchical K-means algorithm for modeling visual area V2 neurons. Neural Information Processing, pp. 373–381. Springer, Heidelberg (2012)

[10] Mantena, G., Anguera, X.: Speed improvements to information retrieval-based dynamic time warping using hierarchical k-means clustering. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 8515–8519. IEEE (2013)

[11] Chen, B., He, J., Pellicer, S., et al.: Using Hybrid Hierarchical K-means (HHK) clustering algorithm for protein sequence motif Super-Rule-Tree (SRT) structure construction. International Journal of Data Mining and Bioin-formatics 4(3), 316–330 (2010)

[12] Ghwanmeh, S.H.: Applying Clustering of Hierarchical K-means-like Algorithm on Arabic Language. International Journal of Information Technology 3(3) (2007)

[13] Chehata, N., David, N., Bretar, F.: LIDAR data classification using hierarchical K-means clustering. In: ISPRS Congress, Beijing, vol. 37, pp. 325–330 (2008)

[14] Chen, B., Tai, P.C., Harrison, R., et al.: Novel hybrid hierarchical-K-means clustering method (HK-means) for microarray analysis. In: Computational Systems Bioinformatics Conferences, Workshops and Poster Abstracts, pp. 105–108. IEEE (2005)
[15] Ying, H., Qin, L.X.: Study on PCA based Hierarchical K-means Clustering Algorithm. Control and Automation 6, 68 (2012)
[16] Lamrous, S., Taileb, M.: Divisive hierarchical k-means. In: 2006 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on Computational Intelligence for Modelling, Control and Automation, p. 18. IEEE (2006)
[17] Zhang, L., Cui, W.D., et al.: Hybird clustering algorithm based on partitioning and hierarchical method. Computer Engineering and Applications 46(16), 127–129 (2010)
[18] Chen, B., He, J., Pellicer, S., et al.: Protein Sequence Motif Super-Rule-Tree (SRT) Structure Constructed by Hybrid Hierarchical K-means Clustering Algorithm. In: IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2008, pp. 98–103. IEEE (2008)