2015 The 5th International Workshop on Computer Science and Engineering
doi: 10.18178/wcse.2015.04.077

An Improved Clustering Algorithm for Big Data Based on K-Means with Optimized Clusters' Number

Lianjiang Zhu[1], Tao Du[2], Shouning Qu[1], Kai Wang[2], Yong Zhang[3]

[1] Information network centre, University of Jinan, China, 250022, Jinan

[2] School of information science and engineering, University of Jinan

[3] School of electrical engineering, University of Jinan

Keywords: Big data, Silhouette Coefficient, clustering, optimized clusters' number.

Abstract. To improve the processing ability of big data, a new clustering algorithm is proposed which is designed based on K-means. In this algorithm, a concept of "Silhouette Coefficient" is defined to estimate the result of clustering. Based on silhouette coefficient, the optimized clusters' number would be chosen, and then K-means algorithm would be operated with this clusters' number. The algorithm is tested by a real production big data set and compared with classical K-means. The result of experiment proves that the improved algorithm has more reasonable result of clustering with little extra calculation.

Introduction

The researches of big data have been one of the most important fields in information science, and there are many technologies have been developed to promote its application [1, 2]. For the features of big data, the key factor of improving technologies' performance is the efficiency of dealing with mass data [3, 4]. To improve the efficiency, many methods have been researched. Among all existing methods, clustering is one of the most efficient means through divided the target data set to several parts according to some special rules. After clustering, big data can be dealt with according to its attributes and the total computation would be much reduced, so the efficiency can be obvious improved. But for the scale of big data, the existing algorithms can't obtain the reasonable result of clustering in application. To improve the clustering efficiency in big data application, a new high efficient algorithm named O-K-means is proposed in this paper based on classical K-means, and the clusters' number would be optimized by the "Silhouette Coefficient" of clustering result. In the rest of this paper, section 2 introduces the existing researches and analyzed these algorithms; section 3 introduces the related conception of clustering algorithm; section 4 introduces the detail of clustering algorithm; section 5 analyzes the advantages of our algorithm.

Related Works

Clustering is one of the most important methods of dealing with big data, and there have been many algorithms proposed [5]. All algorithms can be divided into three types: the ones based on hierarchy, such as HAC, CURE, and ROCK; the ones based on division, such as K-means, DBSCAN, and STING; and the ones based artificial intelligence, such as artificial neural nets clustering and evolutionary computing clustering, but these existing algorithms have some disadvantages in big data application. Typical hierarchy algorithms have advantages in handling data with complex shape and different cluster sizes, but their process of calculation is too complex which lead to poor efficiency. The classical division algorithms have advantage in calculation efficiency, but these algorithms are inefficient in handling data with complex shape. The algorithms based on artificial intelligence have the highest clustering accuracy of all types, but their performance is much influenced by the setting of initial parameters and the efficiency of calculation is relatively low [6].

Among all algorithms, K-means which is designed based on division has been widest applied for its simply structure with relatively high performance in clustering [6]. But the classical K-means algorithms should define the clusters' number before clustering, and this number would influence

the clustering result with the initial centers of every cluster. So many improved K-means algorithms have been proposed [7, 8]. In these algorithms, the key factor is how to defined the number of clustering, but in these algorithms, there are many human interventions in process of calculation, and it would influence the objectivity of result.

The Thought of O-K-Means

According to the analysis of existing researches of clustering, a new high efficient clustering algorithm based on K-means is proposed in this paper, and the key thought of this algorithm is that the number of clusters would be optimized automatically before clustering, so this algorithm is named O-K-means.

In O-K-means, a conception of "Silhouette Coefficient" is defined to judge the result of clustering with candidate numbers of clusters. The main thought of O-K-means is as follows: the standards of clustering effectiveness should include two aspects: one is the inner difference of clusters' member which can be called "cohesion", and the other is the difference between every pair of clusters which can be called "resolution"; and the most idealized result of clustering is that all clusters' cohesion should be the least and all clusters' resolution should be the largest. Based on these two standards, a new conception of "Silhouette Coefficient" is defined, and if there is a single data sample di, and it has been distributed to cluster A, and then the silhouette coefficient of di Si can be described as (1).

$$s_i = \frac{b_{\mathrm{i}} - a_{\mathrm{i}}}{\max(a_i,\ b_i)} \tag{1}$$

In (1), ai means the average distance from di to other members of cluster A, bi means the minimum distance from di to all other clusters' member. Based on the silhouette coefficient of sample data, the clustering result's silhouette coefficient can be computed as (2).

$$s_{\mathrm{k}} = \frac{1}{n}\sum_{i=1}^{n} s_i \tag{2}$$

In (2), n is the number of all data samples and k is the number of clusters. Sk can be used to judge the effectiveness of clustering. According to the definition of Sk, it can be concluded that the range of Sk is [-1, 1]. When the value of Sk is larger than zero, the result of clustering would turn better with the increase of Sk. By comparing the silhouette coefficient of every clustering result, the optimized number of clustering can be set, and the detail of O-K-means would be designed in the next section.

The Detail Design of O-K-Means

In O-K-means, there are four steps: the first step is data cleaning; the second step is choosing sample data; the third step is computing the best number of clusters; the last step is clustering by K-means. The whole process of O-K-means is shown in fig. 1.

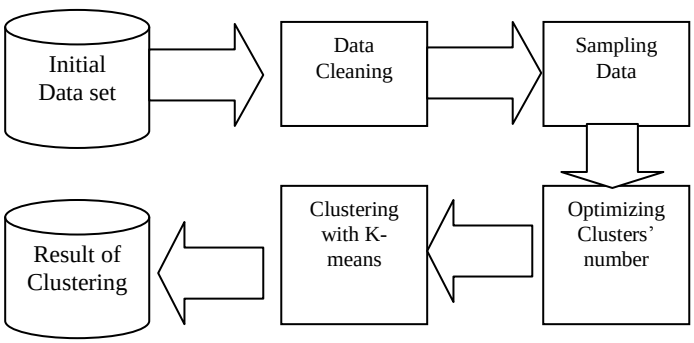

Fig.1: The process of O-K-means

Step1. In initial big data, there must be much invalid or wrong data for restricts of information collection technology and hardware. So before computing the target data, some data cleaning work should be finished [9]. Firstly, the attributes with too much empty value would be deleted for these ones have not enough contribution in application but would influence the final result. Then in the remaining contributes, if still existing continuous null values, these values would be filled by estimated value. The estimating method is to build an arithmetic sequence: some values before null values would be as the first part of sequence and some values after null values would be as the last part of sequence, and the corresponding values would be filled by the order of sequence.

Step2. The scale of big data is too mass to be handled directly with optimized cluster number, so the target data set would be sampled to reduce the computation complexity. In this paper, the data segment which has the largest change would be chosen as the sample data., because changes means information, and the larger means more information. The whole data set's lifetime would be divided into several parts as $T=\{t_1,t_2,\cdots,t_m\}$, and $T_y=\{t_{u_y},t_{u_{y+1}},\cdots,t_{u_{y+h-1}}\}$ is a random time segment of T, and the change amount of data in T_y is defined as (3). And in T there would be a special time segment T_m, and in Tm, the change amount is the largest, then the data in T_m would the sample data to be chosen to computer the optimized k.

$$\Delta x_y=\sum_{i=1}^{n}\sum_{j=t_{u_y}}^{t_{u_y+h-1}}\left|\Delta x_i\left(t_j\right)\right| \tag{3}$$

Step3. After choosing the sample data set, the optimized clustering number could be calculated according to the method mentioned in section 3. By (2), for a random attributes X_i in sample data, its best number of clustering K_m can be computed by (4), and in (4) Ω is the all result of clustering when clusters' number is k.

$$K_M=\min_K\left\{\max_\Omega s_k\right\} \tag{4}$$

Step4. After obtaining the best number of clusters, K-means algorithm can be operated to dealing with the whole data set. And the distance between two data would be computed by typical Euclidean method as (5), and the objective function is designed as (6).

$$EUCLID\left(p_1,p_2\right)=\sqrt{\sum_{i=1}^{t}\left(p_{1_i}-p_{2_i}\right)^2} \tag{5}$$

$$E=\sum_{i=1}^{k}\sum_{p=c_i}\left\|p-m_i\right\|^2 \tag{6}$$

In (5), p1 and p2 mean two sample data, and there are t attributes to be computed. In (6), Ci is a cluster of result, and p is a member of Ci, mi is the center data value of Ci. The whole process of K-means has been well introduced by many researches and this paper does not repeat it.

The Experiment and Analysis

To prove the performance of O-K-means, a series of experiments are designed and typical K-means algorithm would be compared with O-K-means in this section. In these experiments, the testing samples are collected from real-time production monitoring data of a thermal power plant. There are one hundred and sixty three attributes and millions of items, and after data cleaning, there are still 63 attributes, so this is a typical big data application. To overall estimate the performance of O-K-means, two standards are used: one is the silhouette coefficient of clustering; the other is the time cost of clustering.

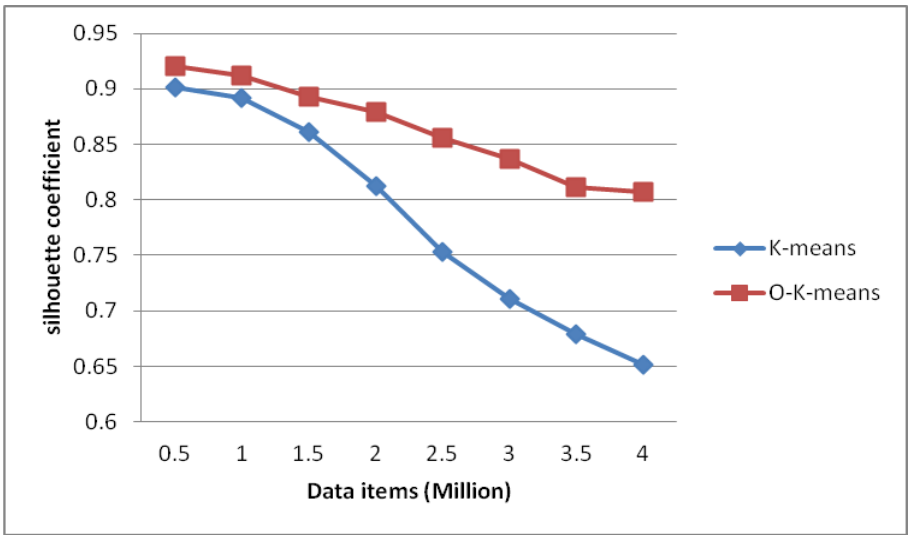

Fig.2: The change tendency of silhouette coefficient

In fig. 2, the silhouette coefficient's change with the increase of data scale is demonstrated. The condition of this experiment is that the classical K-means adopts the clustering number 5 recommended by experience, and O-K-means adopts the clustering number 6 which is computed. According to fig. 2, when the data's scale is smaller than one million items, these two algorithms have similar silhouette coefficient in 0.9; when the data's scale is between 1.5 million to 3 million, O-K-means' silhouette coefficient is still maintain in 0.83-0.87, but K-means' silhouette coefficient is lower than 0.7; and when the data's scale is larger than 3 million, the advantage of O-K-means is more and more clear. It is obvious that O-K-means has more reasonable clustering result and it is more fit for big data application.

In fig. 3, the clustering time cost's change with the increase of data scale is demonstrated. The condition of this experiment is same with last one. According to fig. 3, when the data's scale is smaller than one million items, O-K-means' computing efficiency is obvious lower than K-means; when the data's scale is between 1.5 million to 3 million, the gap between O-K-means to K-means is narrowing; and when the data's scale is larger than 3 million, the time cost of O-K-means is very near to K-means. It can be concluded that O-K-means has similar computing efficiency for big data application.

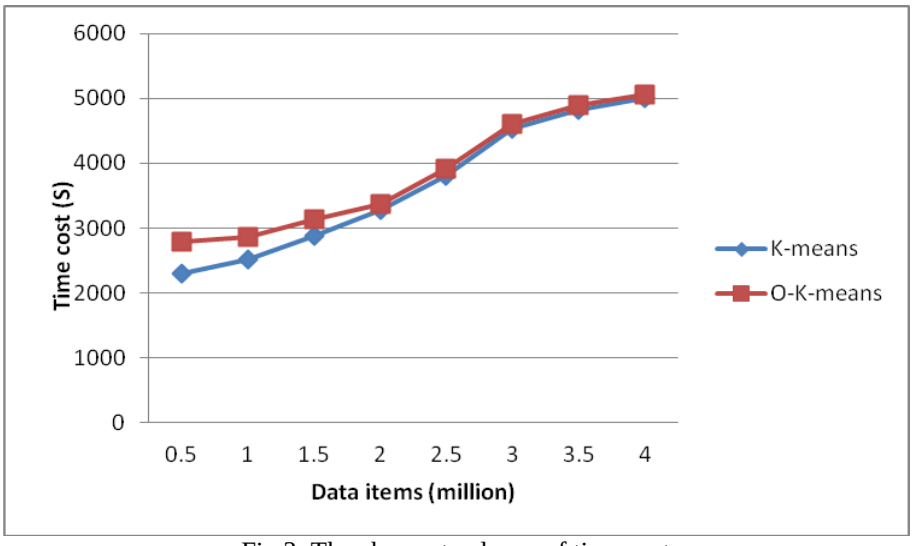

Fig.3: The change tendency of time cost

Conclusion

In this paper, an improved K-means clustering algorithm named O-K-means is proposed. The main contribution of this paper is that the definition of "Silhouette Coefficient". Based on this definition, the clustering result can be estimated, and according to the estimation, optimized clusters' number can be chosen, and the entire clustering algorithm is designed. By data cleaning and data sampling procedure, the calculation efficiency of O-K-means is maintained in a high level with more reasonable clustering, and it is fitter for big data application. At last, the performance of O-K-means is proved by experiment.

Acknowledgements

The research work was supported by Natural Science Foundation of China under the contract number 60573065; and is supported by Natural Science Foundation of Shandong under the contract number 2010ZRQL02 and ZR2012FL12; and supported by independent innovation special programs of Shandong under the contract number 2012CX30302.

References

[1] Gan X., Academician Li G. J. Big data turns into new interesting points in IT. China Science Daily, 2012-6-27

[2] Ibrahim .A. T. Hashem, Ibrar. Y, Nor. B. A, Salimah. M, Abdullah .G, Samee. U. K. The rise of "big data" on cloud computing: Review and open research issues. Information Systems, 47, pp. 98-115, 2015.

[3] Barbierato. E, Gribaudo. M, Iacono. M. Performance evaluation of NoSQL big-data applications using multi-formalism models. Future Generation Computer Systems, 37, pp. 345–353, 2014.

[4] Loshin. D. Big Data Analytics, DOI: 10.1016/B978-0-12-417319-4.00001-6.

[5] Zhao. Y. C. Clustering, R and Data Mining, Examples and Case Studies, Academic Press, 2013.

[6] Lee. I, Yang. J. Common Clustering Algorithms, Reference Module in Chemistry, Molecular Sciences and Chemical Engineering, 2, pp. 577–618, 2009.

[7] Michio. Y, Yoshikazu .T. Functional factorial K-means analysis. Computational Statistics & Data Analysis, 79, pp. 133–148, 2014.

[8] Mikko. I. M, Radu. M. I, Pasi .F. K-means: Clustering by gradual data transformation. Pattern Recognition, 47, pp. 3376–3386, 2014.

[9] Qian. J, Lv. P, Yue. X. D, Liu .C. H, Hierarchical attribute reduction algorithms for big data using MapReduce. DOI: 10.1016/j.knosys.2014.09.001.