A Genetic K-means Clustering Algorithm Based on the Optimized Initial Centers

Min Feng
College of Information Engineering, Taishan Medical University
Taian 271016, China
E-mail:fmxxsc@126.com

Zhenyan Wang
Department of Information Engineering
Shandong Vocational Institute of Clothing Technology
Taian 271000, China

Received: March 27, 2011 Accepted: April 19, 2011 doi:10.5539/cis.v4n3p88

Abstract

An optimized initial center of K-means algorithm(PKM) is proposed, which select the k furthest distance data in the high-density area as the initial cluster centers. Experiments show that the algorithm not only has a weak dependence on the initial data, but also has fast convergence and high clustering quality. To obtain effective cluster and accurate cluster, we combine the optimized K-means algorithm(PKM) and genetic algorithm into a hybrid algorithm (PGKM). It can not only improve compactness and separation of the algorithm but also automatically search for the best cluster number k, then cluster after optimizing the k-centers. The optimal cluster is not obtained until terminal conditions are met after continuously iterating. Experiments show that the algorithm has good cluster quality and overall performance.

Keywords: Clustering, K-means algorithm, Genetic Algorithm

1. Introduction

Cluster analysis is based on "Feather flock together"thinking on classification of samples. Cluster is unsupervised learning, as compared with classification learning. Clustering objects do not have category tags, you need clustering algorithm to cluster according to certain rules. The whole data objects are divided into several different groups, so that differences within the data group are as small as possible and the gap between two groups is as large as possible.

Cluster analysis has been widely used in data mining, which has a prominent feature of handling with large and complex data sets. Clustering requires that input parameters of algorithm weakly depend on domain knowledge and algorithm is not sensitive to the order of input data, it can effectively handle with noisy data.Many different algorithms have been proposed to achieve data clustering, commonly used are K-means algorithm, STING algorithm, CLIQUI algorithm, CURE algorithm etc.K-means algorithm is a common clustering algorithm based on partition, it uses average of data objects as clustering center, so clustering process is easily influenced by noisy data, resulting the algorithm may be trapped into local optimum. To solve this problem, we propose an optimized initial center of K-means clustering method(PKM). The initial clustering centers are consistent with actual distribution of data sets as much as possible, so that clustering is not sensitive to outliers and we can obtain a higher quality cluster. The traditional KM algorithm and PKM clustering algorithm do not consider the validity of indicators---the best number of clusters. For practical problems, appropriate number of clusters is a key factor to obtain high-quality cluster. In general, a optimal cluster is cluster that distance within the class is as short as possible and class distance is maxmized while the number of class is as little as possible. To ensure effectiveness of clustering and weak dependence on initial segmentation, we present a hybrid algorithm named PGKM, combining genetic algorithm with PKM algorithm. In PGKM algorithm, genetic algorithm uses a special fitness function to find a best cluster number of k, and PKM algorithm clusters on optimized individuals selected by genetic algorithm. Both of them complement each other and promote each other, not only speeding up

convergence rate, but also improving cluster quality. Experiments show that the algorithm is effective and we can obtain more accurate cluster.

2. Algorithms

2.1 K-means Algorithm

Traditional K-means algorithm has a basic idea: Firstly, choose a certain distance as similarity measure between every pattern, determine a criteria for evaluating the quality of clustering by function, give initial cluster centers, then use iterative method to find best classification results for extreme value of criterion function. Disadvantage of the method is: (1) It may be affected by initial selected cluster centers and converge suboptimal solution prematurely; (2) It is sensitive to number of clusters, which leading to low clustering accuracy.

2.1.1 Basic Concepts

Supposing that $x = \{x_i | x_i \in R^n, i = 1,2,\cdots, n\}$ is clustering samples, number of clusters is k, and k-cluster centers are $z_1, z_2, \cdots, z_k$.

Definition 1: Calculation of cluster centers

$$z_j = \frac{1}{c_j} \sum_{i \in n_j} x_i, i = 1,2,\cdots,n; j = 1,2,\cdots,k \tag{1.1}$$

c_j is number of samples in class j and n_j represents the sample set of class j.

Definition 2: Criterion function for calculating

$$J = \sum_{j=1}^{k} \sum_{i \in n_j} \left\| x_i - z_j \right\|^2 \tag{1.2}$$

Which indicates that the smaller criterion function is, the smaller error of classification method and the better result of clustering.

2.1.2 Algorithm Description

Input: Iteratively terminal condition is ε, the number of iterating is c, the initial value is 0, the number of clusters is k and a data set containing n objects.

Output: K clusters meeting the iteratively terminal condition and c of the number of iterating.

Procession:

(1)Randomly select k objects as initial cluster centers.(k<n)

(2)For each of the remaining objects, according to its distance from every center, it would be assigned to the nearest cluster.

(3) Re-calculate every class center using formula (1.1), and calculate the value J_1 of criterion function at this time using formula (1.2).

(4)Calculating new distribution:

Supposing that x_i is in class n, if it suits to expression(1.3), x_i will be assigned to class m, and calculate the value J_2 of criterion function at this time.

$$\left\| x_i - z_m \right\|^2 < \left\| x_i - z_n \right\|^2 \tag{1.3}$$

Where, z_m is the center of class m and z_n is the center of class n.

(5) Repeated step (3) to (5) until expression(1.4) is true, otherwise $c = c + 1$.

$$\left| J_1 - J_2 \right| < \varepsilon \tag{1.4}$$

2.2 Optimized K-means Algorithm (PKM)

Traditional K-means algorithm is sensitive to initial clustering centers, which often lead to great volatility on clustering results. In general, we want to choose widely distributed data points as centers, but only consider distance factor, usually affected by noisy points, so we should consider two factors of distance and density when choosing initial points. So we get a optimized initial center algorithm based on these two factors, which choosing k points farest away from each other in high density area as initial central points. Firstly, supposing Density(x) is density of the region in which the sample point x is.

$$Density\ (x) = \{ p \in c \,|\, Dist\ (x, p) \le r \}$$

Where, Dist() is a measure of distance, r is the radius, c is sample point sets representatively. The above equation shows that density of data objects is the number of samples in a sphere, which the center is x and r is the radius. Method for determining the radius r is as follow: (1)Calculate the average distance u between two data objects; (2) Based on experience, giving a constant θ, then the radius of density regions can be calculated by r=θ*u.

Calculate density of each data object, then collect the high density points and marked with set S_A. Choose data point with highest density in S_A as the first clustering center marked with z_1,and choose data point farest away from z_1 in S_A as the second clustering center marked with z_2,and choose x_i satisfied with $\max\ (\min(\ d(x_i, z_1), d(x_i, z_2))), i = 1, 2, \cdots, n$ in data set S_A as the third clustering center, and so forth, choose x_i satisfied with

$\max\ (\min(\ d(x_i, z_1), d(x_i, z_2), \cdots, d(x_i, z_{q-1}))), i = 1, 2, \cdots, n$ in S_A as z_q.According to this method, calculate k-initial clustering centers, and apply K-means algorithm.

This algorithm avoids randomness of selection of initial clustering centers and ensure stability of the clustering results, because it takes points farthest away from each other in high-density area as initial clustering centers, so these centers are basically determined.

2.3 Genetic Clustering Algorithm

Number of clusters (k) is an important parameter to clustering quality. When the sample size is small, using exhaustive method can find the best number of clusters, but when the sample size is large, exhaustive method is almost impossible. In optimized K-means algorithm, although choice of initial clustering centers is helpful to improve clustering results at certain degree, when the sample size is large, the number of identified clusters is random, so the results can not accurately reflect actual data distribution. In order to obtain high-precision clustering results, a modified hybrid algorithm with genetic algorithm and clustering algorithm is proposed. The method uses a special fitness function of genetic algorithm. It not only can automatically search for the best number of clusters, but also can optimize clustering results obtained by PKM algorithm. Experiments show that clustering accuracy is significantly improved because of the combination with GA algorithm and PKM algorithm.

Procession:

(1)Encoding number of clusters (k). k is an integer between 1 and maximum number of categories marked with MaxClassnum that can be binary string, which is chromosome. Reference 12th has proved that the spatial clustering is optimized when $MaxClassnum \le \sqrt{n}$ (n is number of samples). For example, if number of samples is 3600, then MaxClassnum$\le$60 and length of chromosome is 8. The first six genes is the binary representation of number of categories, the seventh is decimal number corresponding to number of categories and the last one is corresponding individual fitness.

(2)Initializing population: Randomly generating population with p chromosomes, value of p is generally [30,150], crossover probability $p_c = 0.6 \sim 0.9$, mutation probability $p_m = 0.001 \sim 0.05$.

(3)Clustering: Decoding each chromosome of the group to get corresponding number of categories (k). For each individual, apply PKM clustering algorithm.

(4)Calculating fitness: E_K of the distance within class is defined as following:

$$E_k = \sum_{j=1}^{k} \sum_{i \in I_j} \left\| x_i - z_j \right\|^2 \tag{2.1}$$

Where, k is number of classifications, I_j is collection of samples of class j, z_j is center of class j.

D_k of the distance of classes is defined as following:

$$D_k = \max_{i,j=1}^{k} \left\| z_i - z_j \right\|^2 \tag{2.2}$$

Where, z_i is center of class i and z_j is center of class j.

Fitness function is defined as following:

$$f(k) = \frac{1}{k} \times \frac{1}{E_k} \times D_k \tag{2.3}$$

Where, $\frac{1}{k}$ decreases with increasing of k, then the value of fitness function will increase. The more compact of all types, the smaller E_k is and the greater $\frac{1}{E_k}$ is, so the larger fitness values f(k) is. And $\frac{1}{E_k}$ is measure of degree of system's reliability. f(k) increases with D_k and D_k is measure of degree of the system's separation.The expression(2.3) indicates that the degree of reliability and separation should be improved in case the number of categories is as little as possible. So the individual of higher degree of reliability and separation can be selected according to its fitness, which optimizing PKM algorithm.

(5) Genetic manipulation: Using "reserving elite individuals roulette algorithm" to select operators, in which selecting those individuals with high fitness value into next generation of groups with a certain probability, selecting single-point crossover operators and single-point mutation operators. After crossovering and mutating, determine whether the number of categories corresponding to individuals is less than or equal MaxClassnum, out of those individuals that number of categories is greater than MaxClassnum.

(6) Termination criterion: If the best fitness value no longer changes, terminating the method, otherwise, repeating (3) ~ (5).

Clustering algorithm has fast local searching capability and rapid convergence speed, and genetic algorithm has strong global searching capability and slow convergence speed. Combination of the two algorithms accelerates convergence speed and ensures that find global optimum at higher probability as well.

Algorithm flow chart is shown in Figure 1.

3. Comparison and Analysis of Experimental Results

To verify validity of the algorithm, we select two data sets of Iris and glass. Comparing KM algorithm, PKM algorithm and PGKM algorithm in six aspects of initial cluster centers, cluster number, iteration number, class distance, distance within the class and accuracy. The number of datas in Iris is 150, which has 4 attributions and the optimal number of clusters is 3.The number of datas in glass is 214,which has 9 attributions and the optimal number of clusters is 6. In genetic algorithm, the size of populations is P=50, crossover probability is P_c=0.8, mutation probability is P_m=0.007.

Iris and glass data sets are tested by KM and PKM algorithm in table 1.As results, KM method converges slowly, the difference of highest accuracy and lowest accuracy rate is relatively large, the average accuracy rate is low. In contrast, PKM method is better. From the table, KM algorithm's iteration number is sometimes less than PKM's, but the clustering accuracy rate is very low. This shows that in KM algorithm, the clustering result is easy to be affected by outliers because of randomly selected initial centers, clustering falls into a local optimum very quickly and clustering quality declines. In PKM algorithm, the optimized initial centers are stable and basically consistent with the actual distribution of datas, clustering results approximate to the optimal value as soon as possible, which enhancing the clustering accuracy.

Compare PKM algorithm and KM algorithm in table 2 and table 3. From the two tables, for PKM algorithm, clusters of different numbers have different classification accuracy, PKM's average accuracy rate is far below PGKM algorithm, which indicating that the improved algorithm automatically searching for optimal cluster numbers significantly improves clustering quality. PGKM combines advantages of the two algorithms PKM and GA, which not only reducing the number of iterations, but also improving the clustering accuracy.

4. Conclusion

K-means clustering algorithm is a widely used algorithm, but the arbitrariness of selecting initial cluster centers affects the instability of clustering results, this feature limits its scope of application. Although PKM algorithm eliminates the sensitivity to initial datas for traditional k-means algorithm and obtains more stable and higher quality clusters, the number of categories must be determined in the traditional algorithm and PKM algorithm, randomly determined number of clusters often results in unsatisfied results. To solve this problem, a genetic K-means hybrid algorithm based on the optimized initial centers is proposed, which guarantees the convergence rate at the same time improving the clustering accuracy. Experiments show that it is effective and feasible.

References

Agrawal R, Gehrke J, Gunopulcs D. (1998). Automatic subspace clustering of high dimensional data for data mining application.*Proc of ACM SIGMOD Intconfon Management on Data,*Seattle,WA,1998:94-205.

Bandyopadhyay S I, Jjwal Maulik U. (2002). An evolutionary technique based on K-means algorithm for optimal clustering in RN.*Information Sciences*,2002, 146:221-237. doi:10.1016/S0020-0255(02)00208-6, http://dx. doi. org/ 10.1016/S0020-0255 (02)00208-6

Guha S, Rastogi R,Shim K. (1998). Cure:an efficient clustering algorithm or large database.*Proc of ACM-SIGMOND Int Conf Management on Data*,Seattle,Washington,1998:73-84.

Guojun Mao,Lijuan Duan & Shi Wang. (2005). *Data mining principles and algorithms*].Beijing:Tsinghua University Press, 2005.

Hall L O,Ozyurt I B,Bezdek J C. (1999). Clustering with a genetically optimized approach. *IEEE Transactions on Evolutionary Computation*,1999,3(2):103-112. doi:10.1109/4235.771164, http://dx.doi.org/10.1109/4235.771164

Jingguang Fu,Gang XU&Yuguo Wang. (2004). Cluster analysis based on genetic algorithm. *Computer Engineering*, 2004,30 (4) :122-124.

Lixin Tang,Zihou Yang&Mengguang Wang. (1997). Cluster analysis using genetic algorithm to improve the K-means algorithm. *Mathematical Statistics and Applied Probability*, 1997,12 (4) :350-356.

Li J,Gao X B,Ji H B. (2003). A feature weighted FCM clustering algorithm based on evolutionary strategy. *Proceedings of the 4th World Congress on Intelligent Control and Automation*, Shanghai, China, 2003:1540-1553.

MacQueen J. (1967). Some methods for classification and analysis of multivariate observations.*Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*,1967.

Pakhiraa M K, Bandyopadhyayb S I, Jjwal Maulikc U. (2003). Validity index for crisp and fuzzy clusters.*Pattern Recognition*,2004,37:487-501. doi:10.1016/j.patcog.2003.06.005, http://dx.doi.org/10.1016/j.patcog.2003. 06. 005

Shoubao Su, Renjin Liu. (2005). Clustering techniques based on genetic algorithm for optimal data set. *Computer Applications*, 2005,25.

Shanlin Yang, Yongsen Li, etc. (2006). Optimization problem of the value in K-means algorithm. *Systems Engineering Theory & Practice* ,2006,26 (2) :97-101.

Zhongzhi Shi. (2002). Knowledge Discover. Beijing: Tsinghua University Press.2002.

Wei Wang. Jiong Yang, Muntz R. (1997). STING: a statistical informationgrid approach to spatial data mining.*Proc of the 23rd International Conference on Very Large Data Bases*, 1997.

Table 1. The results of KM and PKM algorithm

		glass data set			Iris data set		
		Initial centers	Iteration	Accuracy (%)	Initial centers	Iteration	Accuracy (%)
KM	Random	3912769	15	53.7	164553	17	60.3
	Initial	15079200	35	72.2	1303580	12	52.7
	Cluster	6598173	33	68.6	1314060	23	85.3
	Center	45207135	40	78.4	1113045	20	75.4
Average		---------	30.75	68.2	----------	18	68.4
PKM	Optimized	36115198	20	87.3	1857121	15	89.3

Table 2. Clustering Iris data set

	Clustering number	Class distance	Distance within class	Iteration	Accuracy (%)
PKM	7	23.03	75.13	13	72.15
	2	22.50	75.86	12	66.28
	4	24.37	73.51	10	88.33
	10	23.47	74.26	9	78.45
Average	---------	23.34	74.69	11	75.55
PGKM	3	25.68	72.49	5	95.76

Table 3. Clustering glass data set

	Clustering number	Class distance	Distance within class	Iteration	Accuracy (%)
PKM	5	92.16	286.76	12	80.75
	8	81.38	350.36	10	68.07
	10	105.26	273.34	25	85.36
	4	86.54	326.68	11	74.31
Average	-------	91.33	319.28	14.5	77.12
PGKM	6	113.36	256.49	8	96.45

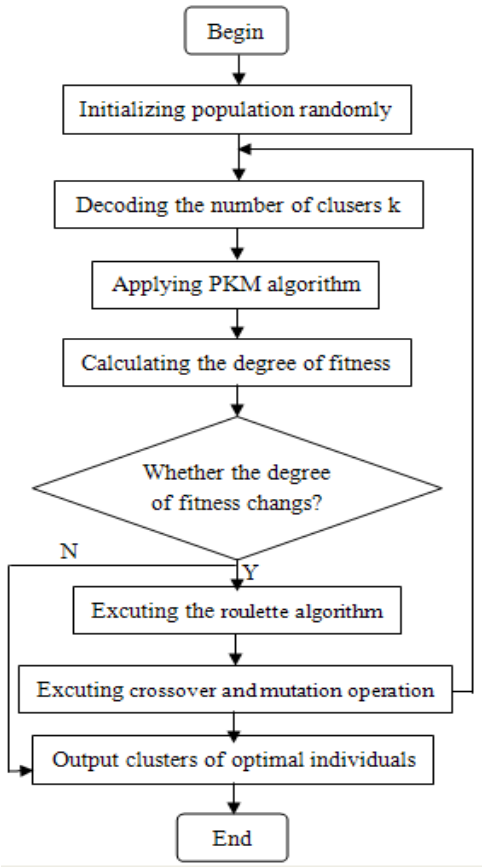

Figure 1. Algorithm flow chart

ISSN 1913-8989 E-ISSN 1913-8997