Available online at www.sciencedirect.com

SciVerse ScienceDirect

Physics Procedia 25 (2012) 1104 – 1109

Physics Procedia

2012 International Conference on Solid State Devices and Materials Science

A Clustering Method Based on K-Means Algorithm

Youguo Li, Haiyan Wu

Department of Computer Science
Xinyang Agriculture College
Xinyang, Henan 464000, China

Abstract

In this paper we combine the largest minimum distance algorithm and the traditional K-Means algorithm to propose an improved K-Means clustering algorithm. This improved algorithm can make up the shortcomings for the traditional K-Means algorithm to determine the initial focal point. The improved K-Means algorithm effectively solved two disadvantages of the traditional algorithm, the first one is greater dependence to choice the initial focal point, and another one is easy to be trapped in local minimum[1][2].

© 2012 Published by Elsevier B.V. Selection and/or peer-review under responsibility of Garry Lee
Open access under CC BY-NC-ND license.

Keywords: cluster analysis; K-Means algorithm; distance algorithm; samples of pattern

1. Introduction

Cluster analysis is based on various kinds of objects' differences and uses distance functions' regulations to make model classification [3]. Whether the classification is really make a difference or not is rest with the distribution form of pattern character vectors. If the contributions of dots of vectors is clustered and sample dots in the same group are concentrated and sample dots in different groups are distant, it will be easy to use distance functions to classify the dots, which will as far as possible make statistics in the same group be similar and statistics in different group be different. The eigenvector of the whole sample pattern congregation can be treated as dots which distribute in feature space. The distance function between dots may act as the measure of similarity of patterns. According to the proximity of dots' distance, the measure can be used to classify patterns.

In this paper we combine the largest minimum distance algorithm and the traditional K-Means algorithm to propose an improved K-Means clustering algorithm. This improved algorithm can make up the shortcomings for the traditional K-Means algorithm to determine the initial focal point. The improved

1875-3892 © 2012 Published by Elsevier B.V. Selection and/or peer-review under responsibility of Garry Lee
Open access under CC BY-NC-ND license. doi:10.1016/j.phpro.2012.03.206

K-Means algorithm effectively solved the disadvantage that the traditional *K*-Means algorithm depends too much on the selection of initial focal points.

2. *K*-Means algorithm

K-Means algorithm based on dividing [4] [5] is a kind of cluster algorithm, and it is proposed by J.B.MacQueen. This algorithm which is unsupervised is usually used in data mining and pattern recognition. Aiming at minimizing cluster performance index, square-error and error criterion are foundations of this algorithm. To seek the optimalizing outcome, this algorithm tries to find *K* divisions to satisfy a certain criterion. Firstly, choose some dots to represent the initial cluster focal points(usually, we choose the first *K* sample dots of income to represent the initial cluster focal point); secondly, gather the remaining sample dots to their focal points in accordance with the criterion of minimum distance, then we will get the initial classification, and if the classification if unreasonable, we will modify it(calculate each cluster focal points again), iterate repetitively till we get a reasonable classification.

K-Means algorithm based on dividing is a kind of cluster algorithm, and has advantages of briefness, efficiency and celerity.

However, this algorithm depends quite much on initial dots and the difference in choosing initial samples which always leads to different outcomes. What's more, this algorithm based on target function always uses gradient method to get extremum. The direction of search in gradient method is always along the direction in which energy decreases, which will leads to the fact that when the initial cluster focal point is not proper, and then the whole algorithm will easily sink into local minimum point.

3. related conception

3.1 Euclidean distance (short for distance)

Suppose that X and Z are two samples of pattern vectors, $X = (x_1, x_2, \cdots, x_n)^T$ $Z = (z_1, z_2, \cdots, z_n)^T$ and we define the distance between *X* and *Z* as:

$$D = \|X - Z\| = [\sum_{i=1}^{n} (x_i - z_i)^2]^{\frac{1}{2}} \tag{1}$$

Easy to know that the smaller D is, the more similar are X and Z (D is the distance of X and Z in n-dimensional space)

3.2 Cluster Criterion Function

The sample pattern congregation is $\{X\} = \{X_1, X_2, \cdots, X_N\}$, and we classify it to C classes, they are $S_1, S_2, \cdots, S_c$. M_j and S_j are mean vectors. So:

$$M_j = \frac{1}{N_j} \sum_{X \in S_j} X, \quad N_j = |S_j| \tag{2}$$

And N_j and S_j are the number of samples. Then we define cluster criterion function as:

$$J = \sum_{j=1}^{c} \sum_{X \in Sj} \left\| X - M_j \right\|^2 \tag{3}$$

J represents the quadratic sum of inaccuracy of all kinds of classes of samples and their mean value. We can also call it the sum of distances of samples and their mean value. So, we should try our best to get the minimum value of [6].

4. Improved K-Means algorithm

To determine the initial cluster focal point

It can't be more essential to portray the selection of the initial cluster focal point, however, usually, the selection of focal point is quite stochastic, which leads to the fact that the outcome of cluster result is also quite stochastic. While in practical applications, we not only want the initial focal points to be decentralized but also want them to be more representative. The largest minimum distance algorithm is based on probe in the field of pattern recognition. Its main thought is to choose the pattern in which pairwise distances are farther apart as much as possible to be cluster focal point. Thus, we can not only determine the best initial cluster focal point intellectually but also increase the efficiency of dividing initial data congregation. What's more, it has no possibility that the initial cluster focal points will be too adjacent, which may happen while using *K*-Means algorithm.

In this paper, at the first place, we use the largest minimum distance algorithm to determine *K* initial cluster focal points, and then we combine it with the traditional *K*-Means algorithm, at last, accomplish the classification of pattern congregation. The improved *K*-Means algorithm is obviously better than traditional one in aspects such as: the precision of cluster, the speed of cluster, stability and so on. In this paper, we adopt Euclidean distance as the criterion, because the calculation of Euclidean distance in both hyperspace and two-dimensional space are similar. So, we will use two-dimensional space as an example to analyze the improved *K*-Means algorithm in this paper.

4.1 Algorithm Description

Given *N* samples of pattern $\{x_1, x_2, \cdots, x_N\}$, which are waiting for classifying, they are. They need to be classified to *K* clusters.

1) Choose any one among $\{x_1, x_2, \cdots, x_N\}$ to act as the role of first cluster focal point z_1, for example, we choose $z_1 = x_1$

2) Choose another point which is as much as possible far apart to z_1 to be the focal point of the second cluster and calculate the distance between each sample and z_1:

$\left\| x_i - z_1 \right\|, \quad i = 1,2,\cdots,N$

If:

$$\left\| x_j - z_1 \right\| = \max \{ \left\| x_i - z_1 \right\|, \quad i = 1,2,\cdots N \}, \quad j = 1,2,\cdots N \tag{4}$$

Then choose x_j to be the focal point of the second cluster, and $z_2 = x_j$.

3) Calculate the distance between each sample among $\{x_1, x_2, \cdots, x_N\}$ and $\{z_1, z_2\}$ one by one.

$$d_{i1} = \left\| x_i - z_1 \right\|, \quad i = 1,2,\cdots N \tag{5}$$

$$d_{i2} = \left\| x_i - z_2 \right\|, \quad i = 1,2,\cdots N \tag{6}$$

Choose the minimum of the outcomes:

$\min(d_{i1}, d_{i2})$, $i = 1,2,\cdots N$

Gather the minimums of all samples of pattern and $\{z_1, z_2\}$. Choose the maximum among the minimums to be the third cluster focal point z_3.

If:

$\min(d_{j1}, d_{j2}) =$

$$\max \{ \min(d_{i1}, d_{i2}) , \quad i = 1,2,\cdots N \}, \quad j = 1,2,\cdots N \tag{7}$$

Then:

$$z_3 = x_j . \tag{8}$$

4) Suppose that we have got r (r <k) cluster focal points$\{ z_i , \quad i = 1,2,\cdots r \}$, now we need to determine the r+1th cluster focal point, namely if:

$\min(d_{j1}, d_{j2}, \cdots, d_{jr}) = \max \{ \min(d_{i1}, d_{i2}, \cdots, d_{ir}) , \quad i = 1,2,\cdots N \} \quad j = 1,2,\cdots N$

Then:

$z_{r+1} = x_j$.

5) Repeat, till $r + 1 = K$.

6) Now we have chosen K initial cluster focal point $z_1(1), z_2(1), \cdots, z_k(1)$. The numbers in parenthesis are serial numbers used in iterative operations to seek cluster points.

7) According to the rule of minimizing distance, allocate $\{x_1, x_2, \cdots, x_N\}$ to one of the K clusters, namely, if:

$\| x - z_j(t) \| =$

$$\min \{ \| x - z_i(t) \| , \quad i = 1,2,\cdots, K \}, \quad j = 1,2,\cdots, K \tag{9}$$

Then:

$x \in s_j(t)$.

The symbol t in the formula is the serial number of iterative operations, s_j stands for the j th cluster, and the cluster focal point is z_j .

8) Calculate the new vector values of each cluster focal point:

$z_j(t+1)$, $j = 1,2,\cdots, K$.

Calculate the mean vectors of samples of each cluster:

$$z_j(t+1) = \frac{1}{N_j} \sum_{x \in s_j(t)} x , \quad j = 1,2,\cdots, K \tag{10}$$

The symbol N_j in the formula above stands for the number of samples of the j th cluster s_j . Calculate the mean vectors of samples of the K clusters respectively. Making mean vectors be new clusters can minimize cluster criterion function J_j.

$$\mathrm{J_j} = \sum_{x \in s_j(t)} \| x - z_j(t+1) \|^2 , \quad j = 1,2,\cdots, K \tag{11}$$

9) If: $z_j(t+1) \neq z_j(t)$， $j = 1,2,\cdots,K$, then turn back to 7）, classify samples of pattern one by one again, and repeat iterative operations. If $z_j(t+1) = z_j(t)$， $j = 1,2,\cdots,K$， then the convergence of the algorithm is finished

5. The analysis of experiment

Aiming at testing the efficiency of the improved K-Means algorithm, we use emulational data congregation presented in table 1. The data congregation is composed of 20 random data and is classified to five classes according to the degree of cluster. We can see that the differences between each class are quite obvious. The experiment takes advantages of Visual C++ 6.0 development environment [7].

Table 1

pattern	abscissa	ordinate	pattern	abscissa	ordinate
X1	1	1	X11	1.69	0.93
X2	1.5	1.5	X12	0.3	1.1
X3	1.5	1.1	X13	7	7.4
X4	81	80	X14	6.9	6.9
X5	7.3	8	X15	22.2	20.5
X6	35.7	33.4	X16	23	21
X7	8	7.3	X17	80.6	73.2
X8	21.2	20	X18	36.7	38.55
X9	81	73	X19	34.76	33.6
X10	6.9	7.6	X20	81	73.6

Table 2

	Standard K-Means	Improved K-Means
iterations	9	6
Cluster Criterion Function J	657.603	58.3263
First class	X1,X12	X1,X2,X3,X11,X12
Second class	X5,X7,X10, X13,X14	X5,X7,X10,X13, X14
Third class	X2,X3,X11	X8,X15,X16
Forth class	X4,X9,X17, X20	X4,X9,X17,X20
Fifth class	X6,X8,X15, X16, X18,X19	X6,X18,X19

The outcomes of two kinds of algorithms are described by Table 2. We may come to the conclusion that the outcome of standard K-Means is not that good, cause its initial cluster focal points are too random, which will cause unstable cluster result. While the improved K-Means gives relatively perfect outcome.

6. Conclusion

According to the academic analysis and result of experiment above, the improved K-Means not only keeps the high efficiency of standard K-Means but also raises the speed of convergence effectively by improving the way of selecting initial cluster focal point. The improved K-Means is obviously better than standard K-Means in both cluster precision and stability. Especially, the advantages will be more obvious

when come to the point of cluster problems which have large scale and completely random distributed data.

References

[1]Usama M.Fayyad cory A.Reina Paul S.Bradley,Initialization of Iterative Refinement clustering algorithms[C].Proc.4th International Conf.On Knowledge Discovery & Data Mining,1998.

[2]DUDA R O, HART P E.Pattern classification and scene analysis[M].New York:John Wiley & Sons,1973.

[3]BianZhaoQi and ZhangXuegong Pattern Recognition Beijing Tsinghua University Press 2000.

[4]JAIN A K, DUBES R C. Algorithms for clustering data[M].New Jersey:Prentice-Hall,1988.

[5]ZhangYufang etc. A kind of improved K-means algorithm [J]. Computer Application,p31～33, 2003, (8).

[6]SELIM S Z, ISMAIL M A.K-means type algorithms: a generalized convergence theorem and characterization of local optimality [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, p81～87,1984,6(1).

[7]SunXin and YuAnping VC++ in depth detailed introduction [M] Beijing: Electronic Industry Press, 2006.