Advanced Materials Research *Online: 2014-05-28*
ISSN: 1662-8985, Vol. 951, pp 227-230
doi:10.4028/www.scientific.net/AMR.951.227
© 2014 Trans Tech Publications, Switzerland

Automatic method for determining cluster number based on silhouette coefficient

Hongbo Zhou[1,a],Juntao Gao[2,b]

[1]Northeast Petroleum University,Institute of Computer and Information Technology,Daqing,Heilongjiang

[a]jiessie9@126.com,[b]gjt@nepu.edu.cn

Keywords: Clustering,K-mean,silhouette coefficient,the sum of the squared error

Abstract. Clustering is an important technology that can divide data patterns into meaningful groups, but the number of groups is difficult to be determined. This paper proposes an automatic approach, which can determine the number of groups using silhouette coefficient and the sum of the squared error.The experiment conducted shows that the proposed approach can generally find the optimum number of clusters, and can cluster the data patterns effectively.

1 Introduction

Clustering[1] is a basic understanding of human activity.Called clustering,the data object is grouped into several classes, the principle of division is between objects of the same class have a high degree of similarity,and different classes of objects vary greatly according to the main idea of clustering algorithm,which can be summarized as:division method[2];gradation method; density-based method; grid-based method; model-based method.

Clustering is a popular data analysis technique, and in solving the clustering problem, A widely used approach is K-means algorithm and its variants.K-means is an iterative hill-climbing algorithm, but the results of K-means algorithm depend on the initial clustering centers and it ends in local optimal value.

It is difficult to tell the right number of clusters.This paper proposes an automatic method for determining cluster number based on silhouette coefficient.Experimental results show that, compared with the traditional K-means algorithm and some improvements clustering algorithm to improve the algorithm time is short.Clustering results has high silhouette coefficient, the accurate rate is higher.

2 Related Concepts

2.1 Degree of cohesion and separation

Measure the effectiveness of the cluster is generally based on the intra-cluster and inter-cluster measure two aspects,the degree of cohesion and separation is often used as a measure of the main characteristics of the clustering effect. the ideal clustering effect should be with the smallest distance inside clusters and the largest distance between clusters, which has the smallest cluster cohesion and the highest clusters separation.Cluster cohesion determining a measure of the closeness of the sample cluster, and the cluster separation determining a measure of the dissimilarity between clusters. Calculated based on the degree of aggregation and separation of the prototype of the following formula (ci is the centroid of the cluster ci, proxim ity (a, b) to achieve a sample of a proximity measure and b):

$$cohesion(C_i) = \sum_{x \in c_i} proxin\ ity(x,\ c_i)$$

$$separation(C_i,\ C_j) = prox\ in\ ity(c_i,\ c_j)$$

The degree of cohesion and separation of clusters are not independent, the sum of both is a constant equal to the total sum of squares,each square of the distance that the total sample mean and prove that literature can be found in [3],which conclusions can be inferred:Minimize cohesion degree is equivalent to maximize separation.Obviously,it is not rigorous that simple using the degree of cohesion and separation as a clustering analysis,because, in general, the more sub-clusters, the smaller the degree of cohesion, and the greater the separation. Effectiveness is mostly a function of the current proposed improvements based on a weighted combination of its degree of cohesion and separation.

2.2 Silhouette Coefficient

Silhouette coefficient is a concept proposed by Kaufman et al, which includes individual silhouette coefficient and cluster silhouette coefficient. Expression of individual silhouette coefficient is as follows:

$$a(i) = \frac{1}{n_c - 1} \sum_{i,j \in C_c, i \neq j} d(i,j) \quad (1)$$

$$b(i) = \min_{p,p \neq c}\left[\frac{1}{n_p} \sum_{i \in C_c, j \in C_p} d(i,j)\right] \quad (2)$$

$$s_i = \frac{b(i) - a(i)}{\max[a(i), b(i)]} \quad (3)$$

Here, a(i) is the average distance between data pattern i and all other patterns in its cluster, b (i) is the minimum average distance between pattern i and all the patterns in any other cluster that not containing the pattern i.

Individual silhouette coefficient si combines distance clustering inside and between clusters, to evaluate a single sample is gathered reasonableness of a class, its value between -1 and 1.For each data pattern i, si can vary between -1 and 1. When si is close to 1, this means a(i) is much smaller than b(i), so the data pattern is assigned to an appropriate cluster, the result of the cluster is good. But when si is about 0, then a(i) and b(i) are approximately equal, it is not clear that the data pattern i should be assigned to which clusters. When si is close to -1, a(i) is much larger than b(i), so the data pattern i is closer to the other cluster, the result of the cluster is bad.

Cluster silhouette coefficient expressions outline is as follows:

$$\mathrm{SC} = \frac{1}{n} \sum_{i=1}^{n} s_i \quad (4)$$

Here n is the number of data patterns in the data set. After setting the number of categories k, the SC said that the cluster silhouette coefficient, can be analyzed by cluster validity SC, for example, used to select the optimal number of clusters. Selection as follows: the number of categories for all possible strike the maximum value of k is the optimal number of clusters, while the maximum is called the optimal cluster SC profile coefficient, this time clustering clustering can be considered the best.

2.3 Sum of the Squared Error

In clustering problem, the measure of similarity is the function to determine how close two data patterns are to each other. If the measure of similarity uses Euclidean distance, then the sum of the squared error can be used as the object function of the optimization, it can be defined as:

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} d(c_i, x)^2 \tag{5}$$

Here K is the number of clusters, Ci is the i-th cluster, ci is the center of the cluster i, d is Euclidean distance between two data patterns. In order to make it cost low computation in program, in practice, the other form of the SSE is used:

$$SSE = \sum_{i=1}^{n} \sum_{j=1}^{K} d(x_i, c_j) \tag{6}$$

Here n is the number of data patterns that need to be clustered, and xi is the i-th data pattern. To a cluster result, if SSE arrives the minimum, the result is best. So we can differentiate SSE, suppose the data is in one dimension:

3 Steps of the Proposed Algorithm

K-means algorithm is simple and runs fast, but it usually gets stuck in local optimal.The new algorithm computes SC for all the clusters to obtain the optimal number. When K changes from 2 to $\lceil \sqrt{n} \rceil$, if there is a knee, peak, or dip in the plot of SSE and SC, the optimal K is obtained.

The proposed algorithm can be described as follows:

Step 1: Initialization. Set K to 2.

Step 2: The values of a (i) and b (i) for each sample point were calculated According to the formula(1) and formula(2).

Step 3: si is calculated according to the formula(3), SC is calculated according to si.

Step 4: SSE is calculated According to the formula (5) or formula (6).

Step 5: Increase K by 1 and repeat step 2- step 4.

Step 6: Compare the different value of SSE and SC upon different K. When there is a knee, peak, or dip in the plot of SSE and SC, the optimal K is obtained.

4 Experimental Results

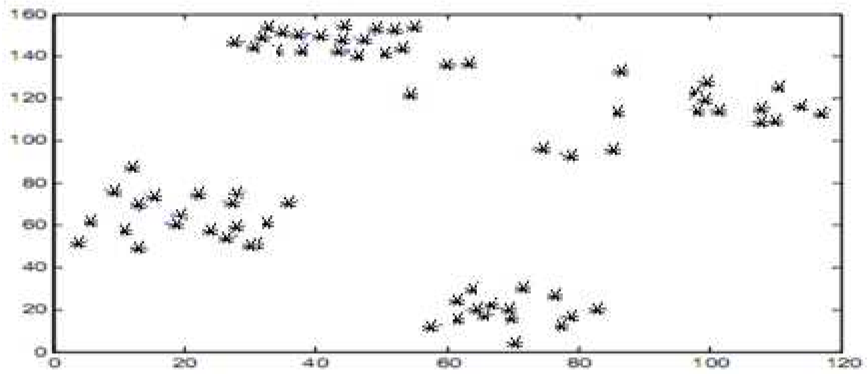
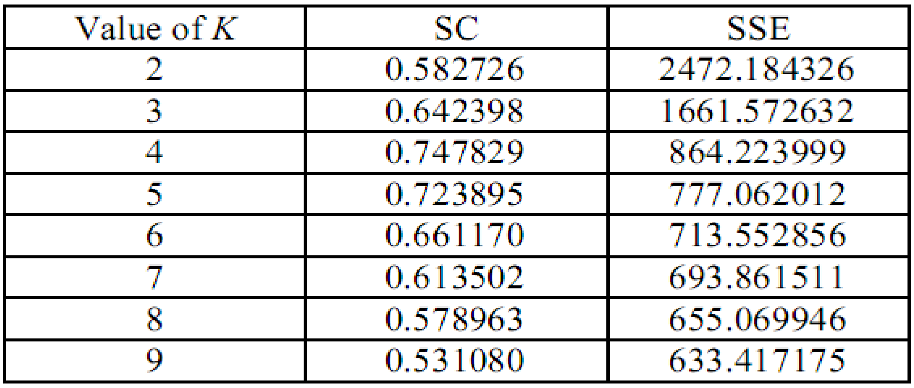


Fig.1 Ruspini data set

Ruspini data set[4] was used to evaluate the new approach. This data set consists of 4 classes, 75 data patterns, and each data pattern has two features. The data set is illustrated in Fig. 1.

The experimental results, as shown in Table 1, were averages of 10 trials.

Table 1 Results of 10 trials



Value of *K*	SC	SSE
2	0.582726	2472.184326
3	0.642398	1661.572632
4	0.747829	864.223999
5	0.723895	777.062012
6	0.661170	713.552856
7	0.613502	693.861511
8	0.578963	655.069946
9	0.531080	633.417175

For the better description of the problem, we plot the evaluation measure against the number of clusters,From the plot, we can see there are a distinct peak in SC and a distinct knee in the SSE when the number of clusters is equal to 4. when K=4,the mean silhouette value 0.7640 is the greatest.So we can conclude that this data set may consist of four clusters with the most possibility.

5 Conclusion

We have presented an automatic approach for solving the clustering problem. The new approach need not the number of clusters predefined. The experimental results show the new approach is robust and effective. The paper is supported by the Education Department of Heilongjiang province science and technology research projects (No.1253G014).

References:

[1] Wang Jicheng,Pan Jingui.Web text mining technology research[J].Computer Research and Development, 2000

[2] HAN J, K AMBER M. Data mining: concep t and technique[M]. Morgan Kaufmann Publishers, 2000 .

[3] HANDD, MANNILAH, SMYTHP.Principles of data mining [M]. Ca m bridge: MIT Press , 2001.

[4] Enrique H. Ruspini, New experimental results in fuzzy clustering, Information Sciences, Vol.6, 1973, pp. 273-284.

[5] Leonard Kaufman, Peter J. Rousseeuw, Finding Groups in Data, John Wiley & Sons,Inc.,2005.

[6] Pang-Ning Tan, Michael Steinbach,Vipin Kumar, Introduction to Data Mining, Pearson Education ,Inc., 2006.