Applied Mechanics and Materials Vols. 411-414 (2013) pp 1081-1084
Online available since 2013/Sep/03 at www.scientific.net

doi:10.4028/www.scientific.net/AMM.411-414.1081

Application of Improved K-means Clustering Algorithm in Customer Segmentation

Gang Li

School of Information Management, Hubei University of Economics, Wuhan, China

wh_lig@sina.com

Keywords: Data Clustering; K-means Algorithm; Customer Segmentation.

Abstract: Market competition is the competition for customers. By adopting customer segmentation model, decision makers can effectively identify valuable customers and then develop effective marketing strategy. Cluster analysis is one of the major data analysis methods and the k-means clustering algorithm is widely used. But the original k-means algorithm is computationally expensive and the quality of the resulting clusters heavily depends on the selection of initial centroids. An improved K-means algorithm is presented，with which K value of clustering number is located according to the clustering objects distribution density of regional space，and it uses centroids of high-density region as initial clustering center points. The proposed method makes the algorithm more effective and efficient, so as to gets better clustering with reduced complexity.

Introduction

Customer segmentation is the subdivision of a market into discrete customer groups that share similar characteristics. Customer segmentation allows a company to target specific groups of customers effectively and allocate marketing resources to best effect. Traditional segmentation focuses on identifying customer groups based on demographics and attributes such as attitude and psychological profiles[1]. Value-based segmentation, on the other hand, looks at groups of customers in terms of the revenue they generate and the costs of establishing and maintaining relationships with them.

Clustering is an unsupervised classification where there are no predefined classes. The data in the data set is assigned to one of the output class depending upon its distance to other data. The data within each class forms a cluster. The number of clusters is equal to the number of output classes. The clustering technique produces clusters in which the data inside a cluster has high intra class similarity and low inter class similarity. Clustering is mainly classified into hierarchical and partitioning algorithms[2]. Nowadays, clustering algorithms are widely used in the commercial field, such as customer analysis, and this application has achieved good effect. K-means algorithm is by far the most commonly used method for clustering.

The K-means Clustering Algorithm

K-means is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid.

The algorithm 1 is composed of the following steps:

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

As shown in Algorithm 1, the original k-means algorithm consists of two phases: one for determining the initial centroids and the other for assigning data points to the nearest clusters and then recalculating the cluster means. The second phase is carried out repetitively until the clusters get stabilized, i.e., data points stop crossing over cluster boundaries.

There is a problem which particularly troublesome, since we often have no way of knowing how many clusters exist. The major drawback of this algorithm is that it significantly sensitive to the initial randomly selected cluster centroids. It produces different clusters for different sets of values of the initial centroids. Quality of the final clusters heavily depends on the selection of the initial centroids.

Improved k-means Clustering Algorithm

Entropy. Entropy depicts the dispersal of objects belonging to the same class being merged into different clusters. According to the distribution of classes，we can calculate the entropy of each cluster by[3]

$$e_i = -\sum_{j=1}^{L} p_{ij} \, log_2 \, p_{ij}$$

While $p_{ij} = \frac{m_{ij}}{m_i}$ is the probability if objects of class j belonging to cluster i; m_i is the number of objects in cluster i; m_{ij} is the number of objects of class j in cluster i; and L is the number of classes. The total entropy of clusters is the weighted sum of each cluster's entropy, and is calculated by

$$e = \sum_{i=1}^{k} \frac{m_i}{m} e_i \tag{2}$$

where k is the number of clusters, and m is the number of objects in the data set.

Silhouette Coefficient. The silhouette coefficient is a function that measures the similarity of an object of an objects of their clusters compared to the objects of other clusters. For object *i*, the value of the silhouette coefficient is defined by

$$\mathrm{s_i} = \frac{\mathrm{b_i} - \mathrm{a_i}}{\max(\mathrm{a_i}, \mathrm{b_i})} \tag{3}$$

Where a_i is the mean of the distance between the object *i* and the objects of their clusters, and b_i is the minimum of the average distance between the object *i* and the objects in other clusters.

Description and Process of Algorithm. Algorithm 2 is as follows:

1. Classify the original data set first, and calculate the silhouette coefficients respectively based on different *k* values. Then choose the *k* value corresponding to the maximum of the silhouette coefficient as the optimal cluster number *k*;
2. Set $m = 1$;
3. Compute the distance between each data point and all other data- points in the set *D*;

4. Find the closest pair of data points from the set *D* and form a data-point set A_m $(1 \le m \le k)$ which contains these two data- points, Delete these two data points from the set *D*;
5. Find the data point in *D* that is closest to the datapoint set A_m, Add it to A_m and delete it from *D*;
6. Repeat step 5 until the number of data points in A_m reaches 0.75*(n/k);
7. If $m < k$, then $m = m + 1$, find another pair of datapoints from *D* between which the distance is the shortest, form another data-point set A_m and delete them from *D*, Go to step 5;
8. For each data-point set $A_m(1 \le m \le k)$ find the arithmetic mean of the vectors of data points in A_m, these means will be the initial centroids.
9. Repeat

Assign each data point to the cluster which has the closest centroid;

Calculate new mean for each cluster;

Until convergence criteria is met.

Algorithm 2 describes the method for finding initial centroids of the clusters. Initially, compute the distances between each data point and all other data points in the set of data points. Then find out the closest pair of data points and form a set A_1 consisting of these two data points, and delete them from the data point set *D*. Then determine the data point which is closest to the set A_1, add it to A_1 and delete it from *D*. Repeat this procedure until the number of elements in the set A_1 reaches a threshold. At that point go back to the second step and form another data-point set A_2. Repeat this till *k* such sets of data points are obtained. Finally the initial centroids are obtained by averaging all the vectors in each data-point set. The Euclidean distance is used for determining the closeness of each data point to the cluster centroids. The distance between one vector $X = (x_1, x_2, \cdots, x_n)$ and another vector $Y = (y_1, y_2, \cdots, y_n)$ is obtained as

$$\mathrm{d(X,Y)} = \sqrt{(\mathrm{x_1} - \mathrm{y_1})^2 + (\mathrm{x_2} - \mathrm{y_2})^2 + \cdots + (\mathrm{x_n} - \mathrm{y_n})^2} \quad (4)$$

The distance between a data point X and a data-point set *D* is defined as

$$d(X, D) = min(d(X, Y), where\ Y \in D) \quad (5)$$

Experimental Results. In this experiment, we choose the IRIS data set which is used specially to test clustering algorithms popularly. IRIS consists of three classes of objects: Setosa，Versicolour and virginica[4]. We compare the results of improved k-means clustering with traditional k-means clustering.

In order to observe the effects of clustering distinctly, we choose the two most sensible attributes (petal length and petal width) from IRIS attributes and place them into the two dimensional coordinate system. Then we adopt traditional and improved k-means algorithms to accomplish clustering respectively. In order to diminish the fluctuation of clustering results caused by the algorithm, we choose the best result from 10 different results generated by k-means. The improved k-means clustering algorithm just executes once. As shown in Table 1, we can contrast the entropy of clusters generated by different algorithms to compare their performances.

Table 1: Comparison between the results of k-means and improved k-means on IRIS data set

Algorithm	Entropy			
	Cluster 1	Cluster 2	Cluster 3	Weighted sum
k-means	0	0.5917	0.2668	0.3012
Improved k-means	0	0.3228	0.2539	0.1893

Application in Customer Segmentation

Marcus's segmentation process using the Customer Value Matrix first requires the calculation of the average values for the Number of Purchases and Average Amount Spent[5]. Once the average values for the axes are determined, each customer is allocated to one of the resulting quadrants. The final step is to obtain quadrant-summary-level information that begins to highlight the key differences between the resulting customer segments.

Real data set of the customer transaction details are used for the clustering algorithms. The data set consists of records of customer transaction for a period of six months form a securities firm. After the pre-process of original data, generate 2000 customer records. For each distinct party id, F is calculated as the number of his/her transaction records and A is calculated as the monthly average of his/her trading commission. Improved k-Means Clustering Algorithm are used to segment the customers for five categories, specifically see Table 2.

Table 2 The segmentation of customers using improved k-Means Clustering Algorithm

Category	Number of customers	Monthly average trading commission	Number of Trades	Proportion(%)
C1	26	12843	506	1.3
C2	92	6471	102	4.6
C3	352	2013	268	17.6
C4	1403	285	185	70.2
C5	127	56	8	6.3

As it turns out, C1 class customer belong to best Customers, which hold more fund and trade frequently; C2 class customer have a stable income and fewer transaction, is a potential good customers; C3 type of client Hold large sums of money, trading relatively frequent, belong to frequent Customers; C4 amount less class customers, trading heavily, is the general small and medium-sized retail; C5 class customers trade rarely, customer belong to sleep. The former three kinds of clients, though accounting for 23.5%, provide 77.9% of the company's profit. The experimental results basically conform to Pareto principle.

Conclusion

In CRM, the customer segmentation plays an important role in identifying the customers by grouping similar. The k-means algorithm is widely used for clustering large sets of data. But the classical clustering algorithm do not always guarantee good results as the accuracy of the final clusters depend on the selection of initial centroids. We have proposed new algorithm to cope the problems, with which K value of clustering number is located according to the clustering objects distribution density of regional space，and it uses centroids of high-density region as initial clustering center points. The results show that the improved algorithm optimizes the parameters of the algorithm, improves customer segmentation's accuracy, provides decision basis for Securities.

REFERENCES

[1]H.W. Shina, S.Y. Sohn: Expert Systems with Applications Vol. 27(2004), p. 27-33

[2]Mohamed Abubaker, Wesam Ashour: International Journal of Intelligent Systems and Applications Vol. 5 (2013), p. 39-47

[3]M. Al- Zoubi, A. Hudaib, A. Huneiti and B. Hammo: American Journal of Applied Science, Vol. 5 (2008), p. 1247-1250.

[4]Information on http://archive.ics.uci.edu/ml/datasets/Iris

[5]Marcus C: Journal of Consumer Marketing, Vol. 15 (1998), p. 494-504

Information Technology Applications in Industry II
10.4028/www.scientific.net/AMM.411-414

Application of Improved K-Means Clustering Algorithm in Customer Segmentation
10.4028/www.scientific.net/AMM.411-414.1081