Hybrid Clustering Scheme for CRM Based on Constraint Optimization RMVHC Algorithm

Qi Yin

School of Computer and Information Engineering, Hunan University of Commerce,, Changsha 410205,Hunan,China

Abstract
According to the present CRM customer information clustering analysis algorithm still has poor clustering effect in the actual operation, this paper proposes a hybrid clustering algorithm based on constraint optimization RMVHC algorithm. First of all, the method based on the examination index and point figure matrix, analysis of main factors to comprehensive comparison and to determine the final number of classification is more objective and more comprehensive. Then the CURE algorithm is introduced with the concept of information gain, to improve the CURE algorithm and then mixed with RMVHC algorithm, to improve the fine of the clustering. Finally from three aspects as the rough classification of clustering, elimination of isolated point and fine cohesion of clustering to cluster analyze hybrid clustering algorithm which based on constraint optimization RMVHC algorithm. Simulation experiments show that the proposed hybrid clustering algorithm based on constraint optimization RMVHC algorithm has a better accuracy in analysis of CRM customer information clustering, and its performance is better than the Ward's method and the constrained optimization RMVHC algorithm.
Key words: RMVHC ALGORITHM, CONSTRAINED OPTIMIZATION, HYBRID CLUSTERING, INFORMATION GAIN, CRM CUSTOMER INFORMATION CLUSTERING

1. Introduction

Customer information is the fastest growing category in the enterprise database data, such data mainly comes from the following two aspects. First of all, the accumulation from enterprise own customer information, the number of enterprise with information system and customer management system is over 70% , and 30% of the enterprises to enter the digital management integration phase with the combination of the commercial automation technology, modern communication technology and network information technology, in these systems, record the customer's personal information and consumer behavior every day, and form a vast amounts of information [1]. In addition with the evolution of marketing, enterprise information is collecting and purchasing a lot of potential customers, then form the target customer database, and do precision marketing campaign on the basis of the database [2].

The research and application of the data mining technology has a long history in worldwide, many data mining prototype system and the development of practical tools are proposed constantly, widely used in the market sales forecast, financial investment, medical research and social insurance and other fields of the CRM system [3]. KDD - R used in market analysis of telecom industry, for example, the AI center development system of the Lock Head Martin Company is applied to the auxiliary to predict the trend of the stock and changes such as the emergence of possibility [4]. And, there are a variety of tools based on the structure of client/server software in the field: IBM Intelligent Miner and SAS Enterprise Miner, SPSS Clementine, etc, they have high processing capaci-

ty, and with support of multiple platforms operating [5]. Another characteristic of this kind of data mining tool is that it usually offers a variety of data mining algorithm, support to solve the problem of wide application, and is a data mining software platform for the needs in various applications [6]. At home, the application of data mining technology in CRM system is also received extensive attention of academia. Many scholars, based on China's national conditions, developed a new data mining tools, relevant software products have been mature [7]. For example, Sadie data that provide mature data mining solutions for the financial industry can realize the function analysis of OLAP, embedded in a variety of commonly used statistical analysis methods, support for multiple relational database and OLAP Server [8]. The enterprise intelligence analysis platform Dminer Enterprisei Suit is put forward by Shuanghai Fudan DE door software company, the platform is a tool sets for data mining algorithm, integrates a variety of popular data mining algorithm, can handle various types of data sources, analyze huge amounts of data, and at the same time provide visual tools to observe and interpret the data mining results [9]. As companies growing demand for in-depth analysis of the CRM data, data mining technology will play a greater role [10].

In this paper, according to the need of the CRM data analysis, we propose a hybrid clustering algorithm based on constraint optimization RMVHC algorithm, improve RMVHC algorithm, and is hybrid clustered with CURE.

2. The comparison between the ward's method and other clustering method

The sum of squared residuals method is put forward by Ward, so people call it Ward's method [11].

n samples can be divided into k categories: $G_1, G_2, ..., G_k$. No. i sample in G_t is represented as the first as X_i^t, at this time X_i^t for p dimensional vector, n_t represents the number of samples in G_t, $\bar{X}^t$ as the center of gravity G_t, get the samples from the sum of squared residuals in G_t is shown in equation (1).

$$S_t = \sum_{i=1}^{n_t} (X_i^t - \bar{X}^t)'(X_i^t - \bar{X}^t) \quad (1)$$

And there are k categories,

$$S = \sum_{t=1}^{k} S_t = \sum_{t=1}^{k} \sum_{i=1}^{n_t} (X_i^t - \bar{X}^t)'(X_i^t - \bar{X}^t) \quad (2)$$

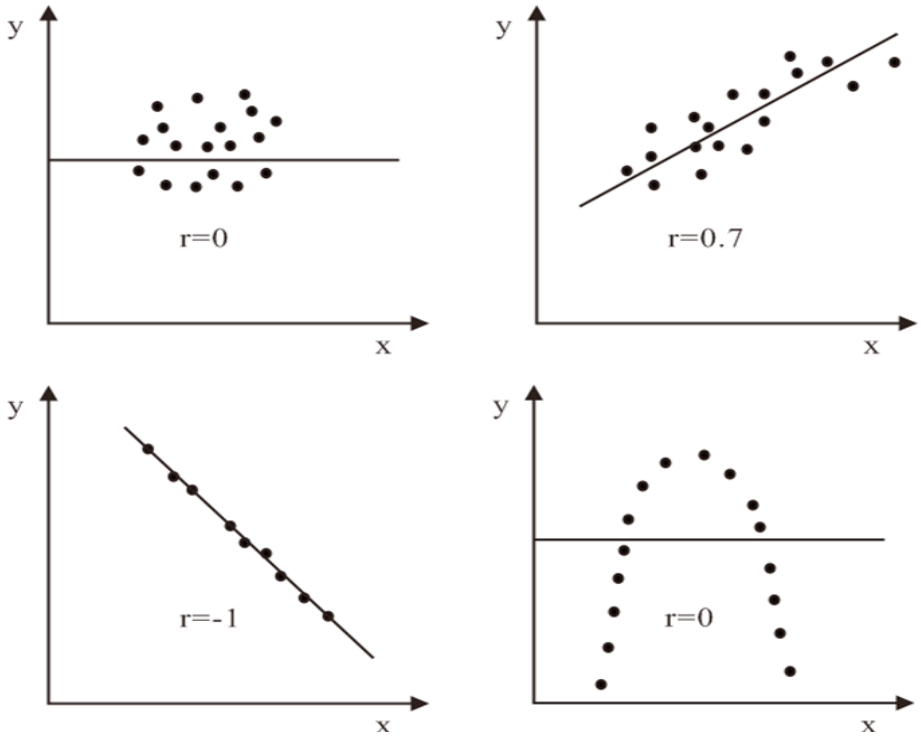

Figure.1. The geometric meaning of each factor in ward's method

General observation, can be found that the Ward's method seems to be difference to early seven methods, but in fact Ward's method and in the seven kinds of system clustering methods can be unified, the distance between G_p and G_q is defined as equation (3).

$$D_{pq}^2 = S_r - S_p - S_q \quad (3)$$

And $G_r = G_p \cup G_q$, in addition to use clustering distance of Ward's method as shown in equation (4).

$$D_{kr}^2 = \frac{n_k + n_p}{n_r + n_k} D_{kp}^2 + \frac{n_k + n_q}{n_r + n_k} D_{kq}^2 - \frac{n_k}{n_r + n_k} D_{pq}^2 \quad (4)$$

The geometric meaning of each factor in Ward's method is shown in the Figure 1.

The traditional method of clustering is mainly manifested in two different aspects: is contraction or expansion, or is properly centered. Due to the low accuracy of classify method with over contraction, maybe causes the samples together in a class but with great difference in actual; otherwise, too expansion is also lead to difficult classify, and go against to obtain practical conclusion.

3. The RMVHC algorithm based on constrained optimization

3.1. Constrained optimization

Set the number of observed as n, the number of variable as m, and the number of classes of a clustering level as G, and x_i is observed of No. i, is the current (level G) No. k class, N_k is the observed number in C_k, $\bar{X}$ is the mean vector, $\bar{X}_k$ is the mean vector in class C_k, $\|x\|$ is European length, $T^2 = \sum_{i=1}^{n} \|x_i - \bar{X}\|^2$ is the total sum of squared residuals, $W_K = \sum_{i \in C_k} \|x_i - \bar{X}\|^2$ is the sum of squared residuals in the class of class C_k, $P_G = \sum W_J$ is the sum of cluster level corresponding to all kinds of classes with the sum of squared residuals. Suppose that one step of cluster combining class C_K and class C_L to the next level class C_M, then define $B_{KL} = W_M - W_K - W_L$ as the decrease of the sum of squared residuals in class from merge result. D_{ij} reperents the distance between the two observations, D_{KI} as the distance between the class C_K and the class C_L of No. G.

Common used distance algorithm are shown as follows.

1) Minkowski distance

$$D_{ij} = \left(\sum_{k=1}^{P} |x_{ik} - x_{jk}|^q \right)^{1/q} \quad (5)$$

2) Mahalanobis distance

$$D_{ij} = (x_i - x_j)^i \sum^{-1} (x_i - x_j) \quad (6)$$

3) Portland's distance

$$D_{ij} = \frac{1}{p} \sum_{k=1}^{p} \frac{|x_{ik} - x_{jk}|}{x_{ik} - x_{jk}} \quad (7)$$

4) Oblique distance

$$D_{ij} = \left[\frac{1}{p^2} \sum_{k=1}^{p} \sum_{i=1}^{p} (x_{ik} - x_{jk})(x_{il} - x_{jl}) r_{kl} \right]^{\frac{1}{2}} \quad (8)$$

In the equation, r_{kl} is the correlation coefficient between variable x_k and variable x_l.

System clustering eventually get a clustering tree, can put all the observations together for a class. How many categories should the observation divide is a difficult problem, because the classification problem is not a certain standard, should be comprehensive analysis based on the specific situation.

According to this problem, we adopt the method of inspection index combining principal factor analysis to comprehensive comparative judgement.The specific method is: take a few more variables to clustering; determine contribution rate the larger number of variables; according to the test values in the process of clustering fluctuation situation preliminarily determine the class number;Make inspection value curve, peak according to the correct class number; as a variable scatterplot matrix and the 3D figure, find out the main factors; according to main factors of scatter plot, and ultimately determine the class number of clustering.

To determine the test index of class number, as follows.

1) R^2 statistics

$$R^2 = 1 - \frac{P_G}{T^2} \quad (9)$$

P_G is classification number of the sum of squared residuals within the total class of G, T^2 is the total sum of squared residuals of all variables. R^2 is larger, but that divided into G class from the sum of squared residuals in each class are small, which divided into G class is appropriate. However, the more clearly classification, each class is smaller, the R^2 is greater, so can only take G to make R^2 small enough, but the G is small itselfe, and R^2 no longer increases greatly.

2) A partial correlation coefficient

Classe C_K and class C_L merge into the next level of class C_M, define a partial correlation coefficient as equation (10).

$$R^2 = \frac{B_{KL}}{T^2} \quad (10)$$

In the equation, B_{KL} is the increment of the sum of squared residuals cause by merger, if partial correlation coefficient is greater, then shows these two classes should not be merged, so when the class $G+1$ into classe G should take class $G+1$ if the partial correlation coefficient is very big.

3) Bimodality coefficient

$$b = (m_3^2 + 1) / (m_4 + 3(n-1)^2 / ((n-2)(n-3))) \quad (11)$$

m_3 is skewness, m_4 is kurtosis. The value b greater than 0.555 may be indicate bimodal or multimodal marginal distribution. The maximum 1.0 only

take two value of overall.

4) False F statistic

$$F = \frac{(T - P_G)(G-1)}{P_G / (n-G)} \tag{12}$$

False F statistic evaluation is divided into G classes effect. If it is divided into G kinds of reasonable from the sum of squared residuals in class should be small, between the sum of squares is opposite bigger. So we should take the false F statistic clustering level and the class number is smaller.

5) False t^2 statistic

$$t^2 = B_{KL} / ((W_K + W_L) / N_K + N_L - 2) \tag{13}$$

With this statistic to evaluate the effect of the combination of class C_K and class C_L, the big value shows that should not combine the two classes, so we should take the level before the merger.

In traditional clustering method, there is no unified pattern to determine class number, this paper put forward based on the examination index and point figure matrix, analysis of main factors to comprehensive comparison, finally determine the number of classification method is more objective and more comprehensive.

3.2. Fine optimization based on hybrid clustering

Compared with other clustering algorithm, CURE algorithm can very good support for different size and complex clustering, and also be applied to large data sets, and in the presence of isolated point can also perform well. In view of the above advantages, we introduce CURE algorithm, and introduce the concept of information gain, to improve the CURE algorithm, and mixed with RMVHC algorithm, in order to improve the fine of the clustering.

Before precisely to define information gain, first we define a concept of entropy in information theory, it as a measure widely used, for any the purity of a data set is described. When there is a given data set S, the data collection has c different target attribute values, there are S relative to the state c of category $(c-wise)$ of entropy as shown in equation (14).

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \tag{14}$$

In the equation, the proportion belongs to the category i of in the data set S is expressed as p_i. As entropy is adopt code length by the number of bits to measure, so the logarithm base is 2 in the equation. Under the condition of the target attribute has c possible values, the maximum entropy is $\log_2 c$.

So the entropy as a standard to measure the purity of the training data set, then we use the standard to define the attribute of the measure of the ability to classify the training data standards. This metric is "information gain". Popular, an attribute of information gain is to choose the attribute partition to the data collection in order to reduce the information entropy. The information gain $Gain(S, A)$ of attribute A for a data set S can be defined as equation (15).

$$Gain(S, A) = Entropy(S) - \sum_{v \in Valite(A)} \frac{|S_v|}{S} Entropy(S_v) \tag{15}$$

In the equation, $Values(A)$ is a collection of all possible value of attribute A. S_v is the subset of value v of attribute A in data set S.

$$S_v = \{s \in S | A(s) = v\} \tag{16}$$

The first item in equation (14) is the entropy of the original data set, the second is the expected entropy after classifying data set S with A, the expected entropy is the entropy of each subset of the set ta- king the processing of a weighted sum, and weight in the original data set S, the percentage of the samples belong to the S_v is occupy $\frac{|S_v|}{S}$. So $Gain(S, A)$ can further understand that because the given value of the attribute A to get information about the objective function value.

4. The algorithm simulation

To verify the effectiveness of the proposed improved algorithm in this paper, simulation experiments on it. The customer information data of a company as an example, using the proposed hybrid clustering algorithm based on constraint optimization RMVHC algorithm for data analysis of CRM.

First of all, from the customer potential spending power, potential consumption frequency and potential index, the three aspects of clustering analysis, CRM customer information below for hybrid clustering algorithm based on constraint optimization RMVHC algorithm for four of the CRM customer cluster analysis results.

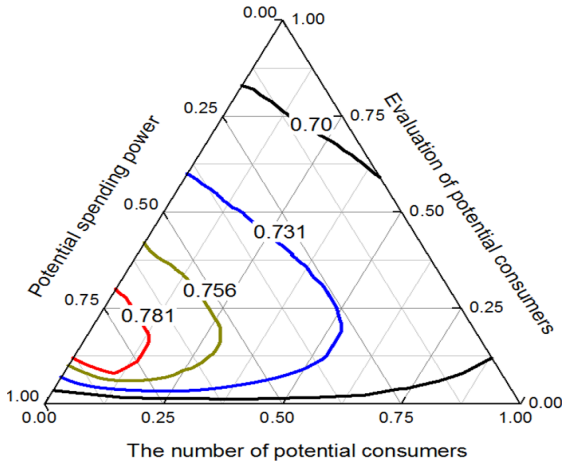

Figure 2. CRM cluster analysis

Seen from the figure, the four customer consumption potential comprehensive score are 0.781, 0.756, 0.731 and 0.781 respectively, so the first customer has the highest potential consumption ability.

Then, using the constrained optimization RMVHC algorithm to check the sum of squares of CRM customer information clustering method to cluster analysis, and compared with the actual result, test the error of clustering, the result is as follows:

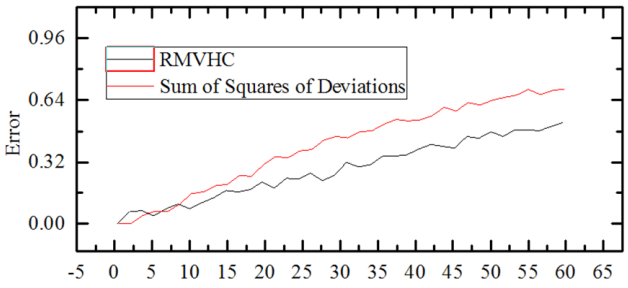

Figure 3. Comparison of the results from the constrained optimization algorithms and search RMVHC square and clustering method

Then, using the constrained optimization RMVHC clustering algorithm and hybrid clustering algorithm to cluster analysis of CRM customer information, and compared with the actual result, test the error of clustering, the result is as follows:

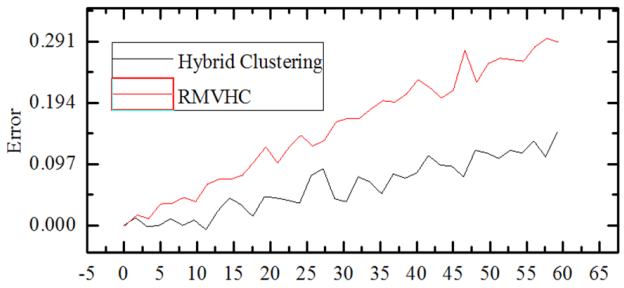

Figure 4. Constrained optimization RMVHC clustering algorithm comparison result mixed clustering algorithm

Seen from the simulation results, the proposed hybrid clustering algorithm based on constraint optimization RMVHC algorithm has a better accuracy analysis of CRM customer information clustering, its performance is better than the sum of squared residuals clustering method and the constrained optimization RMVHC algorithm.

5. Conclusion

Customer relationship management (CRM), as a means of advanced management ideas and technology, all customers such as the enterprise's ultimate customers, distributors and partners as the most important resources, in the process of customer management, found the value of its long-term profit contribution, fundamentally improve enterprise's core competitiveness, make enterprises in an impregnable position in the current fierce competition environment. In this paper, according to the need of the CRM data analysis, we propose a hybrid clustering algorithm based on constraint optimization of RMVHC algorithm. The simulation experimental results show that the improved model proposed has better analysis accuracy of CRM customer information clustering, its performance is better than the Ward's method and the constrained optimization RMVHC algorithm.

References

1. Lifang Lu (2014) Study on Logistics Supply Chain Systems Based on Analytic CRM. *Logistics Technology*, 33, p.p.394-396.
2. Yanhua Wu (2014) Application of Data Mining in CRM System in Modern Open Distance Education. *Information Science*, 32, p.p.141-144.
3. Liu Tao (2014) The Empirical Research on CRM Strategy Micro-Mechanism of City Commercial Banks. *Journal of Shanghai Lixin University of Commerce*, 28, p.p.04-112.
4. Yan Wang (2014) The Correlation Empirical research on City Commercial Bank CRM and Core Competent. *Finance and Economy,* 26,p.p.68-72.
5. Dongmei Zheng (2013) Function design of CRM system in refining company. *Computers and Applied Chemistry*, 30, p.p.1355-1358.
6. Zhiying Dai (2013) Study on Development of Logistics Information System Based on CRM. *Logistics Technology* ,32,p.p.446-448.
7. Liping Wang (2013) Applied Research on Data Mining Technology in Coal Enterprise CRM Construction. *Coal Technology*, 32, p.p. 243-244.
8. Xiaoya Shang (2013) Integration Model of Customer Knowledge Management and Customer Relationship Management: An Empirical Study of Chinese Enterprise. *International Business: Foreign Economic and Trade University*, 5, p.p. 102-111.
9. Lijun Chen (2013) CRM Research Based on Knowledge Management in Electric Power Enterprises. *Science and Technology Management Research*, 33, p.p. 155-158.
10. Humeng Yang (2013) Exploration and Application of an Analytical CRM System for Financial Enterprise. *Computer Applications and Software*, 30, p.p. 259-261.
11. Ronghui Liu (2013) Operator's Comprehensive Intelligent CRM System --Basing on the Design Case of CMCC. *Industrial Engineering and Management*, 18, p.p.117-121.