Journal of Advances in Computer Research
Quarterly pISSN: 2345-606x eISSN: 2345-6078
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 6, No. 3, August 2015), Pages: 51-63
www.jacr.iausari.ac.ir

# A Review of Different Data Mining Techniques in Customer Segmentation

**Tannane Parsa Kord Asiabi[1],RezaTavoli✉[2]**

*1) Department of Computer, PouyandeganDanesh University, Chalous ,Iran*
*2) Faculty ofDepartment of Mathematics, Islamic Azad University, Chalous Branch (IAUC)*
*17 Shahrivar Ave., P.O. Box 46615-397, Chalous, Iran*

Tannane.Parsa@gmail.com; r.tavoli@iauc.ac.ir

**Received: 2014/06/11; Accepted: 2014/09/10**

**Abstract**

*Customers are the most valuable asset of an organization. Due to high contest in the business field, it is necessary to regard the Customer Relationship Management (CRM) of the enterprise. Data Mining and Machine Learning methods been utilized by businesses in recent years in order to improve CRM. CRM is the strategy for building, managing, and strengthening loyal and long lasting customer relationship. Data mining is the knowledge discovery process by analyzing the large volumes of data from various perspectives and summarizing it into useful information. Data mining have a several techniques in CRM but in this article we present the basic classification and clustering techniques that used. The target of this survey is to provide extensive review of different classification and clustering techniques in customer segmentation.*

***Keywords:****Organization, CRM, Data mining, Customer segmentation.*

## 1. Introduction

Today investment in CRM is essential due to increasing of business competition [1]. The organizations understand that they should spend more effort in customer attraction and retention. It is difficult to keep the customers satisfied in this competitive market. Also the company found the database of current customers is a very valuable asset. Therefore the customer retention is a valuable strategy to ensure long time-term profitability and success of organization and large percentage of the profits that obtained of a small percentage of customers and the large number of customer are not profitable, then it is recommended the organizations should focus on profitable customer retention. Trend towards Customer Relationship Management developed from 1990 [1]. CRM is a process that manages the interactions between a company and its customers [2]. The basis of success in CRM is customer data and IT tools. In addition, the rapid growth of the Internet and related technologies that increased marketing opportunities. The goal of CRM is to maximize relationship with the customers over their lifetime [2].

Furthermore data mining is new technology with high potential for market activity to focus on the most important information in the data warehouse [1]. Figure 1 shows the basic of data mining process[23].

Customer Relationship Management consists of four dimensions [3, 5, 6]. This dimensions can contribute to determine a common goals for create a deeper understanding to maximize customer value against the organization in the long term. Data mining techniques can help to do this goal by extraction and identification hidden information about the customer character and manner from large database. As shown figure 2, CRM include the following dimensions:

- Customer Identification: CRM begins with the identification of customers**.** This phase includes the population that can be customer or that are profit for company. The basic concepts of this phase are target customer analyze and customer segmentation.
- Customer Attraction**:** The second phase which is the main discussion with CRM. Customers' satisfactions to the expectations the customers satisfaction with perception is essential to protect customers Jupiter preserve elements include marketing person to person pointing to the marketing person who analyze with support. This Phase companies can attract their resources to target customer. An element of customer acquisition is direct marketing.
- Customer Retention**:** The third phase and the main concern of CRM is Customer Retention. The main elements of this phase are loyalty program, One-to-One marketing and complaints marketing.
- Customer Development: These methods mainly because of the exchange lifetime customer and market basket analysis, it is raised [7].

### 1.1 Applying data mining in CRM

Data mining refers to extracting or "mining" Knowledge from large amount of data [3]. CRM is one of the most important applications of data mining. CRM has an important role in determining the strategies related to companies customers. CRM can be institutionalized with mining customer's data and its analysis. Basically the goal of CRM is establishing a deep and strong relationship with customers to increase their value and lifetime in the organization..Data mining helps companies build personal and profitable customer relationships by identifying and anticipating the needs of customers throughout the customer lifecycle [2]. It is essential to have a complete understanding of customers of company. In this concept customer loyalty is one of the most important concepts in marketing, which can improve the customer relationship management concept [4]. In fact Data Mining can provide the customer insight to generate an effective Customer Relationship Management strategy.This can lead to customer interaction and therefore increased customer satisfaction and good relationship with the customer through data analysis. Data mining process in CRM shown in figure 3:

### 1.2 Related Work

Customer segmentation is the process that divided customers into homogeneous, meaningful and closed based on various features subgroup that used for different marketing activities [8], or is the division of the entire customer population into smaller groups, called customer segmentation [9]. The symbolic model for customer segmentationthat shown in figure 4:

Segmentation also represents the key element of customer identification in CRM [20]. Effective customer segmentation depends on choosing the appropriate segmentation bases. Customers differ in their tastes, needs, attitudes, motivations, life style and so on, and all characteristics can be used to segment customers [21]. In traditional organizations they try to use of marketing segmentation scheme based on survey of individuals and the value of information regardless of industry. After several decades, the organizations decided by using simple segmentation rules and producing new products and expanding them but this strategy is not suitable in today's competitive environment. Today organizations need to have a complete perspective of their customers. Nowadays CLV and RFM models are used for customer segmentation. CLV [10] is a concept in CRM defined as the current value of customer future profit and current value is future profits that generated from customers. The RFM model [10] is the best model based on customer's features shopping. In this model R prefers to last purchase F refers to purchase frequency and M refer to purchase monetary value.

In following cases we will review several cases:

In [11], Chen has tried to examine customer behavior in using of credit cards by integration of data mining concepts and customer behavior scoring model. In this research they used SOM method for customer segmentation and then extracted rules from their job, marital status and their age to predict future customer behavior shopping.

In [12], Lee and Park that present multi agent system for profitable customers segmentation that provide control of customer satisfaction by integrated SOM, DEA and decision tree for profitable customers segmentation.

In [13], Kim.et all that present a model for build a management strategy based on customer value by using of decision tree technique for customer segmentation. This segmentation includes 3 factors: the current value, the potential value and customer loyalty.

In [14], Chen and Cheng that used the data of store transactions. First they calculated customers RFM model by using CLV model and they segmented customers by SOM algorithm. They predicted suitable marketing planning for each group and competitive advantages than obtained for stores.

In [15], Chen and Cheng quantified target variables by RFM concept and scored R, F and M variables according to importance degree and then clustered customers by using K-means algorithm and calculated WRFM value for each Cluster. Finally they extracted meaningful rules by RS theory for improved CRM strategy.

### 1.3. The proposed framework

As shown in figure 5, the main part of this documentary is customer segmentation. Customer segmentation extended process that divided customers based on internal homogeneity and personalities in groups as a distinctive marketing strategy [5].

### 1.3.1 Evaluation of data mining techniques

1.3.1.1 Classification

Classification methods control the features of new entry and assign it to a group then it is necessary to have pre-determined classes[3]. Nowadays according to competitive market, it is necessary to update available information in data base through class label. In this context we examined Decision tree, decision rules, Bayesian network, neural network, support vector machines and logistic regression. As shown in table 1, the following formula that used for evaluate the classifiers:

1.3.1.1.1**Decision Tree (DT):** A DT is a non-parametric, hypothesis-free method for classification. This algorithm is the most famous deductive algorithms, as shown in figure 6, is a data structure that can divided a large set of records to smaller sets of records and extract model from high dimensional data. In addition, understanding the knowledge by using decision tree is easy for humans [17].

1.3.1.1.1.1**CART** Algorithm **(**Classification and Regression Tree**):** This DT is a binary recursive segmentation method. First a subset of the original trees is created as a candidate model then evaluation the accuracy of this subset by using a test data set and the tree with lowest classification error is the final model. The best failure factor that obtained from expression 1 and That S is the best failure rule 2, 3, 4 and 5:

1.3.1.1.1.2$\boldsymbol{C_5}$ Algorithm: This algorithm with adaptive boosting is used to classify high- and low-value customers [21]. It is very effective in Machine Learning and similar to CART. In this algorithm branches are prunes too. But the based on prunes are quite different.

1.3.1.1.2**Decision Rules:** This technique is similar to Decision Tree and present the list of rules have a understandable statement for humans. This rules can generated multiple rules for each record that it is possible applied more one predictive rule for each record but Decision Tree generates the comprehensive rules for all records[8]. Every rule evaluated by two variable Coverage and Accuracy. In expressions 6 and 7 tuple X from labeled class the dataset D and $n_{covers}$ the number of tuples that covered rule R and $n_{correct}$the number of tuples that correctly classified by rule R and |D| the number of tuples in dataset D:

1.3.1.1.3 **Neural Network:** This technique is in order to simulate the operation of human's brain in remembering and learning [18]. These networks consist of the set of nodes that are connected. Every node have input and output and simple calculations performed in nodes based on specific functions present in figure 7.

1.3.1.1.4**Bayesian Network:** Bayesian classifiers are statistical classifier predict membership probability as a tuple belong probability to a special class [15]. The purpose of this algorithm is to create rules to help us to put further members in to the class. Bayesian classified based on Bayesian theory [16]. In Bayesian theory, D is the training data set and h is the pervious a hypothesis, thenP (h| D) is calculated as followsexpressions 8[1] and The maximum of previous (MAP)hypothesis is calculated as follows expressions 9.The purpose is chosen category that have most similarity to the input variables, or in other words, the goal is to maximize the following expression 10.

1.3.1.1.5**SVM (**Support Vector Machines**):** The SVM is a kernel-based method capable of conducting classification and regression tasks [21]. This powerful algorithm is based on statistical learning theory that focus on recognition pattern and the purpose of SVM is to find the best function to classify that recognition members of two classes. Trying to learn $f(x, \alpha)$ obtained from error of training set is calculated as follows expressions 11.

1.3.1.2 Clustering

In clustering, there is no predetermined class. This technique classified data based on similarity [3]. The titles of each group determine by user. The purpose of all clustering algorithms is minimizing the distance in clusters and maximizing distance between clusters. The Set $X = \{x_1, x_2, \dots, x_n\}$is consist of n objects, and every objects have a feature vector with length S. Suppose that we want to clustering these objects in k groups $C = \{C_1, C_2, \dots, C_k\}$that they don't have overlap, so that according to expression 12. In expression 13, the number of different modes for clustering n objects in k cluster is.Based on the above information, it is clear that if the number of clusters is identified, even recognizing the best clustering is difficult. The total number of clustering for n objects into k clusters (when k is unknown) calculated as follows expression 14.

1.3.1.2.1**Hierarchical**: hierarchical clustering is a hierarchy of clusters. First considersome measures to hierarchical decomposition of data, and in each step calculated the distance between each pair of records and the groups more closely similar to each other, thus generate by the methods of classification execution and division.

1.3.1.2.2 **K-means** algorithm: This algorithm is a flat method and based on distance [16] and used for fixed number of cluster that the number of clusters should be specified [19]. Using this algorithm that shown in a set of data in Figure 8.

1.3.1.2.3 **Two Step Cluster**: The name two-step clustering is already an indication that the algorithm is based on a two-stage approach: In the first stage, the algorithm undertakes a procedure that is very similar to the K-means algorithm. Based on these results, the two-step procedure conducts a modified hierarchical agglomerative clustering procedure that combines the objects sequentially to form homogenous clusters [22]. As its name implies, this scalable and efficient clustering model processes records in two steps. This algorithm, the optimal number of clusters can be automatically determined by specific functions.

1.3.1.2.4**SOM (**Self-Organizing Map**):** This algorithm is a popular unsupervised neural network methodology to clustering for problem solving involving tasks such as clustering, visualization and abstraction and market screening [20]. The purpose of this algorithm is to determine the cluster centers, and then assign the objects to the most similar cluster, between objects and cluster centers.**SOM** is a type of artificial neural network with two fully interconnected layers of neurons;the input layer and the output layer that shown in Figure 9.

### 1.3.2. Evaluation Table

In this part we want to evaluation data mining techniques that used in CRM strategy. Each technique has a several advantage and disadvantages that shown in table 2 and 3.

### 1.4. Conclusion

Customer Relationship Management is a technology that manages relationship with customers in order to improve the performance of business. In CRM, the customer segmentation plays an important role in identifying the customers by grouping similar customers.In this paper I tried to assemble data mining techniques that used in CRM. Data mining provides the technology to analyze mass volume of data and detect hidden patterns in data to convert into valuable information

## 2. Figures

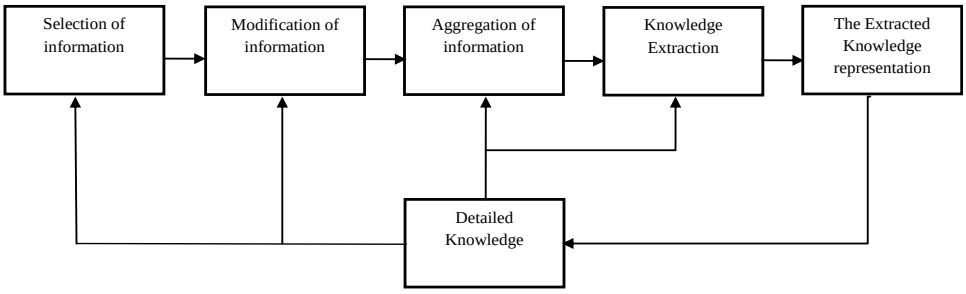


***Figure 1.data mining process[23]***

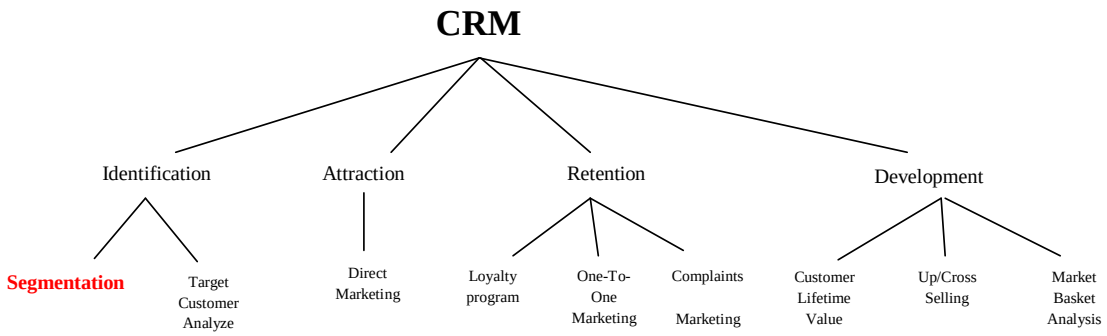


***Figure 2. Customer Relationship Management Cycle [4, 7]***

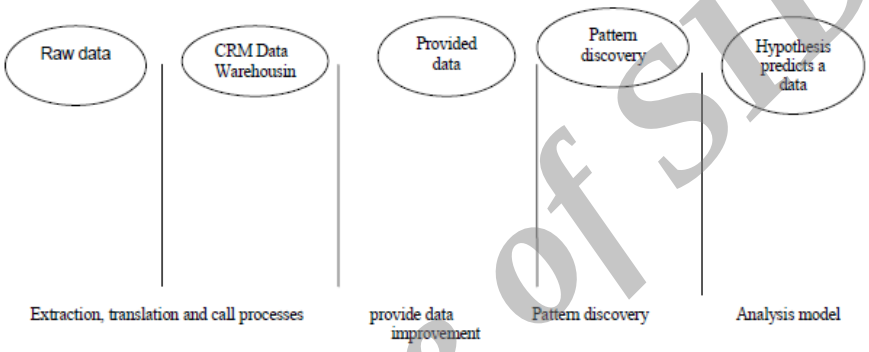


***Figure 3.Data mining process in CRM***

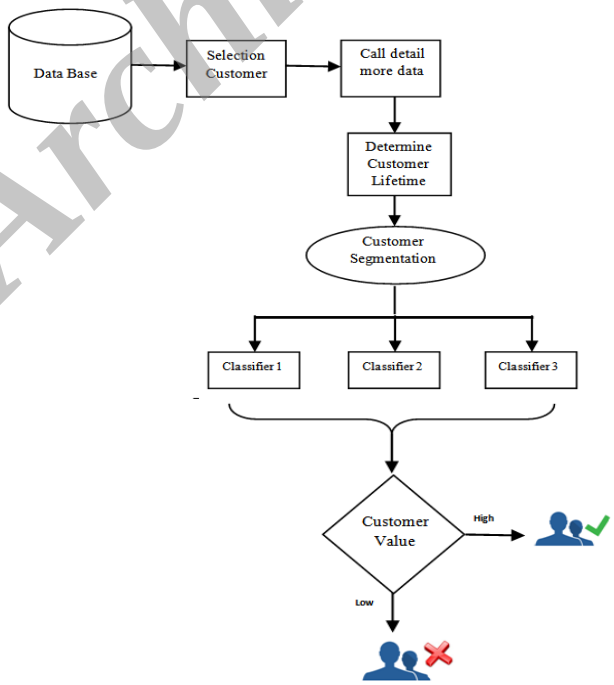

***Figure 4.The symbolic model for customer segmentation***

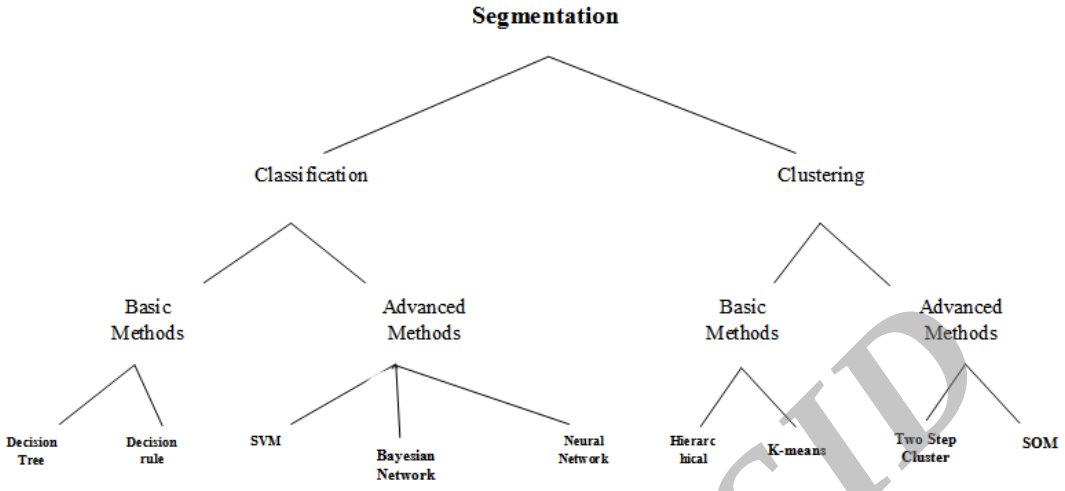


***Figure 5.data mining techniques in customer segmentation [6, 16]***

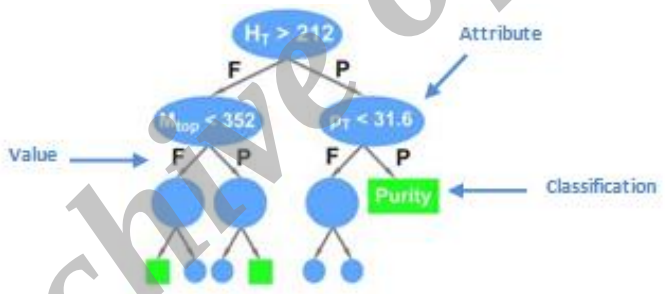


***Figure 6.simple structure of decision tree***

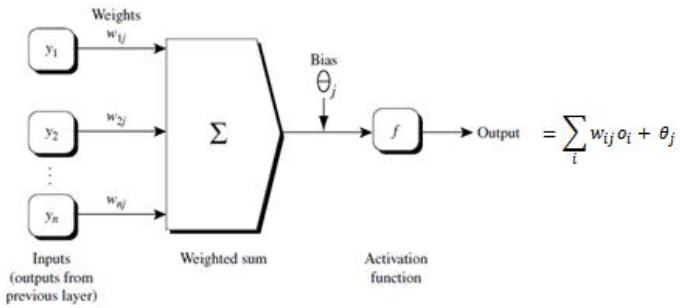


***Figure 7.simple view of neural network [16]***

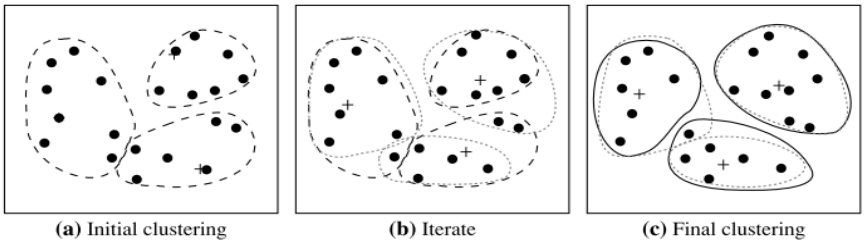


*Figure 8.clustering by K-means algorithm [16]*

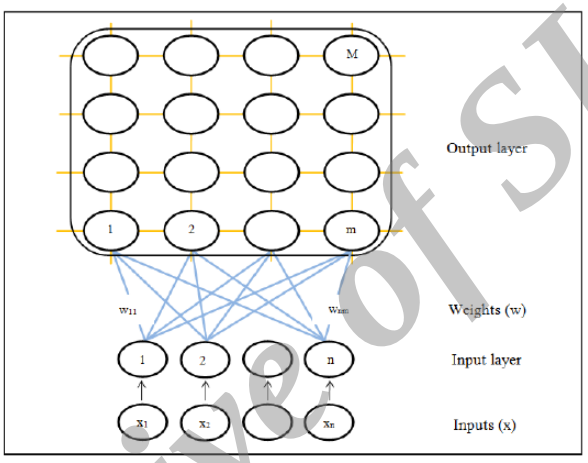


*Figure 9.the structure of Self Organization Map (SOM) [20]*

***Table 1.Evaluation the classifiers of classification techniques [16]***

| Measure | Formula |
|---|---|
| **Accuracy, recognition rate** | $\frac{TP+TN}{P+N}$ |
| **Error rate, misclassification rate** | $\frac{FP+FN}{P+N}$ |
| **Sensitivity, true positive rate, recall** | $\frac{TP}{P}$ |
| **Specificity, true negative rate** | $\frac{TN}{N}$ |
| **Precision** | $\frac{TP}{TP+FP}$ |
| **F, F−*score*, harmonic mean of precision and recall** | $\frac{2\times precision\times recall}{precision+recall}$ |
| **$F_\beta$, where $\beta$ is a non-negative real number** | $\frac{(1+\beta^2)\times precision\times recall}{\beta^2\times precision+recall}$ |

***Table 2.The evaluation classification techniques***

| *Technique* | *Advantage* | *Disadvantage* |
|---|---|---|
| *Decision Tree* | • Diversity of the simplest techniques used to the most complex methods<br>• Used to excavate abundant customers information | • due to lack of stability, stability in some of the methods are simple, yet complex and obscure methods are stable and properly functioning |
| *Decision Rule* | • Provide tangible rules for business field | • Present more than one rule for business strategy |
| *Neural Network* | • provide better strategy for business analyst | • Require numeric data for segmentation of customers |
| *SVM* | • estimate the segment of a customer based on his profile | • It appearedthat the resulting percentage of correctly classified segments was not as highas expected |
| *Bayesian Network* | • Easy implementation | • Not scalable |

***Table 3.The evaluation classification technique***

| *Technique* | *Advantage* | *Disadvantage* |
|---|---|---|
| *Hierarchical* | • Initial evaluation of customer data | • Noise sensitivity |
| *k-means* | • Suitable technique for cluster analyze<br>• The most widely use method | • The resultis not always reliable<br>• Present more than one rule for business strategy |
| *Two step cluster* | • Initial evaluation of customer data | • The result is not always reliable |
| *SOM* | • Provide tangible rules for business field | • performance alone is notsufficient to predict customer behavior |

## 3. Formulas

$$s = \max_i \left( 2 p_L p_R \sum_{i=1}^{classes} \left| p\left( \frac{j}{LeftChild} \right) - p\left( \frac{j}{RightChild} \right) \right| \right) \quad (1)$$

$$p_L = \frac{The\ total\ number\ of\ \ patterns\ in\ the\ left\ child}{The\ total\ number\ of\ \ patterns\ in\ training\ data} \quad (2)$$

$$p_R = \frac{The\ total\ number\ of\ \ patterns\ in\ the\ right\ \ child}{The\ total\ number\ of\ \ patterns\ in\ training\ data} \quad (3)$$

$$p\left( \frac{j}{LeftChild} \right) = \frac{The\ total\ number\ of\ \ patterns\ in\ the\ left\ child\ \ of\ class\ j}{The\ total\ number\ of\ \ patterns\ in\ the\ current\ node} \quad (4)$$

$$p\left( \frac{j}{RightChild} \right) = \frac{The\ total\ number\ of\ \ patterns\ in\ the\ right\ child\ \ of\ class\ j}{The\ total\ number\ of\ \ patterns\ in\ the\ current\ node} \quad (5)$$

$$\mathrm{cov}erage(R) = \frac{n_{covers}}{|D|} \quad (6)$$

$$accuracy(R) = \frac{n_{correct}}{n_{covers}} \quad (7)$$

$$P(h|D) = \left( \left( P(D|h) P(h) \right) / \left( P(D) \right) \right) \quad (8)$$

$$h_{MAP} \equiv \arg\max P(h|D)_{h \in H} = \arg\max P(D|h) P(h)_{h \in H} \quad (9)$$

$$P\left( C = c \,|\, A_1 = a_1 \cup \ldots A_n = a_n \right) \quad (10)$$

$$R_{emp(a)} = \frac{1}{m} \sum_{i=1}^{m} l\left( f(x_i, a), y_i \right) \quad (11)$$

$$C_1 \cup C_2 \cup \ldots \cup C_k = X \ , C_i \neq \varnothing, \ and \ \ C_i \cap C_j = \varnothing \ \ for \quad i \neq j \quad (12)$$

$$NW(n,k) = \frac{1}{k!} \sum_{i=0}^{k} (-1)^i \binom{k}{i} (k-i)^n \quad (13)$$

(14)

$$\sum_{k=1}^{n} NW(n,k)$$

## References


[1] Ms. Neethu, Mrs. Priyanka L.T, "Customer Classification And Prediction Based On Data Mining Technique," International Journal of Emerging Technology and Advanced Engineering (IJETAE), 2012.

[2] Manigamdan .E,Shanthi .V, "Study of Improving the Customer Relationship Management By Data Mining Applications," International Journal of Business Information Systems, 2012.

[3] S. Janakirman, K. Umamaheswari, "A Survey on Data Mining Techniques for Customer Segmentation for Customer Relationship Management," International Journal of Engineering Business and Enterprise Application (IJEBEA), 2014.

[4] Cristian J.Figueroa, "Predicting Customer Loyalty Labels in Large Retail Database: A Case Study in Chile," Department of Electrical Engineering, 2010.

[5] E.W.T. Ngai, Li Xiu, D.C.K. Chau, "Predicting Application of Data Mining Techniques in Customer Segmentation in Customer Relationhip Management: A literature review and classification," Expert System with Application, 2009.

[6] Mohammad BehrouzianNejad, EbrahimBehrouzianNejad and Ali Karami, "Using Data Mining Techniques to Increase Efficiency of Customer Relationship Management Process," Research Journal of Applied Science, Engineering and Technology, 2012.

[7] KeyvanVahidyRodpysh, Amir Aghai, MeysamMajidi, "Applying Data Mining in Customer Relationship Management," Informational Journal of Information Technology, Control and Automation (IJITCA), 2012.

[8] KonstantinosTsiptsis, AntonisChorianopoulos, "Data Mining Techniques in CRM: Inside Customer Segmentation," Wiley, 2011.

[9] JaideepSrivastava, Jau-Hwang Wang, Ee-Peng Lim, and San-Yih Hwang, ,"A Case for Analytical Customer Relationship Management," International Journal of Business Intelligent, 2002.

[10] Beomoo Shim, KeunhoChio, YongmooSuh, "CRM strategies for a small-sized online shopping mall based on association rules and sequential patterns," Expert System with Applications, 2012.

[11] Nan-Chen Hsieh, "An integrated data mining and behavioral scoring model for analyzing bank customer," Expert System with Applications, 2004.

[12] Jang Hee Lee, Sang Chan Park, "Intelligent Profitable Customers Segmentation System based on Business Intelligence Tools,"Expert System with Applications, 2005.

[13] Su-Yeon Kim, Tae-soo Jung, Eui-Ho Suh, Hyun-Seok Hwang, "Customer Segmentation and Strategy Development based on Customer Lifetime Value: A Case Study,"Expert System with Applications, 2006.

[14] Chia-Cheng Shen, Huan-Ming Chuang "A Study on the Applications of Data Mining Techniques to Enhance Customer Lifetime Value,"WSEAS TRANSACTIONS on INFORMATION SCIENCE and APPLICAATIONS, 2009.

[15] Ching-Hsue Cheng, You-Shyang Chen, "Classifying the segmentation of customer value via RFM model and RS theory,"Expert System with Applications, 2009.

[16] Jiawei Han, "Data Mining Concept and Techniques Third Edition,"University of Illinois at Urbana Champaig, 2009.

[17] Raj Kumar, D.r Rajesh Verma, "Classification Algorithms for Data Mining: A survey,"International Journal of Innovations in Engineering and Technology (IJIET),2012.

[18] Shu-Hsien Liao, Pei-Hui Chu, Pei-Yuan Hsiao, "Data Mining Techniques and Applications- A decade review from 2000 to 2011,"Expert System with Applications,2012.

[19] Narander Kumar, Vishal Verma, VipinSaxena, " Cluster Analysis in Data Mining using K-Means Method,"International Journal of Computer Applications ,2013.

[20] VahidGolmah, "A Case Study of Applying SOM in Market Segmentation of Automobile Insurance Customers," International Journal of Database Theory and Application,2014.

[21] Zhiyuan Yao, "Visual Customer Segmentation and Behavior Analysis," TUEKU CENTER for COMPUTER SCIENCE,2013.

[22] Erik Mooi, Marko Sarstedt, "A Concise Guide to Market Research: The Process, Data, and Methods Using IBM SPSS Statistics Hardcover,"2011.
[23] LuanJing ,"data minig as driven by Knowledge management in higher education," Public conferenceUCSF, 2001.