# Design and Implementation of Analytical CRM System Based on Improved Decision Tree

Lan Wang[1, a] and Wenli Gan[1]

[1]College of Information Technology, Luoyang Normal University, Luoyang, 471022, China

[a]cdlichuangdong@163.com

**Keywords:** Decision tree; Customer Relationship Management; Association rule; Classification; ID3

**Abstract.** Decision tree is a popular classification algorithm, which has the characteristics of fast learning, high classification accuracy, and the performance of the classification results. Decision tree structure includes two stages: the training set to generate decision tree; the generated decision tree pruning. This paper improves the traditional ID3 decision tree algorithm, and then puts forward the design and implementation of analytical CRM system based on improved decision algorithm. Finally, the experimental results show that the accuracy of the CRM system based on the improved decision tree algorithm is significantly improved, and it is stable.

## Introduction

Customer Relationship Management (CRM) is also produced. CRM has become an important foundation for enterprises to gain competitive advantage. Foreign CRM has been regarded as a magic weapon to win the competition in the new century; many well-known international companies have fully launched the CRM, and received a significant effect. Similarly, China's enterprises are facing the marketing environment is also undergoing tremendous changes in the economic environment, technological environment, consumer behavior and other aspects of the show a new trend [1]. CRM across the different disciplines, and has almost been applied to all sectors of many companies in different industries.

Decision tree method is one of the commonly used methods for data classification [2]. Compared with other classification methods, the decision tree method has the following advantages: (1) the speed is fast, the computation is relatively small, and it is easy to be converted into classification rules. As long as the roots go down to the leaves, the splitting conditions along the way will be able to determine only a classification of the rules. (2) The accuracy is high. Compared with other classification methods, the classification rules mining by decision tree is more accurate and can guide people's decision.

The decision tree based on ID3 algorithm has good performance and can automatically generate classification rules; here the application of ID3 algorithm to classify the basic customer information, analysis of the characteristics of customer loss, customer retention rate is enhanced by this model, the method of decision tree based on the classification of customer files. In this paper, data mining the number of decision method based on improved design of CRM system, the system is divided into the first group of customers preferred customers (Preferred customers) and the general customers preferred customers for those of the company's most valuable customers, then the application of decision tree classification according to the characteristics of customers, the high value customer recognition, in order to achieve the high value customer retention objective. In order to overcome the inherent shortcomings of the decision tree, the accuracy and the interpretation of the customer churn prediction model can be improved.

## An Improved Decision Tree Algorithm for Mining Association Rules

Mining association rules is the value of the relationship between the data items (items) from a given data set. Decision tree technology has been the focus of many data mining systems, many companies at home and abroad have introduced their own data mining systems, most of which are based on the decision tree method. The more mature decision tree algorithms are ID3, C4.5, CHAID, CART, PUBLIC, SLIQ, SPRINT, etc. C4.5 algorithm is QUINLAN I proposed an improved algorithm for the ID3 algorithm.

Decision tree learning is an inductive reasoning algorithm based on example learning, which is based on the rules of the form of representation of decision tree from a set of non - order and non - regular instances [3]. Traditional decision tree algorithm can be described as a recursive process: first of all, choose a property of the training samples as a node, the attribute of each possible value to create a branch, and accordingly the training samples are divided into several subsets [4].

The origin of the decision tree method is the concept learning system (Learning System Concept, CLS), and then the ID3 method is developed to become the peak. ID3 algorithm proposed by Quinlan in 1986 by learning to generate a decision tree for an example set [5]. In addition, all the properties of the hypothetical example are discrete.

Here through the analysis to improve the following ideas:

Step1: From the root node to every one of the paths to the leaf nodes corresponding to a set of properties test of conjunctive rule, the decision tree on behalf of instance attribute value constraints of conjunctive disjunctive expression rules [6]. This is very easy to convert into THEN - IF form of classification rules, according to the classification rules can be relatively easy to identify and predict unknown data objects, as is shown by equation (1).

$$\begin{aligned} P^{(\beta)}(m|m) &= E\left\{\left[X(m)-\hat{X}^{(\beta)}(m|m)\right]\left[X(m)-\hat{X}^{(\beta)}(m|m)\right]^{T}\right\} \\ &= W_X^{*}\overline{P}^{(\beta)}(m|m)W_X \end{aligned} \tag{1}$$

**Step2:** Because of the complete decision tree to describe the characteristics of the training samples is too precise, can not achieve a reasonable analysis of the new samples, so it is not an analysis of the new number of the best decision tree.

This article in analysis of ID3 and C4.5 algorithm improved by information gain rate to select attributes, overcome with the information gain to select properties tend to attribute values are selected, avoiding the tree height control growth and avoid over fitting the data. The formula used in the method is as follows.

$$k(P,D)=\frac{card(POS_{P}(D))}{card(U)}=\frac{card(\sum_{i=1}^{m}\underline{p}X_{i})}{card(U)} \tag{2}$$

A father node is decomposed in order to K sub nodes. Nodes in the N representation of the number of samples, I (n) said that the n samples, according to the classification of tags to get the entropy =I (n) I (C1, C2,... CH), which CI said n samples belonging to the sample number of CI classification. This formula is represented by a 3 gain attribute information for I (n) - ni/n*I (Ni).

$$\sum_{i=1}^{2}\mu_{p_8}(Xi)=\mu_{p_8}(X1)+\mu_{p_8}(X2)=Bnd_{p}(X_{i}) \tag{3}$$

ID3 algorithm can only construct decision tree for data set which can describe attributes as discrete attributes. The traditional ID3 algorithm selects the attribute A as the test attribute principle: the formula (2) is the smallest. This algorithm tends to be more in the choice of the value of the property, but the value of the property is not always the most optimal attribute. That according to the principle of minimum entropy is shown by equation (4). ID3 algorithm is selected properties; the test does not provide too much information. So, we introduce the attribute importance to improve the fusion degree of the algorithm [7].

$$M:=\frac{\eta_j\hat{v}_X^2(\tau_j)}{v_X^2(\tau_j)}=\chi_{\eta_j}^2 \tag{4}$$

Among them, S1 to Sc is the C A of the different values of the attributes of S and the formation of a subset of the C. If in accordance with the attribute S the A set (including 30 cases) is divided into 10 use cases and two sets of 20 SplitInfo (S,) =-1/3*log (A) -2/3*log (1/3) (2/3).

**Step1:** if the "x, C (x)" is a sample set S in a training instance, but its attribute A value A (x) unknown [8]. One of the strategies to deal with missing attribute values is the most common value assigned to the training instance corresponding to the node n;

**Step2:** attribute generalization, and the concept of threshold control is used to make generalization of the upper layer and lower layer along the attribute concept;

**Step3:** a set of attributes, C is a set of conditional attributes, D is a decision attribute set, the attribute importance is defined as:

$$SGF=(\alpha,D)\,\gamma\,(C,D)-\gamma\,(C-\{\alpha\},D) \tag{5}$$

**Step4:** Narrowing the scope of the negative number: it is worth noting $\sim \forall x P(x) \Leftrightarrow \exists x \sim P(x) \sim \exists x Q(x) \Leftrightarrow \forall x \sim Q(x)$;

**Step5:** IF (a1=3) AND (a4=4) AND (a5=4) THEN d=1.

**Step6:** When $CSD(p,C,D)=0$, $K(P,D)=0$ or $\sum_{i=1}^{m}\mu_p(Xi)=0$;

If $K(P,D)=0$, Then it shows that the attribute P has no influence on the classification;

**Step7:**In decision tree induction method, commonly used information gain method to help determine generated for each node should select the appropriate attributes, so you can choose the attribute with the highest information gain (to reduce the entropy maximum) as the test attribute for the current node.

The improvement of traditional ID3 algorithm and C4.5 is based on the importance of attributes, to improve the selection criteria [9]. Through the weight of formula (4) and increasing the importance of attribute, the label of attribute is strengthened, and the label of non important attribute is reduced ". In this way, the formation of a decision tree, the value of a small number of properties will not be submerged, and ultimately make the decision tree to reduce the occurrence of large data cover up small data.

## Design Key and System Structure of Analytical CRM System

Here the analysis of CRM system three layers of B / S structure based on ASP. Net, under the environment of Microsoft Visual Studio 2013 in language C#, HTML and JavaScript language supplement the code, using Microsoft SQL Server 2008 SqlExpress as database system, ensure the data processing and data access process of efficient, safe; the three-tier architecture to manage code and to isolate the user layer and data layer, provides a reliable guarantee for the project management and the later maintenance.

A database is a collection of related data organized according to a certain structure and rules, and is a repository for storing data. Web database technology of the system uses three layers of architecture; the front desk uses the browser technology based on client, through the network server and middleware to access the back-end database (see Fig. 1).

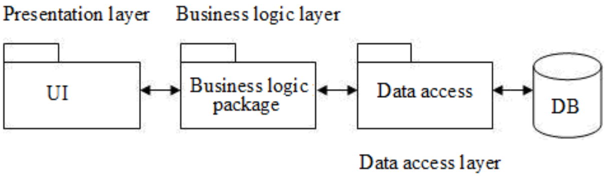

Figure 1. Three layer architecture of analytical CRM system

Depending on the data source, the commonly used Net data provider can be divided into 3 categories: Server SQL data provider, DB OLE data provider and ODBC compatible data source provider [10]. All data source providers are located under the System. Data command space. Each Net data provider has 4 main components. Their functions are as follows: Connection objects: used to connect to the database. Command object: the INSERT, UPDATE, and DELETE commands that are used to perform the data source. DataRead object: a connected forward read-only data set DataAdapter object: used to generate a DataSet from the data source, and to update the data source.

## Experiments and Analysis

The analysis type CRM system development process using a structured system development method, first of all, the "from top to bottom and layer by layer decomposition" development strategy and through the structured design method, based on the improved mining enterprises analysis type CRM management system is divided into internal management system to publicize the company's external website and business.

The functions of each part are described as follows:

(1) In front of the main include: news browsing, product display, online orders, visitors message, bearing knowledge, product search, brand agents, product classification, static display

(2) Customer management: customer information, contact, customer data transfer, customer churn management.

(3) Sales management: sales opportunities, sales orders, online orders.

(4) Customer service management: customer care, customer service records, complaints handling.

(5) Inventory query: warehousing, inventory, inventory list, warehouse management.

(6) Maintenance and management: failure notification, maintenance of the library certification, maintenance report.

(7) Information audit: customer information, dealer information, sales orders, failure notification, maintenance reports, maintenance of a library to prove.

(8) Decision analysis: customer value analysis, sales opportunity analysis, sales order analysis, employee sales analysis, product analysis, customer churn analysis.

In this paper, by using the improved decision tree to carry on the customer churn analysis, the need to deal with the sample (corresponding to the root node) or the sample subset (corresponding to the subtree) are sorted in accordance with the continuous variable size from small to large, assuming that the attribute corresponding to the different attribute values.

With the improved algorithm to construct decision tree, can first calculate each attribute important degree of attributes, if the same, then this case attributes did not differ in the classification ability, you do not need to used to improve; if different, improved ID3 algorithm to construct decision tree.

Decision tree algorithm to increase the depth of each branch of the tree until exactly the training sample can be compared to the perfect classification. In practical applications, the strategy may be difficult when there are too few samples of noise or training in the data to produce a representative sampling of the objective function. At the time of the occurrence of the above, this simple algorithm produces trees will transition fitting the training samples. Experiment according to the above steps selected completed analysis of CRM data, based on improved decision tree algorithm and ID3 algorithm to obtain the decision results and the database. By comparing the experimental results with the ID3 algorithm, the accuracy of the algorithm is improved obviously, and the stability is also improved, as is shown by Fig. 2.

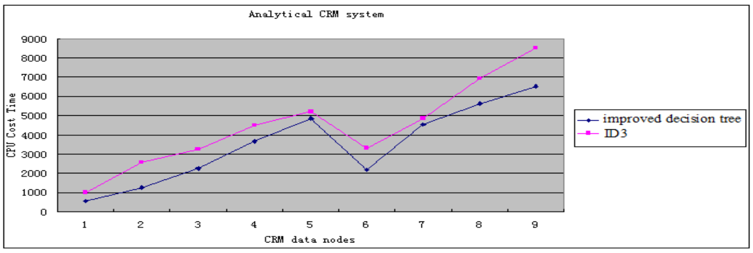

Figure 2. Design and implementation of analytical CRM system based on improved decision algorithm

Below through the experiment to verify the comparison of two algorithms for constructing decision tree we can see that the improved decision tree algorithm to generate the decision tree training model, and then the test set validation rules, than the traditional ID3 algorithm are more able to get effective customer churn rules.

## Summary

This paper improves the traditional decision tree algorithm, and is used to analyze the CRM system, based on the data mining method and the technology of customer churn prediction model. This system is based on the Asp.Net three layers B/S structure, in the Visual Studio Microsoft 2010 environment to C# language based, HTML, JavaScript and other languages for the preparation of the code. Mainly divided into external web site and the internal management system are to deal with the business. Among them, the external web site is divided into foreground and background.

## Acknowledgements

This paper is supported by Scientific and technological projects of Henan Province in China (142102310482).

## References

[1] A.Kusiak, J.A.Kern, K.H.Kernstim, and B.T.L.Tseng, Autonomous Decision-Making A Data Mining Approach. IEEE Transactions on Information Technology in Biomedicine vol.4,No.4,pp.274-284,2000

[2] WANG Cuiru, OU Fangfang. An Algorithm for Decision Tree Construction Based on Rough Set Theory. 2008 International Conference on Computer Science and Information Technology (ICCSIT 2008), Singapore: IACSIT, 2008: P295-298.

[3] Harold Buko Dadye, Richard Rimiru, "Effects of Different Pre-processing Strategies: A Comparative Study on Decision Tree Algorithms", JDCTA, Vol. 7, No. 7, pp. 939 ~ 948, 2013.

[4] Guang-xian Ji, "The research of decision tree learning algorithm in technology of data mining classification", JCIT, Vol. 7, No. 10, pp. 216 ~ 223, 2012.

[5] Siavash Emtiyaz, MohammadReza Keyvanpour, "Customers Behavior Modeling by Semi-Supervised Learning in Customer Relationship Management", AISS, Vol. 3, No. 9, pp. 229 ~ 236, 2011.

[6] Sudheep Elayidom.M, Sumam Mary Idikkula, Joseph Alexander, "Design and Performance analysis of Data mining techniques Based on Decision trees and Naive Bayes classifier For", JCIT, Vol. 6, No. 5, pp. 89 ~ 98, 2011.

[7] Manisha Rathi, Anand Priyadarshini, Ankit Rastogi, "Predictive Analysis for Customer Relationship Management", JDCTA, Vol. 4, No. 2, pp. 95 ~ 99, 2010.

[8] He Bing, Liu Gang, Wang Yuanyuan, Gao Jiang, Wang Hong, "Cooperative Task Planning Using Improved Decision Tree Algorithm", JCIT, Vol. 6, No. 6, pp. 65 ~ 72, 2011.

[9] Lu Jiang, "The Application of Data Mining Technology in the Customer Relationship Management Based on the Customer Loyalty Strategy", IJACT, Vol. 5, No. 7, pp. 89 ~ 96, 2013.

[10] Hyun Gi Hong, Je Ran Chun, "The Performance of Customer Relationship Management System: antecedents and consequences", JCIT, Vol. 8, No. 12, pp. 385 ~ 390, 2013.