International Journal of Smart Home
Vol. 10, No. 1, (2016), pp. 251-266
http://dx.doi.org/10.14257/ijsh.2016.10.1.23

Application of Improved Decision Tree Method based on Rough Set in Building Smart Medical Analysis CRM System

Hongsheng Xu*, Lan Wang and Wenli Gan

Luoyang Normal University
Henan Luoyang, 471022, China
** xhs_ls@sina.com*

Abstract

Medical Customer Relationship Management (CRM) is a kind of study method for the patient and potential patient carries on the exchange, timely access to and convey information, tracking to give the necessary guidance. The purpose of community hospital CRM is the daily business management and decision analysis of the hospital with the relationship between doctors and patients. Decision tree learning is an inductive learning algorithm based example. Rough set theory is used to process uncertain and imprecise information. In this paper, a decision tree algorithm based on rough set is proposed, and the improved decision tree algorithm based on rough classification is better than the standard C4.5 algorithm in classification accuracy and regression rate by experiment. Finally, the improved decision tree method is applied to the smart medical analysis CRM system. The experimental results show that the method can improve the management efficiency of CRM.

Keywords*: Customer Relationship Management; Decision tree; Rough set; Smart medical treatment; Attribute reduction*

1. Introduction

Customer Relationship Management (CRM) is the customer relationship management. From the word meaning, refers to the relationship between the management and the enterprise CRM customers. CRM is customer value and its relationship to a business strategy selection and management, CRM requirements to the customer as the center of the business philosophy and corporate culture to support effective marketing, sales and service process. If the enterprise has the correct leadership, strategy and enterprise culture, CRM application will realize the effective customer relationship management for the enterprise.

State Council in 2009 health care reform ideas proposed "establish and perfect the basic medical and health system covering both urban and rural residents, long-term goals for the masses to provide a safe, effective, convenient and affordable medical and health services", a few years later, medical reform gradually on the right track, it not only affects the with the public nature of the state owned medical units, but also bring great impetus to the development to the community health care. To solve the problem of the common people, the doctor, the expensive, the role of community health care is not negligible.

Due to the lack of relevant systems and the lack of funds, resulting in the medical community is not optimistic reality, such as poor medical hardware, community medical institutions practitioners personnel vacancies, the masses lack of trust on the community health care, neither do the "Wumart", and doesn't do it "cheap", community medical care and basic medical insurance of disconnection, community medical institutions lack of effective supervision and so on. However, China should take all measures to develop

* Corresponding Author

ISSN: 1975-4094 IJSH
Copyright ⓒ 2016 SERSC

community medical treatment to resolve the problem of "hard medical treatment and expensive medical treatment". One way is to apply computer information technology to community health care, so that it can meet the needs of more residents in the limited funds.

With the reform of medical system, community hospital is becoming the important barrier for the health of community residents, and becomes an important part of community life. It is responsible for the treatment of common diseases of community residents, and the prevention of common diseases. The relationship between the community hospital and the residents it serves is gradually deepening [1]. The goal of customer relationship management system in community hospital is to make the relationship between doctors and patients, and manage the other tangible assets of the hospital. System reflects the patients for the center for advanced business philosophy, through the establishment of all residents in the community with the health file database, on the health status of residents of subdivision, implementation of every resident in the personalized medical service plan, implementation of the residents to focus on disease tracking survey and disease prevention programs to develop, and ultimately achieve a win-win situation between the hospital and the residents.

Decision tree learning is an example based inductive learning algorithm, J.R.Quinlan has made a detailed theoretical description of the. Decision tree learning focuses on the classification rules of decision tree representation from a set of non order and irregular instances.. It uses a top-down recursive way, in the internal nodes of the decision tree attributes. According to the different property values to determine from the junction, a branch of the downward in the leaf nodes of the decision tree, we obtain the following conclusions. So from the root to the leaf node of a path corresponds to a conjunctive rule, the whole decision tree corresponding a set of disjunctive expression rules.

Rough set theory is a mathematical theory of the analysis of data from the Poland mathematician Pawlak Z. in 1982, which is mainly used to deal with uncertain and imprecise information.. Its characteristics is does not need to be pre given some attributes and characteristics of the number of description, but directly from the given problem description set to find the inherent law of the problem [2]. The basic idea is closer to the reality. Now has been part of the research on rough set theory is applied in the decision tree, such as the first of the data sets of attribute reduction, and decision tree is constructed based on the core, the method to construct the decision tree by using attribute reduction to remove the noise and redundant attributes. Resolution is defined, resolution is used as the criterion to construct decision tree.

Using the rough set attribute classification rough degree as the splitting attribute standards, according to the attribute classification rough degree by constructing decision tree, also in this paper proposed using variable precision rough set noise removal method. The standard of dividing the attribute is used in the literature, and the suppression factor is introduced to avoid the decision tree. When the restraining factor is less than a certain value, the decision tree is no longer. Proposed in the literature using core attributes and identify the matrix to select the largest contribution to the classification of attributes. In the literature, the dependence of the attribute of decision attributes on the conditional attribute is proposed as the heuristic information to select attributes.

C4.5 algorithm for decision tree learning problems and it is from the ID3 algorithm to expand and come. These problems include: determine the depth of decision tree growth; to deal with continuous valued attributes, selection of an appropriate attribute selection measure; processing attribute value incomplete training data. To deal with the consideration of various attributes; improve the computational efficiency. A improved decision tree community medical analysis type CRM system research based on rough set, from large amounts of data quickly mining user feeling in rules and its application to the analysis of smart medical CRM system has a very important theoretical value and

 Copyright ⓒ 2016 SERSC

practical significance is proposed, based on the rough set and decision tree theory of these two methods.

2. Method of Improved Decision Tree C4.5 based on Rough Set

Decision tree (decision tree) is used for the main technology of classification and prediction, decision tree learning is by example based inductive learning algorithm, through the example of a group of out of order, no rules to infer the decision tree classification rule.

Decision tree algorithm is a method of approaching the value of the discrete function. It is a typical classification method, first of all data processing, using inductive algorithm to generate readable rules and decision tree, and then use decision to analyze new data. Essentially, a decision tree is a process of classifying data through a series of rules [3].

The basic algorithm of decision tree induction is the greedy algorithm, which is based on top-down recursive way by constructing a decision tree. The basic strategies of the algorithm are as follows:

(1)The tree begins with a single node representing the training sample;

(2)If the samples are in the same class, the node becomes the leaves, and it is used to mark;

(3)The otherwise, the algorithm uses information gain based on entropy measurement as heuristic information, select a sample classification attributes can best be called. The attribute becomes the test or decision attribute of the node.

(4)The every known test attribute value is to create a branch, and then divide the sample.

(5)The algorithm uses the same process, the formation of the sample decision tree recursively on each partition. Once a property appears on a node, it is not necessary to consider it in any descendant of the node.

(6)It can divide step only if one of the following conditions set up stop.

Rough set theory as a computational intelligence science research, whether it is in theory or in practice has made great progress, and it has been successfully used in artificial intelligence, knowledge and data discovery, pattern recognition and classification, fault detection and it.

Definition 1: Information system $S = \{U, Q, V, f\}$, including U: a finite set of objects; Q: a finite set of attribute $Q = C \cup D$, C: condition attributes subset, D: decision attribute subset; V: range of attributes, Vp attribute range; $f : U \times A \to V$ is a total function, making for each $X_i \in U, q \in A$, there are $f(X_i, q) \in Vq$.

In the information system $S = \{U, Q, V, f\}$, $X \subset U$ is a subset of the global, individual attribute subset $P \subseteq Q$, then:

X Lower approximation set: $\underline{P}X = \{Y \in U / P : Y \subseteq X\}$

X Upper approximation set: $\overline{P}X = \{Y \in U / P : Y \cap X \neq \varphi\}$

X Boundary region: $Bnd_P(X) = \overline{P}X - \underline{P}X$

The collection $\overline{P}X$ of $X \subset U$ those elements that are bound to be classified P, U is based on the subset of attributes, and all of the collection of elements X that can be included in the collection, that is, the maximum defined set within it.

A collection of $Bnd_P(X)$ those elements $X \subset U$ that is neither classified nor classified in the $U - X$ upper [4]. The larger $Bnd_P(X)$ the boundaries of the collection and it are the smaller the degree of the determination.

In the rough set theory, knowledge is considered as the ability to classify objects of real or abstract objects. A knowledge base of U can be understood as a relational system,

Copyright ⓒ 2016 SERSC

where U is the domain, and R is the equivalent relationship of U. Decision table information system and decision table, he is a kind of special and important knowledge expression system, is also a kind of special information table, it said when certain conditions are met decision (behavior, operation, control) should be how to it [5]. It is a two-dimensional table, each row describing an object, each column describing an attribute of the object. Attribute is divided into conditional attribute and decision attribute. The object of the domain is classified into decision making with different decision attributes according to different conditional attributes, as is shown by Figure 1.

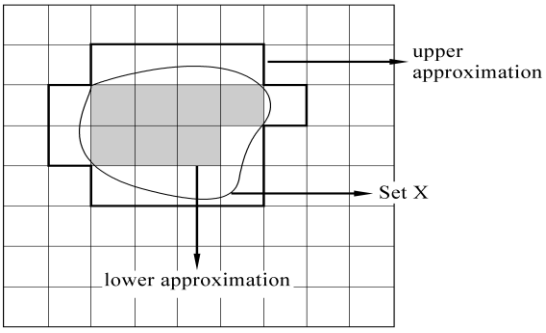

Figure 1. Rough Set Classification Chart

Definition 2: the information system $S = \{U, Q, V, f\}$, U is for the object of the finite set $X_1, X_2, \cdots, X_m$, is divided into a finite sample set of examples, making, $X_i \subseteq U$, $X_i \neq \varnothing$, $X_i \cap X_j = \varnothing \ (i \neq j)$, $i, j = 1, 2, \cdots, m$, $\bigcup_{i=1}^{m} X_i = U$ 。 $Q = C \cup D$, C is Attribute set, which sets the attribute set for the decision attribute set. $p \in P \subseteq C$ Attribute, then rough classification is defined as:

$$CSD(p, C, D) = K(P, D) * \sum_{i=1}^{m} \mu_P(X_i) \tag{1}$$

The classification accuracy $\sum_{i=1}^{m} \mu_p(X_i)$ of each attribute set p obtained by the attribute is demonstrated. $\mu_p(X_i)$ Values indicate that the attributes to decision attribute sample set of classification accuracy, larger values, that value $\mu_p(X_i)$ is greater, that $\frac{card(\underline{P}X_i)}{card(Bnd_p(X_i)) + card(\underline{P}X_i)}$ is caused by uncertain factors less, the effect of classification, $Bnd_p(X_i)$ the better; on the contrary, that the attribute set of classification results is not obvious, namely the classification uncertainty [6]. Thus, $\sum_{i=1}^{m} \mu_p(X_i)$ it is demonstrated that the P attribute is a measure of the classification accuracy of all sets.

The decision tree uses the gain information metric to select test attributes at each node of the tree.. This metric is called an attribute choice metric or a split measure of the pros and cons.. Select the attributes of the highest information gain (or maximum entropy compression) as the test attribute of the current node. This property makes the amount of information needed is for the division of the sample classification minimum, and reflect the division of minimum random or "impurity". This information theory makes the desired number of the desired test minimum for an object classification, and is ensured to find a simple tree.

Definition 3: let S be a collection of s data samples. If the attribute of the class label has a different value, the m m is defined as different Ci (i=1,... , m). Let Si be the sample

 Copyright ⓒ 2016 SERSC

number of class Ci. The desired information for a given sample classification is given by the following:

$$I(S_1, S_2, \ldots, S_m) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (2)$$

Among them, PI is the probability of any sample belonging to Ci, and estimated by S / Si. Note that the logarithm to the base 2, because information in binary code.

Usually C4.5 algorithm is the most suitable for the following problems:

The example is the "attribute value pair said: instance is to use a fixed set of attributes and their values to describe. C4.5 algorithm not only can deal with discrete values, also allows the processing domain for real property.

Definition 4: Generating total-1 segmentation points is in the sequence of values. The value of I (0<i<total) is set to Vi= (Aic+A (i+1) C) /2, which can be divided into two subsets of the data set on the node [7].

$$K_E(x) = \begin{cases} \frac{1}{2} c_d^{-1} (d+2)(1-\|x\|^2), & if \ \|x\| < 1 \\ 0, & if \ \|x\| \geq 1 \end{cases} \quad (3)$$

The objective function with discrete output value: function C4.5 algorithm can learn more than two discrete output values. But it can't learn the function that has real number value output.

It may require disjunctive description (disjunctive description): the decision tree C4.5 algorithm to generate naturally represents the disjunction expression.

Definition 5: let P. R, when P is independent, and Ind (P) =Ind (R), then said R is a reduction of P, denoted as Red. R R all the non relational composition of the collection known as the nuclear Core. Push and prove: Core= Red.

In the internal nodes of the decision tree for comparison of attribute values, and according to the different attribute value judgment, in the leaf nodes of decision tree, we obtain the following conclusions from the node, a branch of the downward; the whole process is repeated with a new node to the root of the subtree. For example, an example of classification is from the root node of the tree to test the nodes represent attributes, then along a branch of the attribute value moving down, repeat the process until the leaf node is reached, the instance belongs to class.

Definition 6: let s be a contain s a sample data set, class attribute can take M different values, corresponding to m different categories of Ci, I epsilon {1,2,3... M}. If Si is a sample number in the class Ci, then the amount of information required for a given data object is the follow equation (4).

$$C(t) = \frac{E[B(t), B(-t)]}{E[B(t)^2]} = 2^{2H-1} \quad (4)$$

Where H is the Hurst exponent and C is the correlation coefficient.

In rough set theory, "knowledge" understanding is the ability of classification, the division of the data, the available set representation, for example, assuming a given data set u and equivalence relation set P, if P divides u, it is called knowledge. Knowledge reduction is refers to in the insurance to the classification or decision ability of a knowledge base invariant conditions, delete the irrelevant or unimportant knowledge, which can simplify the judgment rules, to improve the efficiency of decision-making. In practical application, the decision table is usually used to describe each object in the domain.

Using the decision tree to carry out the classification mainly contains two steps.

Copyright ⓒ 2016 SERSC

Step (1): to construct a decision tree model using the training data set. This process is actually a process of machine learning from the knowledge acquired from the data.

Step (2): to classify the unknown input data by using the decision tree model. On the input record, the attribute values of the records from the root node are sequentially tested until a leaf node is reached, thereby finding the class of the record. The key of the two processes is the construction of decision tree.

Based on the above analysis, we use degree of rough classification as the standard splitting attribute CSD (p,C,D), is able to reflect the attribute classification accuracy is guaranteed to construct the decision tree classification,, but also take into account the dependence of condition attribute and decision attribute of the decision tree classification more effective.

3. Smart Medical Analysis CRM System

The medical field of CRM is a research method for the patient and potential patient carries on the exchange, timely access to and convey information, tracking to give the necessary guidance. From the perspective of a non-profit organization, medical institutions should be to an insurance or uninsured patient with quality of medical service [8]. In order to achieve a balance in terms of profitability, managing the relationship with the patient to hospital is particularly critical, lock those payments for Medicare patient, and increase their loyalty, in order to obtain more profits to cover the uninsured patient.

With increasing competition in medical institutions, medical institutions have begun to focus on how to improve the medical institutions of the core competitiveness of the problem, began to make various efforts, try to provide differentiated and personalized service for the patient. Community medical patient relationship management is a real "take the customer as the center" of the management system, the investment is a kind of effective management philosophy is not only for patients to provide perfect personalized service and cultivate loyal quality patients, more can promote community health comprehensive competitiveness, bring the long-term economic benefit is best hospital profits rising breakthrough.

In the cost reduction, the customer relationship management makes the sales and marketing process automation, greatly reducing the sales expenses and marketing expenses. And, because the customer relationship management to enterprises and customers have highly interactive, help the enterprise to realize customer more accurate positioning, so that enterprises retain old customers, gain new customers the cost decreased significantly. In hand, increase income, due to the process of customer relationship management in the hands of the large number of customer information can be through data mining techniques to discover customer potential demand, cross selling can bring additional new sources of income. And, due to the use of customer relationship management, can more closely relationship with customers, increase the number and frequency of orders, reduce customer loss.

Customer churn analysis and modeling is a new application of application data mining technology. In short, the prediction model is a pattern of discovery from the database, and is used to forecast the future [9]. Customer churn prediction model of simple said is from the customer data warehouse in extraction of a certain amount of training samples, after pretreatment of training set is formed, by using data mining methods, the formation of predictive models, predicted by the model to the new sample classification, predict whether a customer has the loss of possibility, as is shown by equation (5).

$$\begin{aligned} E_{\beta} &= \left\| \Sigma_{a} - C(a(k) \mid \mathcal{R}_{k}) \right\|_{2}^{2} \\ &= \sum_{i,j} (\Sigma_{a} - C(a(k) \mid \mathcal{R}_{k}))_{i,j}^{2} \\ &= \mathrm{tr}\{(\Sigma_{a} - C(a(k) \mid \mathcal{R}_{k}))^{2}\} \end{aligned} \tag{5}$$

 Copyright ⓒ 2016 SERSC

Where, E is that Customer relationship management of the relationship between the enterprise CRM to manage and customer, C(ak) is a business strategy for choosing and managing a valued customer and its relationship. CRM requires a customer centered business philosophy and corporate culture to support effective marketing. Above we can make a simple understanding for e-commerce and CRM, we now the commercial software market, CRM trend can be described is like a duck to water, recent national policies tend to emphasize the in the small and medium-sized enterprise, in CRM users inside the small and medium-sized enterprise user occupies a large part of the points, and CRM itself has a good flexibility.

Using information technology to transform the enterprise management mode, establish to office automation, financial management information system implementation in Enterprise Resource Planning (ERP), supply chain management (SCM), customer relationship management (CRM) as the target of an integrated management system, build a web site or through the intermediary of the network to carry out information exchange and the development of electronic commerce, realize the management innovation of enterprises and medical and public health. In 2013, the state will continue to accelerate transformation of the mode of economic development, adjust and optimize the industrial structure, improve the overall quality of the industry, started the "Twelfth Five Year Plan" national major scientific and technological infrastructure construction, accelerate the promotion of major projects of strategic emerging industries, implementation of a number of high technology major projects, to accelerate the development of a major information technology.

The significance of optimizing customer value is through a series of activities, so that we gradually become the important environment in the value chain of the other party. This not only keeps the low cost of continuous sales, but also enables us to control the value chain of the other party, so that we can get the maximum profit [10]. Optimize product / service in customer value chain space and position points, optimize customer value first step is to optimize our products in customer value chain space position. From the secondary position gradually to the main, key position for replacement.

$$f(x) = \frac{1}{nh^{d}} \sum_{i=1}^{n} k\left(\left\| \frac{x - x_i}{h} \right\|^2 \right) \tag{6}$$

Community hospital customer relationship management system (HCRM) to the hospital based on the doctor-patient relationship of daily affairs management and decision analysis, in equation (6) f(x) is mainly for customers, x is follow-up services, k is complaints and other aspects of the data were collect and collate, to improve the patient's degree of satisfaction and the loyalty.

Customer relationship management is the process of the enterprise in the face of customer, from the judgment, choice, strive to develop and maintain the whole process to implement. The first country in development of CRM is the United States, domestic CRM started late, but shows a strong momentum of development, in foreign countries, computer technology has been used in the hospital for more than 40 years of history, the United States is about in the early 1960s, and the earliest began his research. In recent years, the hospital information system in China has great development. But the application of CRM system in community medical system has just begun.

Hospital (Medical Management System HMMS) based on the computer grid is the main force for hospital management and operation. After ten years of development, has begun to take shape and the paperless economic accounting automation, office treatment, medical electronic files, graphics image digitalization, integrated information network of principle, to the unification of the system standard direction.

Customer relationship management is using modern techniques, the customer, competition, brand, three elements of coordinated operation and realize the optimization

Copyright ⓒ 2016 SERSC

of the whole system, the goal is enhance the competitive ability of the enterprises in the market and support long-term customer relationships, continue to tap the new sales and service opportunities, so that enterprises and ultimately achieve sustained growth in sales revenue, profits and shareholder value.

4. Smart Medical Analysis CRM System based on Rough Set Improved Decision Tree

Decision tree is a similar to the flow chart of the tree structure, which each internal node said test on an attribute, each branch represents a test output, and each node represents classes or class distributions, the topmost node of the tree is the root node. More explicitly said that the decision tree is according to the root node to leaf nodes of the order of examples classified. Among them, each node represents an attribute and each branch represents that it is connected to the node in the attribute value.

In the basic structure diagram of the decision tree, the middle node is often expressed in rectangular nodes, and the leaf nodes are represented by ellipse, as is shown in Figure 2.

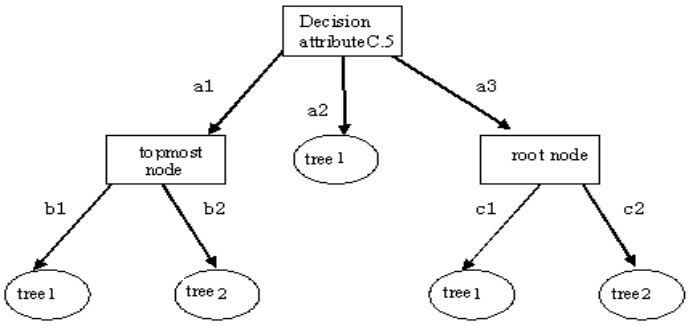

Figure 2. Basic Structure Diagram of Decision Tree

The C4.5 algorithm uses the top-down greedy search to traverse the possible decision tree space, which can be described as a hypothesis from a hypothetical space searching for a fitting training sample [11]. The assumption that the C4.5 algorithm searches is that the decision tree is possible. C4.5 algorithm to a from simple to complex hill-climbing algorithm to traverse the hypothesis space, starting from the empty tree, then gradually consider more complex assumptions to search to decision tree with a correct classification of the training data.

This paper is to solve the traditional mining algorithm efficiency is not high, redundancy rules to be big, the user only to them were interested in a part of the rules and other issues, rough set and traditional decision tree mining algorithms are combined, in order to improve the efficiency and practicality of decision tree mining and medical institutions in the community analysis CRM system application.

In this paper, we use rough classification to construct decision tree. The decision tree based on rough classification is the standard of attribute classification accuracy and conditional attribute and decision attribute.. The greater the roughness of attribute, the more the determination of the attribute, and the dependence of the attribute and decision attribute. After a large number of examples of the analysis, it is in the process of splitting the attribute; it is based on rough classification decision tree algorithm for the selected attribute classification accuracy to better than C4.5 algorithm selection with the maximum information gain properties.

Rough set theory can analyze based on past a large amount of empirical data to find these rules, rough set based decision support system in this area makes up the shortage of

 Copyright ⓒ 2016 SERSC

conventional decision-making methods, allow the decision objects exist some not too clear, less complete attribute and after reasoning that almost certainly the conclusion. Knowledge discovery from the database, modern society, with the rapid development of information industry, a large number of information from the financial, medical, scientific research and other fields of information is stored in the database.

The data mining process based on rough set includes data preprocessing, reduction (including attribute reduction and attribute value reduction) and rule extraction.

Firstly, the training set is classified according to the attribute and the category, and the classification rules are generated according to the relationship between the subsets of the attribute subsets and the approximate and the lower approximation of the target attribute subsets. In practice, the advantage of rough set knowledge reduction and other classification techniques are used to classify incomplete data [12]. Application of rough set attribute significance of training samples of 17 attributes by learning to form a training sample of 12 attributes, based on using C4.5 algorithm modeling, greatly improve the efficiency of learning. Experimental results show that the model is robust and stable.

Attribute reduction: a notable feature of data mining method based on rough set is that it has explicit knowledge expression form. According to the definition of information system in rough set theory, the attribute A is divided into C and decision attribute D, so we can get the C Then D If according to the information table. In theory, we can get a rule for each record in the information system. But the rule obtained by the information table is more conditional, the generalization ability of the rule is weak and the application is narrow.

Definition 7: let.U X be a collection, R is an equivalence relationship defined on U. A: 1 if R (x) =U{Y epsilon U / R:Y epsilon x}, R (x) for X r approximation set; (2) if R (x) =U{Y epsilon U / R:Y U-shaped; X = Phi}, R (x) for X r approximation set; (3) if R (x) = a (x) - R (x) is called R (x) as the set X the boundaries of the domains. If the R is empty (X), called X for the set of set R is clear; on the other hand, call set X is about rough set R.

Definition 8: let R is a family of equivalence relations, and {r} R, if ind (R) = ind R-{R}, is called {r}, R, otherwise known as {r} R can be omitted.

$$k(P,D) = \frac{card(POS_P(D))}{card(U)} = \frac{card(\sum_{i=1}^{m} \underline{P}X_i)}{card(U)} \quad (7)$$

The algorithm of improving decision tree based on rough set is as follows:

Algorithm: Generate_decision_tree generates a judging tree from the given rough training data.

Input: training sample samples, represented by discrete value attributes; the collection of candidate attributes attribute_jist.

Output: a decision tree.

Method:

(1) Create node N;

(2) the nodes for all data samples in a continuous type description attribute specific values and ascending sort attribute values the value sequence {A1c, A2c... Atotalc}.

(3) Returns N as a leaf node to class C Tags:

(4) attribute_list If is empty. Then

(5) Returns N as the leaf node, marking the most common class of samples;

(6) If attribute_list is empty, return N as the leaf node, and tag the most common class of Samples;

(7) Calculate the roughness of each attribute in attribute_list;

(8) S in S1 S2,,..., by the value of t split (according to k t may be Sk, S S1, S2, Sk,);

Copyright ⓒ 2016 SERSC

(9) Select the best segmentation point from the total-1 segmentation point. For each split point data set, the C4.5 calculates its information gain ratio, and selects the segmentation point to partition the data set.

(10) If Si is empty, and a leaf is added, the most common class of Samples is marked. Otherwise, add a node returned by _Tree Generate_Decision (Si, attribute_list-test_attribute).

C4.5 can handle the discrete description attributes and also handle the continuity description attribute. In selecting a node branching attributes for discrete attribute description, C4.5 and ID3 is the same, according to the number of the attribute values of the parameters were calculated; for a continuous description of the properties of AC, assuming that the data in a node set number of samples for total.

The establishment of decision tree consists of two stages: the first stage, the stage of building. Select the training data set for learning, export decision tree. Decision tree induction of the basic algorithm is a greedy algorithm, it uses is top-down recursive divide and conquer approach to construct decision tree, algorithm is outlined as follows. The second stage: the pruning stage. Testing decision tree with test data set, if the established decision tree can not correctly answered [13]. We want to decision tree pruning algorithm to solve the over adaptation problem data until the establishment of a correct decision tree.

Definition 9: (equivalence relation) design knowledge representation system s = (U, a, V, f), if the attribute set P A, called P not resolved between ind (P) is the equivalence relation on u, which ind (P) = {(x, y) epsilon U * u |. Epsilon P and f (x, a) = f (y, a)}. The set of all the equivalence classes derived from Sx(f) is denoted as P / U, which constitutes a division of the domain and contains the equivalence class of X, denoted as p [x]:

$$S_x(f) = C \mid f \mid^{\alpha},\ -\tfrac{1}{2} \le f \le \tfrac{1}{2}\ C > 0 \tag{8}$$

Attribute reduction, the basic theory of rough set and some expansion of the relevant theory of data analysis and reduction. So called knowledge reduction is to reduce the time complexity of decision tree generation, which is based on the same ability of keeping the classification ability of knowledge base on it. On a table information data mining decision rules, if the attribute reduction can reduce decision tree mining algorithm for the calculation of the amount, can also reduce the redundancy of decision rules.

Using a reduction set RED from the decision system S= (U, A) to generate the rules of the process is quite direct. Intuitively, each reduction is used to form a decision rule for each object in the decision table, simply read from the appropriate attribute values from the table. In the form of equation(9), f(a) is in the similar logic language, f(b) with the decision rule is expressed as x is the antecedent of the decision rule and the combination of the conditional attribute value.

$$f(b) - f(a) = \frac{f'(x_3)}{\frac{1}{x_3}}(\ln b - \ln a) = x_3 f'(x_3) \ln \frac{b}{a} \tag{9}$$

Decision rules mining value reduction): value reduction rule acquisition, reduction of decision rules is to elimination of decision rules in the necessary condition attribute value that is to calculate each rule of nuclear and simplified. After the reduction of attributes, the redundant values are eliminated. In the algorithm of decision tree model mining, the attribute of the rule conclusion is reduced, and the decision rule is reduced.

Using a community medical institutions database simulation experiment is carried out, in practice on the basis to find the best solution; in the optimization improved the experimental results were analyzed, the algorithms or methods improved better [14]. Of analytical CRM function design, focusing on from the target patients, patients with potential, the opportunity for patients, patients, referral and management of patients with six function modules were functional division, in analysis module in the application of improved decision tree based on rough set mining algorithm of potential patients,

 Copyright ⓒ 2016 SERSC

opportunity patients and management with data for extracting the decision, analysis, to provide basis for decision making for managers.

Definition 10: a variable X, it may have a variety of values, namely x1, X2,..... , xn, the probability that each one is P1, P2,..... , Pn, then the entropy of X is defined as:

$$x_1^2(t+1) = \beta^2(t)\sum_{i=i_1}^{i_2} b_i^2(t)^* x_i^2(t) + \left(\frac{N^2(t)}{N^3(t) + N^2(t)} \right) s_0 \alpha N^1(t) \tag{10}$$

For the classification system, the category C is a variable, and its possible value is C2, C1,.....,Cn, while the probability of each category is P (C1), P (C2),..... , P (Cn), therefore n is the total number of categories.

The end user of the system is divided into two categories: a class is user community hospital, community hospital department will use the system records and call the community residents and health related information services for the community residents, and thus the health of residents for scientific and effective for management and help. The other is community residents. Use the system to obtain the effective data about the health status of oneself, and participate in the plan and the health related activity arrangement.

5. Experiments and Analysis

This system is divided into customer management, customer service management, marketing management, statistical analysis and other modules.

The first step of the system is data preprocessing: the part of the input data processing, including the data missing value, the attribute discretization and generalization. The actual storage of data in the database is affected by the factors such as human or physical, and there are some interference factors such as noise data, vacancy data and inconsistent data. So it is necessary to pre process the data in the database before data analysis and mining, and provides a flexible and convenient data abstraction platform for application development.

Theorem 1: let a property V take a different value {a1, A2, A,... , av}, the use of attribute A can be divided into S collection V sub set {S1, S2,... And Sv}, including SJ contains a set s of attribute a in AJ value of sample data, if attribute a was selected as the test attribute (for the sample set division), let SIJ subset SJ belong to Ci of the sample set, using properties of dividing a current sample collection of information entropy.

Selected from the community medical institutions in the database part of the record and interpret data, choose the important 26 effects explain the property of the conclusion of condition attribute set was composed, a decision attribute, 158 object constitute training sample set. The decision tree is constructed by using C4.5 and improved decision tree algorithm based on rough classification, and the two value discretization of sample set is carried out before constructing decision tree. The process of building a decision tree is as follows.

Step1: pruning (pruning) method, the main purpose is to remove the noise or abnormal data, make the decision tree algorithm has better generalization ability. Pruning often using a statistical metric, branches cut off the most unreliable, leading to rapid classification, ability to raise the independent test data to carry on the card to make a classification tree. According to the implementation of the pruning time is divided into two methods: pre pruning method and post pruning method.

Step2: after every record in the attribute reduction of decision table can be used as a rule, but which contains a large number of redundant information, namely in the reduction of information system, and not every record every attribute values are for information system and decision rules extraction work for a must for attribute reduction results

Copyright ⓒ 2016 SERSC

continue to simplify. The redundant information in the decision table after attribute reduction is the attribute value reduction. Actually, attribute value reduction is further reduction, as is shown by equation (11).

$$q_{ii} = \lim_{h \to 0^+} \frac{p_{ij}(h)}{h} = \begin{cases} \lambda_i, & j=i+1, \\ u_i, & j=i-1, \\ 0, & |i-j| \geq 0. \end{cases} \tag{11}$$

Where q shows the importance of attribute importance of decision tree, p is the bigger, h shows that the correlation degree of attribute set and decision attribute u is higher. When the attribute set has only one attribute.

Step3: After the decision tree is constructed, the classification rules can be extracted directly from the decision tree, and the rule is expressed in the form of IF-THEN. Create a rule for each path from the root to the leaf node. Along the path of each attribute value on the form before rule (part IF) a conjunction. The leaf nodes contain classes of projections that form the rule (THEN part). As the following diagram is a decision tree that has been generated, the follow equation (12):

$$\Pr\left[\frac{f-q}{\sqrt{a(1-q\times p)/N}} \succ z\right] = c \tag{12}$$

Where n is the number of instances, f=E/N for the observed error rate (which e n instances classification error number) and Q is the true error rate and C for the confidence (C4.5 algorithm of an input parameter. The default value is 0.25), Z correspond to the confidence degree C standard deviation, its value can be set according to the C value by looking up table of normal distribution is obtained. Through this formula, a confidence limit of the true error rate Q can be calculated.

This experiment environment is: the hardware environment: CPU 2G, P43.8G memory, software environment: Windows7 flagship version, the design of the CRM system based on the Struts framework for the J2EE platform. The experimental data is the data of the disease of a community medical institution after data preprocessing, the data set contains 5866 samples, each sample has 9 attributes, and the target is classified into 4.The experiment is divided into two steps, first step test when alternative to generate binary decision tree of decision tree classification performance influence; second step test when paying attention to small classes (ISPASS=2) of decision tree classification performance. The experiment makes use of 20 times cross validation to evaluate the classification effect. The results were 1 and 2 respectively.

Table 1. Comparison of Classification Results when Decision Tree Generated

Classification method	Category		The minimum sample number for the stop split			
			60	70	80	100
StandardC4.5	ISPASS=2	Precision	0.831	0.913	0.928	0.933
		Recall	0.920	0.937	0.947	0.957
	Node number		1256	885	365	158
Improved decision tree by Rough Set	ISPASS=2	Precision	0.898	0.925	0.933	0.959
		Recall	0.939	0.940	0.958	0.982
	Node number		564	231	105	48

 Copyright ⓒ 2016 SERSC

Selection of ISPASS=2 class classification accuracy higher resolution to stop the minimum sample number ranges 80, set C1 is more than or equal to 0.8, C2 is more than or equal to 1.6, C3 = 2.5 for the second step experimental. Results are shown in Table 2 shows:

Table 2. ISPASS=2 Class Classification Results Concern

Classification method	ISPASS=2		ISPASS=1		Node number
	Precision	Recall	Precision	Recall	
StandardC4.5	0.926	0.969	0.956	0.988	658
Improved decision tree by Rough Set	0.986	0.992	0.966	0.992	36

In Table 1 and Table 2 can be seen when flexible application based on rough set improved C4.5 algorithm to generate non equilibrium data for the binary decision tree, not only ISPASS=2 small class classification accuracy and regression rate improve and ISPASS=1 large class of accuracy of classification and regression rate also increased slightly and decision tree complexity has a significant decline. Thus, in does not change the distribution of samples is flexibly based on rough set improved decision tree algorithm for imbalanced data set of decision tree generation for a community medical institutions disease patient data has a better classification results.

From the experimental results, it can be seen that based on rough classification of the improved decision tree algorithm in terms of the number of generated rules to more than the standard C4.5 algorithm, the algorithm for constructing decision tree is relatively complex, but based on rough classification algorithm of decision tree C4.5 algorithm with an average accuracy of 6 percentage points higher than. The experimental results are tested repeatedly and the stability of the decision tree algorithm based on rough set is improved.

6. Summary

Customer relationship management is to customers as the center of business strategy. It uses information technology means, to restructure the enterprise work flow, so that enterprises and customers better communication, to achieve customer profitability is maximized. Decision tree is a data mining in a very effective classification method, classification, prediction, rule extraction in sometimes interested to some association rules. Therefore, the improved algorithm has become the hot research. Rough set is proposed by Z. Pawlak in the early 1980s, a for dealing with uncertain and vague knowledge of the mathematical tools, the basic idea is in the premise of keeping the ability of classification, through the reduction of knowledge, derived concept classification rules, suitable for to find hidden in the data, potentially useful rules, find out the relationships and characteristics in its internal data, has been widely used in knowledge acquisition, decision analysis, machine learning and other fields.

In this paper, we firstly propose an improved decision tree algorithm based on rough set, which is based on the attribute division, considering both the attribute classification accuracy and the dependence of the conditional attribute and the decision attribute.. The analysis of the large number of examples proves that the improved decision tree algorithm based on the rough classification algorithm is better than the standard C4.5 algorithm in the classification accuracy of the classification accuracy. Finally, the decision tree based on rough set improved mining application to the analysis of CRM system analysis management function module, the whole data mining process is divided into three steps:

Copyright ⓒ 2016 SERSC

data preprocessing, attribute reduction, decision rule mining. Through the system can effectively find, maintain and retain patients, mining new patients, provide personalized service for patients, so as to realize the community medical institutions in the limited capital and technology support issued to patients with better service and realize the profit maximization objective, and provide a scientific basis for decision-making in the management, improving the intelligent management level of the medical community.

Acknowledgments

This paper is supported by Scientific and technological projects of Henan Province in China (142102310482), and also is supported by the science and technology research major project of Henan province Education Department (13B520155) and Henan Province basic and frontier technology research project (142300410303).

References

[1] S. Tsumoto, "Automated discovery of positive and negative know ledge in clinical databases", IEEE Engineering in Medicine and Biology, **(2000)**, pp. 56-62.
[2] A. Kusiak, J. A. Kern and K. H. Kernstim, "Autonomous Decision-Making A Data Mining Approach", IEEE Transactions on Information Technology in Biomedicine, vol. 4, no. 4, **(2000)**, pp. 274-284.
[3] H. Xu and R. Zhang, "Semantic Annotation of Ontology by Using Rough Concept Lattice Isomorphic Model", International Journal of Hybrid Information Technology, vol. 8, no. 2, **(2015)**, pp. 93-108.
[4] W. H. Sang and Y. K. Jae, "A New Decision Tree Algorithm Based on Rough Set Theory", International Journal of Innovative Computing Information and Control, vol. 4, no. 10, **(2008)**, pp. 2749-2757.
[5] K. D. Supriya and P. R. Krishna, "Clustering web Transactions Using Rough Approximation", Fuzzy Sets and Systems, vol. 148, **(2004)**, pp. 130-139.
[6] W. H. Sang and Y. K. Jae, "Rough Set-based Decision Tree using the Core Attributes Concept", Second International Conference on Innovative Computing, Information and Control, Japan: IEEE, **(2007)**.
[7] M. Yahia, R. Mahmodr and N. Sulmmann, "Rough neural expert systerm", Expert system with Application, vol. 18, **(2002)**, pp. 87-99.
[8] Y. Fang, "An Approach to Evaluating the Effectiveness of Customer Relationship Management with Interval Grey Linguistic Variables", JDCTA, vol. 7, no. 2, **(2013)**, pp. 372 - 378.
[9] G. Alvatore, M. Bentto and S. Roman, "Rough set theory for multi criteria decision analysis", European Journal of Operational Research, vol. 129, **(2001)**, pp. 1-46.
[10] H. Hong and J. Chun, "The Performance of Customer Relationship Management System: antecedents and consequences", JCIT, vol. 8, no. 12, **(2013)**, pp. 385 - 390.
[11] M. Sonajharia and J. Rajni, "Rough Set Based Decision Tree Model for Classification", 5th International Conference on data warehousing and knowledge discovery, DEXA Society, **(2003)**; Prague, Czech Republic.
[12] A. A. Estaji, M. R. Hooshmandasl and B. Davvaz, "Rough set theory applied to lattice theory", Inf. Sci. vol. 200, **(2012)**, pp. 108-122.
[13] A. A. Bakar and A. Arshad, "Rough Set and Decision Tree Model for Determining Scholarship Award Qualification", RNIS, vol. 12, **(2013)**, pp. 65 - 70.
[14] J. Wei, "Rough Set based Approach to Selection of Node", International Journal of Computational Cognition, vol. 1, no. 2, **(2003)**, pp. 25-40.

Authors

Hongsheng Xu, he was born on December 28, 1979.

Educational background: master, Henan University, Kaifeng, China, 2007;

Major field of study: data mining, Knowledge discovery, artificial intelligence, Customer Relationship Management.

 Copyright ⓒ 2016 SERSC

Lan Wang Professor, she was born on Nov 23,1967.

Educational background: master, Northwestern Polytechnical University,Xian, China, 2007;

Major field of study: Customer Relationship Management, Decision tree, Rough set, data mining, artificial intelligence.

Copyright ⓒ 2016 SERSC

Copyright ⓒ 2016 SERSC