Analysis of Decision Tree Mining Algorithm Based on Improved Rough Set Classification

Lan Wang[1] and Hongsheng Xu[1, a*]

[1]Luoyang Normal University, Luoyang, 471934, China

[a]85660190@qq.com

*The corresponding author

Keywords: Rough set; Decision tree; Data mining; Classification; ID3

Abstract. Classification is to find a set of models (functions) that describe the typical features of data set. In this paper, firstly, the rough set classification algorithm is improved. Then, analysis and application of classification algorithm based on improved rough set are described. This paper analyzes the advantages and disadvantages of the traditional decision tree algorithm, and puts forward the improvement method. The paper presents analysis of decision tree mining algorithm based on improved rough set classification.

Introduction

CRM was initially applied to service enterprises to help enterprises maintain a large number of customer resources. Manufacturing enterprises are different from service enterprises; manufacturing enterprises generally use a relatively fixed production process and production process of mechanical parts. In management, focus on the construction of sales network, their number of customers is relatively small, but the contact with the customer's frequency is high, its sales growth depends on the sales channels and sales performance, and large customer relationship maintain good.

Rough set theory is a mathematical tool, which was first proposed by mathematician Z Paw in Poland in 1982, which is a kind of mathematical tool which is not proficient and uncertain [1]. Through the study of rough set theory and ID3 algorithm, we use the knowledge of the importance of rough set attribute to select the attributes with high degree of importance.

Regression analysis is a widely used prediction technique. The purpose of regression analysis is to find out the relationship between numerical variables, and to express them through the function relation. The relationship between the prediction effects of regression analysis depends only on the prediction variables and other variables, namely, the accuracy of the model depends on the independent variables and the dependent variable distribution accords with the model, the independent variables and the dependent variable distribution for the selected model has good prediction effect.

The decision tree data mining technology has been widely used in various fields based on its application in the retail sector, mainly related to customer segmentation and cross selling and so on; in the financial industry, the application of outstanding performance in the aspects of preventing fraud and credit evaluation; in the telecommunications industry, the most prominent manifestation is to maintain its applications in customer; in the field of electronic commerce, which is widely used in many aspects of online sales, online advertising, digital sales and customer relationship management.

One of the most important steps in data mining is to find the right data for the mining algorithm. In the process of the customer's transaction through the e-commerce website, there are two main sources for the enterprise to obtain the relevant data: (1) server data; (2) customer registration information. The paper presents analysis of decision tree mining algorithm based on improved rough set classification.

Analysis and Application of Classification Algorithm Based on Improved Rough Set

At present, the ideal method is to construct the decision tree with better heuristic function and extract the decision rules. The method is based on attribute importance evaluation index as the information entropy to select the condition attribute, attribute of between dependency and redundancy are fully considered in inconsistent decision table of correct classification, which make up the attribute dependence of insufficient emphasis on the shortcoming of ID3 algorithm, to solve the decision tree repeat neutron and some properties have been repeatedly selected in the same decision tree.

Rough set theory not only provides a new scientific logic and research method for information science and cognitive science, but also provides an effective processing technology for intelligent information processing [2]. At present, rough set theory has become a new hot spot in the field of artificial intelligence at home and abroad. The rough set theory answers how people will find useful knowledge from these vast data in the face of a growing database.

The construction of decision tree can be divided into two stages: building stage and adjustment stage. This paper discusses the application of decision tree classification based on a teacher's classroom teaching evaluation system. This paper mainly discusses data mining and knowledge discovery in classroom teaching evaluation database, as is shown by equation (1).

$$S_j^{\,2} = \frac{\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)}{n-1} \tag{1}$$

In the construction of decision tree, ID3 algorithm is the most influential decision tree generation algorithm, which was proposed [3]. The basic idea of ID3 algorithm is based on the information entropy to measure the attributes for decision tree node selection, each attribute is selected the most amount of information or to make the minimum entropy into attributes, to construct a decision tree entropy decrease most quickly, to the leaf nodes of the entropy is 0.

Application of decision tree to classify new samples, from the root node to begin testing the sample, determine the next node according to the test results, or until it reaches the node, prediction category belongs to a leaf node is the new node. Decision tree algorithm is ID3, 4.5, C5.0, CART, as is shown by equation (2).

$$R_2 = R_1 \cup \{c_j\} \rightarrow \min(\Delta\gamma_{P\cup C}^{\beta}(R_1 \cup \{c_j\}, D)), c_j \in C - R_1 \tag{2}$$

Attribute deletion: the property that has a large number of different values and no generalized operators or other attributes that can be used to replace its higher level concepts. For example, the user information in the user ID, ID number, etc., they are too many values and can not be found in the value of the domain operator, should be removed.

There are a lot of attribute reduction algorithms. In this case, the attribute reduction is used, and the decision table after reduction [4]. In the decision table, out of about sex, marital status two attributes, clients in gender, marital status is not whether or not to purchase the company's products a decisive factor, age, education and income is the decisive factor.

Select the reference value in the classification table of the small layer attribute data, the reference classification of each layer of the log interpretation results as the last column to join the selected logging data, constitute a training sample. Finally, according to the training samples, the logging interpretation results are taken as decision attributes.

$$\xi_{ij}(k)=\left[1+\left|\frac{\Delta x_i(k)}{\sigma_i}-\frac{\Delta x_j(k)}{\sigma_j}\right|\right]^{-1} \tag{3}$$

There are some limitations in the practical application of classification [5]. Because of the rough set theory to deal with the classification is accurate, only consider completely "contains" data set does not contain "and", there is no ambiguity, and its processing object is known, and extracted from the model conclusion applies only to those objects.

The lower approximation set is obtained from the collection of all of the knowledge contained in the X, while the upper approximation is the intersection of the set of knowledge contained in the X. The results show that the lower approximation set, the upper approximation set and the boundary domain. In rough set theory, when the boundary region is empty, rough set. On the contrary, the upper and lower approximation sets are equal [6].

Step1: An approximation space (or knowledge base) is a relational system K={U, P}, where U is a domain, and P is an equivalence relation on U. If Q is P, the intersection of the equivalence relation in Q is called the Q, which is called IND (Q);

Step2: NO said customers do not drain, YES said the loss of customers. It can be seen from the picture, the customer cost change rate for 100% customers must have loss; and the cost change rate is less than 30% of the customers; customer monthly tariff is relatively stable in general will not be lost, cost change rate from 30% to 99% customers have lost;

Step3: Plus a leaf, marked as the most common class in samples:

$$L_B^{\beta}=(\underline{R_B^{\beta}}(D_1),\cdots,\underline{R_B^{\beta}}(D_r)) \tag{4}$$

Step4: Given a Boolean attribute A, if the node n contains 6 instances of known A=1 and 4 A=0, then the probability of A (x) =1 is 0.6, while the probability of A (x) =0;

Step5: Rule extraction instability problem. The decision mechanism of rough set theory is very simple, so the decision rules generated by rough set are unstable and have poor classification accuracy.

As a kind of rough set theory to deal with imprecise and inconsistent (imprecise) and incomplete effective tools and other incomplete information, on the one hand, thanks to his mathematical foundation of mature, does not require a priori knowledge; on the other hand, it is easy to use.

A New Method of Data Mining Based on Decision Tree

Classification and data mining prediction is one of the important part, there are many classification algorithms, there have appeared a lot of new algorithm (rough set parallel decision tree algorithm, TAN algorithm based on Bayesian based); decision tree classification method has a practical use for classification and decision.

Rough control and fuzzy control are based on the knowledge rule based control, but rough control is more rapid and simple, easy to realize (because sometimes rough control can save fuzzification and defuzzification step); another advantage is that control algorithm can completely from the data itself, so from the perspective of software engineering, and its decision with the fuzzy reasoning process (or neural network) control algorithm can easily be verified as is shown by equation (5) [7].

$$f(x)=1-R_0-x+\frac{1}{\sigma}\ln(\frac{x}{S_0}) \tag{5}$$

In the field of medical diagnosis, the rough set method was used to induce the new cases according to the previous cases. The accuracy rate of artificial abortion is only 17%-38%, and the application of rough set theory can be increased from 68% to 90%. In addition, the application field of the above examples is: software engineering data analysis, approximate reasoning, rough control, image processing, earthquake prediction, power system analysis, the application of a very broad space [8].

The development of data mining, due to the huge data resources in the world and will have to convert these data resources huge demand for information and knowledge resources, knowledge and information on demand from all walks of life, from business management, production control, market analysis to engineering design, scientific research [9]. Data mining can be regarded as the natural evolution of data management and analysis.

The use of data mining technology can help to obtain the knowledge needed for decision making. In many cases, users do not know what the value of the information data exist, so for a data mining system, it should be able to search and found a variety of patterns of knowledge, to meet user expectations and the actual needs.

Classification is to find a set of models to describe the typical features of the data set (or function), in order to be able to identify the classification of unknown data attribution or category, the unknown image to one of the discrete categories.

Experiments and Analysis

The first step is to preprocess the data when the rough set is used for automatic rule acquisition. Data preprocessing mainly includes two aspects: the discrete data and data missing. When rough set theory is used to process decision tables, the values in decision table are expressed by discrete values.

The classification model can be described in many ways [10]. The main methods include classification rules, decision trees, mathematical formulas and neural networks. Decision tree is a tree structure with hierarchical structure:

(1) X, U, X of each concept (sample set) and not clear the relationship between B, X included in the smallest definable set containing X maximum definable sets.

(2) For a hypothetical h, when fitting the training examples of other assumptions worse than it is, but in fact in the case the whole distribution (including training examples set outside) was better, then the assumption of over fitting the training examples.

(3) Let S be an instance set, A is an attribute of the instance in S. H (S) and H (S|A) are known as the information entropy and conditional entropy of S, respectively.

$$\alpha_{R_i} = \frac{\sum_{j=1}^{n} | \underline{R_i} x_j |}{\sum_{j=1}^{n} | \overline{R_i} x_j |} \tag{6}$$

(4) Compared with other methods, the computational cost of decision tree method can reduce the computation time greatly and improve the efficiency of the system.

(5) Let S1 be a collection of samples of test - attribute=a. in samples. Decision rules based on rough set theory are based on the analysis of empirical data.

(6) The implementation of object data mining research for relational database, relational tables can be seen as a decision table in rough set theory, and bring great convenience to the application of rough set method, there is uncertainty in the real world rules, there is uncertainty.

Can automatically extract features of control rules provides a new means to solve the problem of rough set. A new control strategy of fuzzy rough control (fuzzy-rough control) has emerged as an attractive development direction. If the range of decision table attribute or attribute is to certain conditions for continuous values (floating point expression), and it is then before processing must be discrete. The discretization plays an important role in the whole process of rule acquisition. Good discretization algorithm not only has little loss of information, but also has strong adaptability.

The classification of commonly used categories belonging to predict unknown data instances (finite discrete values), such as a bank customer credit rating are 6, grade). In some cases, however, it is necessary to predict the value of a numeric attribute, which is called a prediction.

Found that the uncertainty of knowledge from the database provides a handy method of rough set. Other techniques used in data mining, such as the neural network method, can automatically select the appropriate set of attributes, and using the rough set method for pretreatment to remove redundant attributes, can improve the discovery efficiency, and reduce the error rate.

Summary

The paper presents analysis of decision tree mining algorithm based on improved rough set classification. Rough set is an important research method of the relational data model and information table in the relational database it is very similar, which is convenient for the embedded database management system based on the algorithm of rough set. Introducing the kernel of rough set is such as simplifying the powerful concepts and methods.

Acknowledgements

This paper is supported by Scientific and technological projects of Henan Province in China (142102310482) and the science and technology research major project of Henan province Education Department (13B520155, 17B520026).

References

[1] Ju-Sheng Mi, Wei-Zhi Wu and Wen-Xiu Zhang, “Approaches to knowledge reduction based on variable precision rough set model”, Information Sciences, vol.159, (2004),pp. 255-272

[2] A.Kusiak,J.A.Kern,K.H.Kernstim,et al., “Autonomous Decision-Making A Data Mining Approach”, IEEE Transactions on Information Technology in Biomedicine vol.4,No.4, (2000), pp.274-284.

[3] Beynon M, “Reduces within the variable precision rough sets model: A further investigation”, European Journal of Operational Research, vol. 134, no. 3,(2001), pp.592-605.

[4] SANG W H, JAE Y K, “Rough Set-based Decision Tree using the Core Attributes Concept”, Second International Conference on Innovative Computing, Information and Control, Japan: IEEE, (2007),pp.298-301.

[5] A.A. Estaji, M.R. Hooshmandasl, B. Davvaz, “Rough set theory applied to lattice theory”, Inf. Sci. vol.200, (2012),pp.108-122.

[6] Alvatore G, Bentto M, Roman S, “Rough set theory for multi criteria decision analysys”, European Journal of Operational Research, vol.129,(2001),pp.1-46.

[7] Lei Lin, Dai Chuanlong, Wang Houjun, “Rough set theory based fault diagnosis of node in wireless sensor network”, Journal of Beijing University of Posts and Telecommunications, vol. 30,no.4,(2007), pp. 69-73.

[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] Guo Xiaohui, Ma Xiaoping, “Fault Diagnosis Based on Rough Set and Neural Network Ensemble”, Control Engineering of China, vol.14,no.1,(2007), pp.53-56.

[10] SANG W H, JAE Y K, “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.