Construction of Decision Analysis System Based on Improved Decision Tree Pruning Algorithm and Rough Set Classification Theory

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

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

[a]85660190@qq.com

Keywords: Pruning algorithm; Decision tree; Rough set; C4.5 Algorithm; Classification reduction

Abstract. Decision trees are generally generated from top to bottom, and each decision or event may lead to two or more events. Rough set is characterized by the use of imprecise, uncertain, partial real information to obtain easy to process, robust, low-cost decision-making scheme. Decision tree method is widely used in enterprise decision-making. Rough sets use upper and lower approximations to describe uncertainty, which makes the boundary clear and reduces the randomness of algorithm design. The paper presents construction of decision analysis system based on improved decision tree pruning algorithm and rough set classification Theory.

Introduction

In the decision tree classification algorithm, C4.5 algorithm is the most commonly used and the most classical one. It inherits the advantages of ID3 algorithm and improves and complements the ID3 algorithm. C4.5 algorithm adopts information gain rate as the criterion for selecting branch attributes. It overcomes the disadvantage of the information gain selection in the ID3 algorithm, which tends to select the attributes with more values, and can complete the discretization of the continuous attributes and deal with incomplete data.

The main idea of rough set theory is to derive the classification rules of concepts through the reduction of knowledge on the premise of keeping the classification ability unchanged. Knowledge is considered as the ability to classify abstract or realistic objects [1]. According to the characteristics of the objects under discussion, the ability to classify them can be regarded as some kind of knowledge. In the process of classification, individuals with little difference are grouped into one category, and their relationship is indiscernible, also called equivalence relation.

One of the important contents of decision table reduction is to simplify the conditional attribute of decision table so that the decision table before and after reduction has the same function. The same decision can be based on a smaller number of conditions, so that we can obtain the same required results by some simple means, which is a good thing with half the effort.

Using modern information technology and decision analysis methods, decision analysis system provides timely and reliable business information for enterprise decision makers by establishing database and analysis model. It helps the decision makers to make quantitative analysis and demonstration on the future management direction and objectives of the enterprises, so as to make a scientific decision on the production and management activities of the enterprises.

The decision tree should be pruned because of avoiding the decision tree from over fitting the (Overfitting) sample [2]. The decision tree generated by the previous algorithm is very detailed and huge, each attribute is considered in detail, and the training samples covered by the leaf nodes of the decision tree are "pure". Therefore, if you use this decision tree to classify the training samples, you will find that for the training samples, the tree is well represented, the error rate is very low and the samples in the training sample set can be correctly classified. The error data in the training sample will also be learned by the decision tree and become part of the decision tree, but the performance of the test data is not as good as expected, or extremely poor, which is the so-called over-fitting (Overfitting) problem.

DOI: 10.25236/icess.2019.273

Rough set Theory is a mathematical tool for the study of incomplete and uncertain knowledge. In information systems, the understanding and representation of knowledge is the first problem that people think about, and it is also a difficult problem to solve. From the current research, rough set theory and technology is an ideal method to solve these problems.

The variable precision rough set model is an extension of the rough set model. It introduces parameters on the basis of the basic rough set theory, that is to say, the existence of a certain degree of error classification rate is allowed. On the one hand, the concept of approximate space is improved. On the other hand, it is helpful to use variable precision rough set to find the relevant data from the data that is considered irrelevant.

Pruning Technique Analysis of Improved Decision Tree

Decision tree is generally composed of square node, circular node, scheme branch, probability branch and so on. The block node is called decision node, and several fine branches are drawn from the node. Each fine branch represents a scheme, which is called scheme branch. A circular node is called a state node, from which a number of fine branches are drawn to represent different natural states, which are called probabilistic branches [3]. Each branch of probability represents a natural state. Indicate the content of the objective state and the probability of its occurrence on each twig. At the end of the probabilistic branch, the result (gain or loss) of the scheme in the natural state is indicated.

Decision tree technology is the most widely used classification technology. It has the following advantages: first, the decision tree method is simple and easy for people to understand; secondly, the decision tree model is efficient and suitable for the situation where the training set has a large amount of data. Thirdly, decision tree method usually does not need knowledge outside the trained data; fourth, decision tree method has higher classification accuracy.

The data in a database generally can not conform to the model obtained by classification prediction or cluster analysis [4]. Those data objects that do not conform to the laws (models) of most data objects are called heterogeneity. Many data mining methods have excluded these anomalies as noise or accident before the formal data mining.

The real test of the success of a learning algorithm is its performance on data not seen in the training. The training process should include training samples and validation samples. Validation samples are used to test trained performance. If the verification results are poor, we need to use different structures to re-train, such as using larger sample sets, or changing the conversion from continuous value to discrete worth data, and so on. Usually, a validation process should be established to test the generalization of training results after the training is finalized.

In order to find the decisive Eigen values and divide the best results, we must evaluate each feature. After completion of the test, the original data set is divided into several subsets of data. These data subsets are distributed across all branches of the first decision point. If the data under a data branch belongs to the same type, there is no need to further split the data set. If the data in the data subset is not of the same type, the process of dividing the subset of data needs to be repeated. If the algorithm for partitioning a subset of data is the same as the method for partitioning the original data set until all data of the same type are in one subset, as is shown by equation (1) [5].

$$H = -\sum_{i=1}^{n} p(x_i)\log_2 p(x_i) \tag{1}$$

The C4.5 algorithm adopts the information gain rate as the criterion of selecting the branch attribute, which overcomes the disadvantage of the ID3 algorithm that the information gain selects the attribute with more values. If we have an attribute D, we divide T into the set T 1 / T 2. When all the records in each set have the same result, that is, when the value is "yes" or "no" at the same time, Info (DNT) is 0, and the gain Gain (D, is then. T) Take the maximum value. So the gain ratio is used instead.

On the basis of the original over-fitting decision tree, a simplified version of the decision tree is generated. The over fitting decision tree can be simplified by pruning. Pruning can be divided into two types: pre-pruning (Pre-Pruning) and post-pruning (Post-Pruning). Let's take a closer look at these two methods: PrePrune: prepruning, stopping tree growth early, the pruning decision tree of the simplified version can be obtained by pruning after PostPrune: and pruning on the generated fitting decision tree.

Decision tree method, as a kind of decision technology, has been widely used in investment decision making of enterprises. It is the most common and popular method in stochastic decision making model. This method effectively controls the risk brought by decision making. The so-called decision tree method is to use the tree diagram to express the expected value of each decision. By calculation, the decision method with the greatest benefit and the lowest cost can be selected. Decision tree method belongs to risk decision making method.

The basic strategies of the algorithm are as follows:

1) The tree starts with a single node representing the training sample.

2) If the samples are all in the same class, the node becomes a leaf and marked with that class.

3) Otherwise, the algorithm uses a metric as the heuristic information and selects the best attribute to classify the sample as the test attribute of the node [6].

4) Creating a branch for each known value of the test attribute and dividing the samples accordingly.

5) Using the same procedure, the algorithm recursively forms the sample sub-decision tree on each partition. When one of the following situations occurs, the recursively stopped: (a) has samples of a given node belonging to the same class of. (b) And has no remaining attributes to further partition the samples or no samples in the branch, the majority vote is used. Converts a given node to a tree leaf and marks it with most classes in the parent node.

The utility function is generally between 0 and 1, so that the minimum return value of the problem is 0, and the maximum income value is 1. The utility value corresponding to other income values needs to be determined by experiments, that is, the analyst repeatedly questions the decision maker to determine the utility value corresponding to each income value and draw points in the right-angle coordinate system, and then connect them with curves. A good utility curve should be smooth and monotonous, as is shown by equation(2) [7].

$$\text{E(age)}=\frac{11}{26}I(s_{11},s_{21})+\frac{12}{26}I(s_{12},s_{22})+\frac{3}{26}I(s_{13},s_{23}) \quad (2)$$

In order to minimize the information needed to classify a subset of training samples that are later partitioned, that is, to partition the current (node) set of samples using this attribute, This minimizes the "different types of mixing" of the resulting set. Therefore, using such an information theory method will help to reduce the number of partitions required for object classification.

At the beginning of the decision tree, the decision tree as a single node (root node) contains all the training samples. 2, if the sample of a node is the same class, the node becomes the leaf node and is marked as the class. Otherwise, the information entropy method (information gain) will be used as the heuristic information to help select the appropriate attributes, so that the samples can be divided into several subsets. This property becomes the test attribute of the node. In the algorithm, all attributes are discrete values. If there are attributes with continuous values, they must be discretized first.

In fact, the criterion of pruning is how to determine the size of the decision tree. The pruning ideas can be referenced as follows: 1. Use the training set (Training Set) and verify the set (Validation Set), To evaluate the utility of pruning methods on pruning nodes 2: use all training sets to train, but use statistical tests to estimate whether pruning a particular node improves the evaluation performance of data outside the training set, such as using Chi-Square (Quinlan,) 1986) to further extend whether the node can improve the performance of the whole classified data or only improve the performance of the current training set data.

Classification Reduction Method Based on Rough Set Theory

Rough phenomenon is based on incomplete information or knowledge to deal with the unclear phenomenon, so it is necessary to classify the data based on the observation or part of the measured information, which requires a different processing method than probability statistics and fuzzy mathematics. This is rough set theory. Intuitively, rough sets are based on a series of data or descriptions that neither know more or less, whether useful or useless, incomplete or even partial, to analyze data and speculate on unknown information.

Definition 1: let U be the domain, P, Q is two equivalent relation families on U, and Q P, if satisfied, 1) ind (Q) = ind (P); 2) Q is independent, then Q is a reduction of P.

Definition 2: DS = (U, A, V, f) is called a knowledge representation system, in which U is a nonempty finite set of objects, called a domain, A is a set of object attributes, and A = C> D, that is, attribute set consists of conditional attribute set and decision attribute set, C, D satisfies C< D = A, C =D; V is the set of attribute values; f: U A # V is a monojection, specifying the attribute value of each object x in U. According to the definition of knowledge representation system, it can be easily expressed in a two-dimensional table, which is called the knowledge representation system (KRS). In this table, columns represent attributes and rows represent instances so that knowledge can be reduced to attribute reduction.

Let S be an information system, X a non-empty subset of U, and P A, Then the P- lower approximation and the P- upper approximation of X are defined as follows: PX= ≡ {y ∈ U I (P): Y X} X = Ⅹ {y ∈ U I (P): Y X ≠ φ} 4 . The approximate quality is such that X = {X1 + X2, … , X n} is a partition of U, where a subset of Xi,i=1,. N, a class of X, the approximate quality is defined as

$$\gamma_P(X)=\frac{\sum_{i=1}^{n}|\underline{P}X_i|}{|U|} \tag{3}$$

The research direction of rough set theory: 1 using abstract algebra to study the special algebraic structure of rough set algebraic space [8]. 2 describing rough space by topology. 3 and studying rough set theory and other soft computing methods A combination of methods of law or artificial intelligence, For example, fuzzy theory, neural network, support vector machine, genetic algorithm, etc. The classical rough set theory based on the equivalence relation is extended to the rough set theory based on the similarity relation or even the general relation.

Ziark takes the error classification rate as the error classification rate, and the An takes the correct classification rate as the correct classification rate, and the value is as follows: this project adopts the method of An. As can be seen, variable precision rough sets become standard rough sets. The main task of variable precision rough set is to solve the problem of data classification with no function or uncertain relation between attributes.

We call a subset of attributes indistinguishable. The indiscernibility relation is a kind of equivalent relation (it is easy to verify that it satisfies the mathematical axiom of equivalent relation), so the elements in the domain can be divided into several equivalent classes, each equivalent class is called the knowledge grain of knowledge base [9]. The set of all equivalent classes is described as a basic set. If the set X can be expressed as the union of some basic sets, then X is called B-exact set, otherwise it is called B-rough set, as is shown by equation(4).

$$a_R\left(X\right)=\frac{Card\left(R_{-}\left(X\right)\right)}{Card\left(R^{-}\left(X\right)\right)} \tag{4}$$

Therefore, the hierarchy of system structure determines the hierarchy of system faults. The failure of the system may be caused by the failure of the elements that make up the system, that is, the propagation of the fault is a layer by layer process from the lower level to the higher level. Considering the hierarchical classification model as a conceptual hierarchical tree, it is known that each conceptual node has only one parent node, but there can be several child nodes.

Because the definition of reduction based on rough set model depends on lower approximation, and the computation of lower approximation is sensitive to noise data, the result of attribute reduction is greatly affected by noise data, and many valuable rules cannot be extracted. In order to deal with noise data better, Ziarko et al proposed a variable precision rough set (Variable Precision Rough Set, VPRS) model, which greatly improved the coverage and generalization ability of extraction rules, and better reflected the data correlation in data analysis. Thus, it lays a foundation for obtaining approximate decision rules.

Experiments and Analysis

The goal of decision analysis system mainly includes data warehouse system and auxiliary decision support system. Its overall goal is to construct the basic frame of "integrated information base" of enterprise comprehensive information transmission and information sharing by using internet technology WEB technology, data warehouse technology and information security technology. This paper focuses on the comprehensive information resources, digitization, and database and network development of the information. The decentralized comprehensive information database is linked to standard, optimized in structure, expanded in scale, interconnected in network, shared in information and developed in comprehensive application. Improve the system, authority, applicability, timeliness and sharing of comprehensive information resources.

The decision process of the decision tree method is as follows: (1) drawing the tree diagram and arranging the various natural states of each scheme according to the known conditions. (2) The probability of each state and the value of profit and loss are marked on the probabilistic branch. (3) Calculate the expected value of each scheme and mark it on the corresponding state node of the scheme [10]. (4) Pruning, comparing the expected value of each scheme, and marking the expected value on the project branch, the final scheme left by the low expectation value (that is, the inferior scheme) is the best scheme.

The concept represented by the parent node of a node is the generalization of the concept represented by the node; similarly, the concept represented by the child node of a node is the specialization of the concept represented by the node. In this way, rough set theory can be used to reduce the complexity of system diagnosis problem.

A very important problem in practical application is attribute dependency. If I (P) I (R), all attribute values in R are uniquely determined by the attribute values in P, that is, the attribute set R A is completely dependent on the attribute set Pa, which is expressed as P R. Another important problem in rough set theory is attribute reduction. The classification quality of attribute set is the same as that of original attribute set. If the minimal subset of attributes P C A, as is shown by equation(5), then the set V is called a reduction of C, denoted as RED (v). In short, reduction is the minimum set of conditional attributes that do not contain redundant attributes and ensure correct classification.

$$\hat{\nu}_{X,j}^{2} \equiv \frac{1}{M_j}\sum_{k=L_j-1}^{N-1} w_j^2(k) = \frac{1}{M_j}\sum_{k=L_j-1}^{N-1} \overline{w}_j^2(k) \tag{5}$$

Because of the over-fitting of the training set, the data of the verification set can be modified, repeatedly operated on, the bottom-up processing node is removed, and those nodes that can maximize the accuracy of the verification set are deleted. Until further pruning is harmful (that is, pruning reduces the accuracy of the validation set) REP is one of the simplest post-pruning methods, but in the case of less data, the REP method tends to be overfitted and used less. This is because the characteristics in the training data set are ignored in the pruning process, so we should pay attention to this problem when verifying that the data set is smaller than the training data set.

C4.5 uses this method to process missing attribute values. The accuracy of a given taxonomy is for correct labeling of future data. Keep (holdout) and 2. 5 tree pruning and avoid overfitting. The K-fold cross validation (K-foldcross-vaildation) method is based on a given number of data because of the noise and outliers, many branches reflect the training samples are randomly divided according to the sample, commonly used to assess the accuracy of classification techniques. The pruning method deals with this problem of overadaptive data. There are pruning square decision tree accuracy, can use bagging or boosting two techniques.

Summary

The method is to prune the branches from a fully grown tree. The tree node is clipped by deleting the branches of the node. On the basis of allowing the decision tree to grow to the fullest extent and according to certain rules, the decision tree is cut off from the decision tree which is not representative of the leaf nodes or branches. After pruning, the pruned branch node becomes a leaf node and marks it as the category with the largest number of categories in the sample. With rough set theory as the main tool, combined with the concepts of generalization and specialization of knowledge, the classification objects are classified from high level universal pattern to low level concrete pattern one by one, which reduces the searching amount of pattern matching in classification. The problem of combined explosion in classification space is solved effectively.

Form enterprise information resources integration processing, exchange and release, decision-making consulting, technical support center. A batch of application systems, such as enterprise management monitoring and warning system, model forecasting system and leader aided decision support system, are gradually formed to provide information support and decision support for enterprise management and regulation.

Acknowledgements

This paper is supported by Henan key Laboratory for Big Data Processing & Analytics of Electronic Commerce, and also supported by the science and technology research major project of Henan province Education Department (17B520026).

References

[1] Y.M. Li, S.J. Zhu and X.H. Chen. Data mining model based on rough set theory. Journal of Tsinghua University,2015,39 (1): 110-113.

[2] [An A, Shan N, Chan C, Cercone N, Ziarko W., Discovering rules for water demand prediction: an enhanced rough-set approach, Engineering Applications in Artificial Intelligence, 2014,9(6):645-653.

[3] C.Q. Miao, Wang Jue. Concept and operation information representation in rough set theory. Journal of Software, 2013, 10 (2): 113-116.

[4] B.S.Ahn,SS.cho, C.Y.Kin.,2000,The integrated methodology of rough set theory and artificial neural netwouk for business failure prediction, Expert Systems with Application, 2011,18:65-74.

[5] H.L. Zeng. Rough set Theory and its applications. Chongqing, Chongqing University Press, 2016.

[6] Greco S., Matarazzo B., Slowinski R., Rough sets theory for multicriteria decision analysis, Eropean Journal of Operational Research,2014,129:1-47.

[7] PAWLAK Z. Rough Sets. Communication of the ACM, 2015, 38(11) : 89-95.

[8] Tsumoto S. Automated discovery of positive and negative know ledge in clinical databases. IEEE Engineering in Medicine and Biology, July August, 2000:56262.

[9] Jiang Liangxiao, Cai Zhihua, Liu Zhao. A decision Rule mining algorithm based on rough set. Microcomputer and applications, 2014 (3): 7-8.

[10] Pawlak Z, Grzymala-Busse J, Slowinski R, Ziarko W., Rough sets. Association for Computing Machinery,Communications of ACM , 2015,38(11):89-96.