Clustering Reduction Method Analysis of Rough Set and Decision Tree based on Weight Matrix Analysis

Lan Wang¹ and Yi Zhang¹
¹Luoyang Normal University, Luoyang, China

Abstract. Decision table is a kind of special and important knowledge representation system. In this paper, rough set theory is used to discretize data and reduce data and rules. From the definition of resolution matrix, it can be seen that the shorter the length of an item in the resolution matrix, the greater the effect of this item on classification. And the more frequently the item appears, the more important it is decision tree. Construction of decision tree is a cyclic recurrence process. The paper presents clustering reduction method analysis of rough set and decision tree based on weight matrix analysis.

Keywords: Weight matrix analysis, Rough set, Decision tree, Clustering reduction, Classification

1. Introduction
Generally speaking, the relationship between one classification and another classification is considered, and the concepts of relative reduction and relative kernel are derived. In rough sets, the concept of relative reduction is the reduction of conditional attributes relative to decision attributes.

Based on the characteristics of rough set theory, it can be applied to the field of knowledge acquisition and can support many steps of knowledge acquisition, such as data preprocessing, data reduction, rule generation, data dependency acquisition and so on.

The establishment of Bayesian belief network can be divided into two stages. In the first stage of network topology learning, that is, the learning of directed acycloïd graph, the structure of Bayesian network composed of all attribute variables and class variables is established from the example data by using the learning algorithm of Bayesian network [1]. In the second stage, the learning of the local conditional probability distribution of each variable in the network is carried out, and the maximum posterior probability of the class variable is calculated when the value of the attribute variable is given. The algorithm based on this classification idea is TAN algorithm. However, there are still some problems in the accuracy of classification for nonlinear sample data, data with noise and outliers.

Based on the consistency search index and minimum value of rough set, these three indexes are used to calibrate the intermediate concept level. In this way, the original data table is decomposed into small data tables for hierarchical classification, and because of the clear physical meaning of the intermediate concept level, the comprehensibility of the model is greatly increased.
The PEP method does not need to separate the training set from the verifier, which is beneficial to the case of less data. Moreover, the pruning strategy is more efficient and faster than other methods. Because in the process of pruning, each subtree in the tree needs to be accessed at most once, and in the worst case, its computational time complexity is only linearly related to the number of non-leaf nodes of the non-pruning tree.

2. Decision Tree Reduction Method based on Rough Classification Degree

The basic principle of cluster analysis and the measurement of distance are between samples. The distance is absolute distance, Euclidean distance, Chebyshev distance and so on. Among them, the similarity between different samples can be measured by distance. When measuring the distance, it is often necessary to carry out standardized transformation in order to eliminate the influence of dimension.

After measuring the disorder degree of the data set, it is necessary to divide the data set and measure the entropy of the partition data set in order to determine whether the data set is divided correctly or not [2]. We calculate the information entropy for the result of each feature partition dataset, and then judge which feature partition dataset is the best way to divide the data set.

Because the essence of the diagnosis problem is a kind of mapping, the corresponding relationship between the symptoms of the fault and the fault is regarded as a mapping, so a feedforward network can be used to approximate the mapping relationship and realize the classification of the fault. However, in the diagnosis of complex systems, the practicability of the system is often reduced because of the large scale of the network and the long learning and training time.

An element in a resolution matrix is a collection of all attributes that can distinguish objects from each other. However, if it belongs to the same decision class, the value of the element in the discriminant matrix is an empty set. As can be seen by definition, it is a symmetric matrix and the elements on the main diagonal are empty sets. Therefore, as long as the upper half or the lower half of the triangle is considered, as is shown by equation (1) [3].

\[
(1 - B)^{\delta} = \sum_{k=0}^{\infty} \binom{\delta}{k} (-1)^{k} B^{k}
\]

In some cases, the data available may lack the value of some properties. You can also cross-use pruning and then pruning to form a combination. One of the strategies for missing attribute values is to use the most common value of this attribute to supplement the accuracy evaluation and improve the value of the missing attribute 2.6 classification. This method is simple and easy to use. Another strategy is that it is important to assign the accuracy of the estimation classification for each possible value of an attribute, which allows us to estimate a probability.

The judgment matrix \( A = (a_{ij}) \ n \times n \) is subjective and difficult to satisfy consistency. However, as long as the estimation is more reasonable, we can still take the corresponding eigenvector of its maximum eigenvalue as the weight vector of the index. This is the basis for determining the index weight by eigenvector method. The calculation method of the maximum Eigen value and the corresponding are eigenvector of the judgment matrix.

In essence, correlation analysis is a comparative analysis of geometric shapes between geometric curves [4]. The closer the geometric shape is, the closer the development and change trend is, and the greater the correlation degree is. The calculation data preprocessing of grey correlation degree. The original data can not be compared directly, because the order of magnitude difference is different, the unit of measurement is also different. In order to make it comparable.

The division point with the largest entropy of the classification class will be selected as the division of the next stage, and the entropy value will be used to represent the contribution of the partition variables to the output results. Further divide the divided area until the stop standard is met. In this algorithm, the missing value is ignored when searching for the cut-off point, and if the truncation point is not found for a property, the property is not processed.
Based on the adaptive value function, the genetic operation is applied to the population and the individual to realize the optimization and recombination of the individual structure in the population, and the optimal solution is approximated in the global range. Genetic algorithm combines the advantages of directional search and random search, avoids the defect that most classical optimization methods are easy to fall into local optimization based on the gradient or higher order derivative of objective function, and can achieve a good balance between region search and spatial expansion, as is shown by equation (2).

$$E_j = -\ln m \sum_{i=1}^{m} r_{ij} \ln r_{ij}, \quad j = 1, 2, \ldots, n$$ (2)

Knowledge element subject extraction needs the help of industry vocabulary, organization feature thesaurus, and the construction of index database to assist the decision-making of subject extraction. In the process of regional subject recognition, it also needs the regional database to assist the administrative regional level and the judgment of the region to which it belongs.

According to the definition of the general value reduction algorithm, the general value reduction algorithm is described as follows: for each rule in the rule set, for any conditional attribute in the rule, if the property is removed and the rule does not conflict with other rules in the collection, the conditional attribute can be removed from the rule [5]. The information table after attribute reduction and value reduction can be extracted by decision rules, and the process of generating rules from decision system $S = (U, A)$ by using a reduction set RED is quite direct.

3. Analysis of Data Reduction Method based on Rough Set Theory

There are three pruning strategies for decision tree: pruning based on cost complexity, pessimistic pruning and MDL pruning. Pruning based on cost complexity uses a separate set of samples for pruning, that is, unlike the set of samples used in tree construction, it is called pruning samples. In many cases, especially when the training set is very small, it is expected that all the samples will be used for both tree construction and tree pruning.

When the indexes are non-numerical index, and it is the measurement of similarity. The shortest distance method is a method to measure the similarity. The shortest distance method is used to carry out the basic process of cluster analysis. The results of cluster analysis may be different by using different methods to measure the degree of similarity [6]. In cluster analysis, the desired class number can be determined in advance.

Because feedforward neural network is one of the most widely used networks in many neural network models, and has strong pattern classification ability, it can obtain satisfactory convergence accuracy while realizing mapping from $R^n$ to $R^m$. It should be an ideal diagnostic model, because the conventional BP algorithm has some problems, such as slow convergence speed and may fall into local minima. Therefore, the three-layer forward BP network with additional momentum term is selected as the diagnostic model of the system, as is shown by equation (3).

$$\gamma_0(k) = \frac{m+\sigma M}{\Delta(k)+\sigma M}, \quad \sigma \in (0, 1), \quad k = 1, 2, \ldots, n, \quad i = 1, 2, \ldots, p$$ (3)

The enterprise analytical CRM management system based on association rule mining is an effective rule set for customer turnover analysis, which should be able to accurately reflect the customer turnover and predict whether the existing customers are lost or not, and what is the probability of loss, so as to take some measures to retain customers in time.

There are many complex objects in the actual system that are difficult to establish strict mathematical models, so the traditional control methods based on mathematical models are difficult to work. Fuzzy control simulates the fuzzy reasoning and decision-making process of human beings, and summarizes the control experience of operators into a series of language control rules, which has the characteristics of robustness and simplicity.
At present, with the continuous expansion of knowledge base, the complexity of knowledge reduction is becoming greater and greater. It can be said that it is a great problem to calculate all reduction and optimal reduction. Through analysis and research, the ideal method is to use genetic algorithm to find better knowledge reduction.

Most of the current decision tree algorithms use the theory based on information entropy in attribute selection. Because of the high time complexity of calculating information entropy, and the construction of decision tree is a cyclic recurrence process, multiple calculations will lead to complex calculation costs.

ID3 algorithm takes the maximum mutual information between attributes as the criterion for selecting attributes, and the greater the mutual information between attributes. It shows that the more dependent it is, the stronger it is. It is proved that the mutual information and compatibility relationship between attributes is unchanged for an information system [7]. Therefore, there is a compatible relationship with mutual information, that is to say, it can well reflect the dependence of attributes on decision attributes, as is shown by equation (4).

\[ u^{(0)} = \sum_{i=1}^{N} \alpha_i e_i + \sum_{i=N+1}^{M} \alpha_i e_i \]  

(4)

In the first pruning method, the improved classification C 'pruning of trees is given by stopping the construction of trees in advance. However, it is difficult to choose an appropriate threshold [8]. Higher appendix values will lead to simplified scalability of decision trees, while lower closed values may make trees simplified too little.

Decision tree is a simple way to determine the production capacity scheme. Decision trees can not only help people understand problems, but also help people solve problems. Decision tree is a method of listing the relevant steps of solving problems and the conditions and results of each step by illustrating them.

Once a set of attribute values are determined, a specific sales market can be identified, in which the default value of each attribute is a set of all possible values of this attribute. B is the set of sales transactions that specify specific bearing equipment, that is, pre-set bearing equipment, and all sales transactions in the history can be divided into multiple grades through data preprocessing technology.

4. Clustering Reduction Method Analysis of Rough Set and Decision Tree based on Weight Matrix Analysis

Although rough theory and fuzzy theory are both theories to describe the uncertainty of sets, fuzzy theory focuses on describing the uncertainty of elements in the set, while rough set theory focuses on the uncertainty between sets. The two are not contradictory and highly complementary.

Starting with an empty decision tree, select an attribute (classification attribute) as the test attribute [9]. The test attribute corresponds to the decision node in the decision tree. According to the value of the attribute, the training sample can be divided into the corresponding subset. If the subset is empty (no sample takes the value in the sample space), or the sample in the subset belongs to the same class (all the worthy samples in the sample space belong to the same classification), then the subset is the leaf node.

In this method, the weight of one index is related to two factors, one is the standard deviation, the other is the correlation coefficient between the index and other indicators. The smaller the standard deviation of a certain index is and the higher the correlation degree between the index and other indexes is, the smaller the weight of the index is, and the greater the weight of the index is. The correlation degree of the two indexes was measured by correlation coefficient.

Based on the theory of dominant relation rough set and grey correlation degree, a new method for comprehensive evaluation of weighted number is proposed. The dominance relation is used instead of the equivalent relation in classical rough set theory, which avoids the discretization of the original data and ensures the initial amount of information [10]. The grey system finds out the law by generating
and processing the small sample data, determines the correlation degree, and then determines the importance of each index.

It is actually a Boolean function with Meta variables; it is the combination of elements in the matrix term. According to the corresponding relationship between the resolution function and the reduction, the method of calculating the S reduction Red (M) of the information system can be obtained as follows by equation (5).

$$M = \frac{\eta_{ij}^2(\tau_j)}{\mu_{ij}^2(\tau_j)} = \chi^2_{ij}$$

The minimum E (C, Ai) value is used to divide the attribute interval E (C, Ai), (value is the most suitable for attribute division interval). Numerical attributes are usually represented as intervals. Depending on the position of the training object in the interval, the interval is divided into two parts. Then all the training objects are sorted according to the value, and we use the function (5) as the threshold to evaluate all the candidate objects. As we have already mentioned, the minimum E (C, Ai) value is the most suitable attribute for dividing the interval.

In the operation, random diversity groups and cross operations are conducive to expanding the search space, and with the acquisition of high fitness values, cross operations are conducive to exploring around these solutions. Genetic algorithm (GA) has the ability to jump out of the local optimal solution by searching in multiple directions by maintaining a potential solution.

According to the size of the index frequency, the numerical knowledge element agent extraction in the training database is processed in batches, and the subject extraction experiment and rule learning are carried out with the small but high frequency indexes, and then the index selection quantity is gradually increased and the rule learning is carried out again.

According to the obtained frequent itemsets, the corresponding strong association rules are generated. According to the definition of these rules, the minimum confidence threshold must be met. In addition, interesting metrics can also be used to help mine valuable association rule knowledge. Because the corresponding operation is very simple, the whole performance of mining association rules is determined by the operation processing of step one.

The problem of data classification can be regarded as a search problem, the database is regarded as a search space, and the classification algorithm is regarded as a search strategy.

5. Conclusion

After mining all the frequent itemsets from the database, it is easier to obtain the corresponding association rules. That is to say, it is necessary to produce strong association rules that meet the minimum support and minimum confidence. Random noise and some decisions depend only on a small amount of training data, which will lead to the decrease of classification accuracy of decision tree and excessive fitting of training data set. Solving these problems is mainly realized by pruning the decision tree. The current pruning algorithm is not ideal in accuracy and complexity and needs to be further improved.

References

[1] Kuncheva L I. Editing for the k-nearest neighbors rule by agentic. Pattern Recognition Letters, 2015,16:809-814.
[2] ZDZISLAW P. Rough set theory. Kunstliche Intelligenz, 2011, 5(3): 38-39.
[3] Zhao Weidong, Sheng Zhaohan, He Jianmin. The application of rough set in decision tree generation. Journal of Southeast University, 2000, 30 (4): 132-137.
[4] BLEYBERG M Z, ELUMALAI A. Using Rough Sets to Construct Sense Type Decision Trees for Text Categorization. IFSA World Congress and 20th VAFIPS International Conference, Vancouver, BC, Canada: July, 2001: Vol1, 19-24.
[5] Yuan Qinqin. Improvement and Application of decision Tree algorithm. Changan University,
2006.

[6] Qiao Mei, Han Wenxiu. Decision tree algorithm based on Rough set. Journal of Tianjin University, 2005, 38 (9): P842-846.

[7] Hong Jiarong, Ding Mingfeng, Li Xingyuan, A new decision tree inductive learning algorithm, Journal of computer Science, 2015, 18 (6): 470-474.

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

[9] WEI Jinmao, HUANG Dao, WANG Shuqin. Rough Set based Decision Tree. Proceedings of the 4th World Congress on Intelligent Control and Automation, Shanghai, China: IEEE, Jan, 2002: P426-431.

[10] Zhang Wenxiu, Qiu Guofang. Uncertain decision based on rough set. Beijing: Tsinghua University Press, 2005.