International Journal of Database Theory and Application
Vol.9, No.5 (2016), pp.1-8
http://dx.doi.org/10.14257/ijdta.2016.9.5.01

Research of Decision Tree Classification Algorithm in Data Mining

[1]Qing-yun Dai, [2]Chun-ping Zhang and [2]Hao Wu

[1]Dept. of Electric and Electronic Engineering, Shijiazhuang Vocational and Technology Institute, Shijiazhuang, HeBei, China, 050081
[2]Dept. of Information Engineering Shijiazhuang Vocational and Technology Institute, Shijiazhuang, HeBei, China, 050081
dddqqyy@126.com dqy_2003@126.com

Abstract

Decision tree algorithm is one of the most important classification measures in data mining. Decision tree classifier as one type of classifier is a flow- chart like tree structure, where each intenal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node represents a class. The method that a decision tree model is used to classify a record is to find a path that from root to leaf by measuring the attributes test, and the attribute on the leaf is classification result.

Keywords*: data mining; classification; decision tree*

1. Introduction

Recently, with the wide application of database and data warehouse as well as the rapid development of computer technology, the ability of people using information technology to collect data increases significantly. A large number of databases have being used for business management, government offices, scientific research and engineering development. It is a very difficult task to find valuable information or knowledge from huge mass storage of data, which raises Data Mining formed and developed rapidly. Data Mining is acquire and mine interesting knowledge from mass data, which is implicit, previously unknown and potentially useful information, then be represented as concept, rules, law, patterns and other forms.

Classification is a very important task in data mining. Target of classification is to find out a classification function or model, which can map a single record in database to a pre-assumed class. Categories can be used to forecast future data, based on trend description of the given data inferred automatically from historical data records. The most widely known classification algorithm is decision tree method, which is a tree structure of classifications.

2. Introduction of Decision Tree

Decision tree is the main technology used for classification and prediction. Decision tree learning is a typical inductive algorithm based on instance, which focus on classification rules displaying as decision trees inferred from a group of disorder and irregular instance. In top-down recursive way, it compares attributes between internal nodes of decision tree, judges the downward branches according to different attribute of the node, and draws a conclusion from leaf nodes in the decision tree. So from a root to a leaf node corresponds to a conjunctive rule, and the entire tree corresponds to a group of disjunctive expression rules [1]. Take the decision tree as a Boolean function. The input of the function is the object or all property of situation, and the output is the "yes" or "no" decision value. In the decision tree, each tree node corresponds to a property test, each leaf node corresponds to a Boolean value, and each branch represents one of the possible

ISSN: 2005-4270 IJDTA

values of testing attribute.

The most typical decision tree learning system is ID3, which originated in the concept learning system CLS, and finally evolved into C4.5 (C5.0), which can deal with continuous attributes. It is a learning guide, based on a decision tree composes of training subsets. If the tree fails to classify correctly all the given training subset, choose other training subset adding to the original subset, repeat it until the correct decision set. To train a number of training instance classification, a decision tree which can classify an unknown instance classification based on specific occurrence of attribute value sets. Using decision tree to classify instances, you can test gradually the value of the objects' properties starting at roots, and then going down the branch until reach a leaf node, in which class is the class of the object.

Decision tree is a widely used method of classification. There are multiple decision tree methods, such as ID3, C4.5, PUBLIC, CART, CN2, SLIQ, SPRINT *etc.* Most developed decision tree is a variant of the core algorithm. The following will introduce firstly the basic idea of the decision tree classification, the construction and pruning of the decision tree, and then describe in detail the algorithm of ID3 and C4.5, and the analysis and improvement of the decision tree algorithm [2].

3. Construction and Pruning of Decision Tree

Decision tree classification algorithm is usually divided into two steps: to construct and prune Decision Trees.

3.1. Construction of Decision Tree

The input of decision tree construction algorithm is a set of classic labeled examples. The result of structure is a binary tree or ternary tree. The internal nodes of a binary tree (non-leaf nodes) are usually represented as a logical judgment, such as in the form of (AI=VI) logical judgment, AI is the attribute, of which VI is a value. Tree's edge is the branch outcome of logic. The internal node of ternary tree is an attribute, of which edges are all values. Where there are several attribute values, there are several sides. Tree leaf nodes are category tag. Method of constructing decision tree is a top-down recursive structure. With the ternary tree as an example, its structural idea is starting to establish decision tree with single node represented the training sample. If the samples are all in the same class, it can be leaf nodes, and contents of nodes are the category tags. Otherwise, select an attribute based on certain strategy, divide example collections into several subsets In accordance with the attribute and values, and make all the examples in each subset has the same attribute value. Then deal with recursive process of each subset one by one. This idea is actually "divide and rule". So does the binary tree, and the only difference is how to choose better logical judgment.

The key to construct decision tree is how to choose better logical judgment or attribute. There can be many choices to the same set of examples. Research shows that, in general, the smaller the tree, the stronger forecasting ability. The key of constructing decision tree as small as possible, is to choose the proper attribute caused branch. Attribute selection depends on Impurity measurement method on the various examples of subset. Impurity measurement method includes Information Gain, Gain Ra-tio, Gini-index, distance measurement, X2 statistics, the weight of evidence, the minimum description length *etc*. [3]. Different measurement brings different effects, especially for multi valued attributes, select appropriate measurement method is greatly affected the result. ID3, C4.5, C5.0 algorithm use the concept of information gain to construct decision tree, while Gini-in-dex is used for CART x, where each classification decisions is relative to previously selected target classification.

3.2 Pruning of Decision Tree

The object of Data mining is real world data, which are generally not perfect. Maybe there are some missing values in attribute field; or lack essential data resulting to incomplete data; or data are inaccurate even wrongly, or containing noise, so it is necessary to discuss the problem of noise.

The basic decision tree construction algorithm does not consider the noise, so the generated decision tree fits completely with training examples, which will lead to excessive fitting and will destruct predictive performance. Pruning is a technique to overcome noise, at the same time it also can make the tree simplified and easy to understand.

3.3 Two Basic Pruning Strategy

1) Forward- Pruning is pruning before the decision tree's growth process is completed, when decide to continue dividing the impure training subset or shutdown.

For example, when some efficient statistic reaches a predetermined threshold, the node is no longer to continue splitting, and the internal node becomes a leaf node. Leaf nodes take the maximum frequency class as its logo, or may simply store the probability distribution function of examples. Forward- Pruning will stop the tree's work before it reaches full maturity, that is, the tree stops extension when it should not, or called horizon effect. Moreover, it is difficult to select an appropriate threshold. A higher threshold may lead to overly simplified tree, while a lower may make the tree simplify too little. Even so, it is worthy of studying the large-scale practice of Forward- Pruning, because it is quite efficient. It is expected to solve horizon effect in future algorithm.

2) Post-Pruning is pruning after the decision tree growth process is completed. It is a Fitting- and-simplifying of the two stage method. First generate a decision tree fitting completely with training example, and then trim the leaves of the tree from the bottom to the top, gradually to the root. When Pruning, it is used a test data set. After a leaf is cut, if the accuracy or other measure of the test set does not reduce (or worse), you can do it, or stop.

If the sub tree of the node should not be pruned according to certain rules, the evaluation rule from bottom to top will avoid the node to be pruned in the same way. Against it is the top-down strategies, that is, trim the node gradually from the root until no more can be done. The risk is a node is cut off according to certain rules, to which its subclasses should not be cut, however.

3.4 Principle of Tree Pruning Optimization

The Minimum Description Length Principle (MDL). The simplest explanation is expected as its idea, and approach is to get binary code of decision tree. The tree required least binary code is called "the best pruning tree".

The Expected Error Rate Minimization Principle. The idea is to select a branch of sub tree expected error rate minimization, that is, compare the possibility of error rate between prunning or not pruning by calculated internal node of tree, then choose.

The Principle of Occam's Razor. If not necessary, don't add entity. "Choose the simplest decision tree in compatible theories", the little a decision tree is, the easier to be understood, and the smaller cost its storage and transmission is.

4. ID3 Algorithm

The ID3 algorithm is a famous decision tree generation method proposed by Quinlan in 1986, which is now cited in the high rate. By using information entropy theory, ID3 algorithm select the gain property value of the maximum information in current sample concentration as the test attribute; Divide sample set based on the testing attribute value,

where there are different value in testing attribute, there are different sub sample set, at the same time, the nodes corresponding to of sample set on decision tree grow new leaf node. As a rule, the simpler decision tree structure is, the easier to generalize the law of things in nature. Expectations of average path from non-leaf node to descendant node is always the shortest, namely average depth of the generated decision tree is minimum, requires selecting better division in each node. The less uncertainty of the system, the fully transmit of information .By information theory, ID3 algorithm take uncertainty of divided sample sets as a standards to measure the quality of division. The bigger information gain value, the less uncertainty. Therefore, algorithm select information gain maximization in each non-leaf node as the testing attribute

ID3 algorithm divides mining sample set into training set and test set. Decision tree builds on the learning of training sample set. After decision tree generated, post- pruning the tree using the test sample set, and cut the leaf nodes which do not meet the conditions of the predictive accuracy of classification

4.1 Information Theory and Entropy

Information theory is established to solve the problem of information transfer (Communication) process by C.E Shannon. An information transmission system is composed of a sending end, a receiving end and a connection between the two channels [4]. Entropy is a measure of uncertainty of an attribute corresponding to the event. The smaller of an attribute's entropy, the higher of purity the subset divided. The bigger of an attribute's entropy, the greater the uncertainty information become.

4.2 Information Gain

Information gain is based on concept Entropy of information, refers to the reduce weight of desired information or entropy (usually use a "byte" to measure), sample classification according to determine to choose what kind of variable on what level. ID3 always choose the attribute of the highest information gain (or maximum entropy) as the test attribute of current node. This attribute makes information required by the results divided of classification minimum, and reflects minimum random or "impure" of the division. Information gain is calculated as follows:

Let S be s training dataset, m different values in S class identifier attribute, defining m be different set C_i (i = 1, 2, 3, m). Let Si be the sample number of C_i. The desired information required by a given sample classification can be get as following:

$$I(s_1, s_2, \ldots, s_m) = -\sum_{i=1}^{m} p_i \mathrm{lb}(p_i) \quad (1)$$

P_i is an arbitrary sample, belonging to the probability of C_i, generally estimated by s_i /s.

Set attribute A has n different values $\{a_1, a_2, \ldots, a_n\}$, using the attribute A ,divide S into n subset $\{S_1, S_2, \ldots, S_n\}$, in which S_j contains some samples of S, who have the value A_J in A. If A is to be the test attribute, the subset corresponds to the branches which grow out of nodes contains a collection of S.

Let s_{ij} be a sample number of class C_i of subset S_j. the entropy of subset divided by A is given by:

$$E(A) = -\sum_{j=1}^{n} \frac{s_{1j} + s_{2j} + \ldots + s_{mj}}{s} I(s_{1j}, s_{2j}, \ldots, s_{mj}) \quad (2)$$

According to the above desired information calculation formula, the desired information to a given subset of S_j is calculated by the following formula:

$$I(s_{1j}, s_{2j}, \ldots, s_{mj}) == -\sum_{i=1}^{m} p_{ij} \mathrm{lb}(p_{ij}) \quad (3)$$

As the metric of decision classify attribute, attribute A is information gain, which can

be determined by the following formula

$$Gain(A) = I(s_1, s_2, \cdots, s_m) - E(A) \quad (4)$$

4.3 Analysis of ID3 Algorithm

ID3 algorithm refines decision tree step by step with constant circulation treatment, until find a complete correct decision tree. The decision tree that ID3 algorithm constructs from top to bottom is a group of rules similar as" If ... Then rules". The original program is used to distinguish the chess walking, so there are only two categories, true or false, and the attribute value is a number of discrete finite value. While ID3 algorithm has been developed to allow more than two categories, and the attribute value can be integer or real. Its advantages and disadvantages are summarized as follows:

4.3.1 Advantages

ID3 uses the concept of information gain in the selection of important features from. The basic theory of algorithm is clear, and the algorithm is relatively simple. Its calculation time is the linear function of product of a number of examples, the characteristic number, and node number. Moreover, the search space is a complete hypothesis space, in which the objective function must be, so there is no danger of non-solution [5]. Overall using of the training data, rather than considering training cases one by one in the candidate cut algorithm, we can enjoy the advantage of the decision made by using statistical characteristics of all the training cases, against the noise.

4.3.2 Disadvantages

It is not quite reasonable that Information gain calculation depends on a large number of features of eigenvalue. A simple solution is to decompose the characteristics and transform all eigenvalue into binary feature. ID3 is more sensitive to noise. The mistake in the training examples of Quinlan define is noise. It includes two aspects, one is wrong of the characteristic value, the other is wrong of categories. When the training set is increased, ID3 decision tree will also change. It is inconvenient for asymptotic learning that features of information gain change with the increased example in the process of constructing decision tree.

5. C4.5 Algorithm

C4.5 (Classsification4.5) algorithm is introduced by Quinlan in 1993.It is the successor to the ID3 algorithm, and also be the foundation of many algorithms. Besides having the function of ID3 algorithm, C4.5 has new method and new function. When applied to a single decision tree algorithm, C4.5 algorithm has high classification accuracy rate and speed.

C4.5 algorithm add the concept of information gain proportion and continuous attribute as well as the treatment for vacancy of attribute value to ID3, also have a more mature approach to tree pruning. With different pruning techniques to avoid the overfitting of the tree, it overcomes the problem in the application of ID3. Firstly, choose attribute by information gain ratio rather than information gain, it overcomes the bad tendency to select more attribute. Secondly, it can deal with continuous attribute with the discrete way, and also can process deficient data. The knowledge representation of C4.5 algorithm is decision-making tree, which ultimately can form rules of production.

5.1 Concept of Information Gain Ratio

Information gain ratio develops on the basis of is the concept of information gaint. Information gain ratio of an attribute can be given by the following formula:

$$\text{GainRatio}(A) = \frac{\text{Gain}(A)}{\text{SplitI}(A)} \quad (5), \quad \text{in which}$$

$$\text{SplitI}(A) = -\sum_{j=1}^{V} p_j \log_2(p_j) \quad (6)$$

Here set attribute A having V different values of $\{a_1, a_2, \ldots, a_v\}$. Using attribute A, divide S into V subsets of $\{S_1, S_2, S_3\}$, in which S_j contains such samples in S , who have value a_j on A. If we split samples with the benchmark of the value of attribute A, SplitI(A) is the concept of entropy as above.

5.2. Combined with Continuous Value Attributes

ID3 algorithm assumes initially attribute as discrete value, but in practical application, a lot of attribute values are continuous. For the continuous attribute values, C4.5 first sort the data collection according to the attribute value, then divide them dynamically with different threshold values. When the input changes, take the midpoint of two actual values as a threshold, then take two divisions, in which all the samples are. Get all possible thresholds, gain and gain ratio, and every attribute will become two values, which is less than or more than or equal to the threshold value [6].

Under condition of the attribute having continuous value, for example A has continuous attribute value, $a_1, a_2, \ldots a_m$. can be arranged in ascending order in the training set. If A has N values, all records were divided with each value of v_j(j = 1, 2, ... , n). These records are divided into two parts: one part is in the range of VJ, while another portion is greater than v_j. Calculate respectively the gain ratio to each of the division, and choose the division of the maximum gain to discretize the corresponding attribute.

5.3. Production of Regular

Once the tree is established, we can transfer the tree into If- Then rule, which is stored in a two-dimensional array. Each row represents a rule of a tree, namely a path from root to leaf. Each column in the table stores nodes of a tree.

6. Analysis for Decision Tree Classification Algorithm

Since been introduced, Classifying algorithm based on decision tree has dozens of species by now. The various algorithms have their own merits on execution speed, scalability, comprehensibility of output result and accuracy of the Classification forecasting [7]. The development of decision tree classification algorithm can be divided into the following stages:

First, the earliest ID3 algorithm select of sample set of test attribute using information entropy principle, which can only deal with samples of discrete attribute and complete attribute value range of, generate decision tree shaped like multiway tree. Subsequently been improved, the C4.5 algorithm can directly deal with continuous attribute, and handle training samples of vacancy of attribute value.

ID3 series algorithm and C4.5 series algorithm mine information as much as possible in learning of training set, but the generated decision tree has too much branches and larger scale. In order to simplify the decision tree algorithm and improve its efficiency, a number of other decision tree algorithm emerged [8].

The principal advantage of decision tree classification algorithm is to generate understandable rules, to handle various data types without relatively large computation. The decision tree can show clearly which attributes are more important. At the same time,

it also has some shortcomings such as it is more difficult to forecast continuation field. It needs to do a lot of pretreatment work for the chronological data. When the categories are too much, errors may increase faster

7. Decision Tree Classification Algorithm

In view of the decision of tree classification algorithm, the following will introduce algorithm improvements about attribute selection and discretization of continuous attribute

1) Attribute Selection. In order to avoid the noise and the interference attribute affecting the classification of data, sort the importance of attribute before establishing of a decision tree ,and then train the most important attributes and test their prediction accuracy with neural network technology. Then plus-minus respectively an adjacent attribute to both ends according to importance order ,train and test, and compare with original inspection results, repeated this several times until you find n attributes of the best classification results, then get it.

2) Discretization of continuous attributes. Discretization is an effective method to deal with continuity in classification process. The efficiency and effectiveness of discretization directly affects the efficiency and performance of subsequent machine learning algorithm. Many classification rules, such as ID3 algorithm, can only deal with discrete attribute, while the discretization of continuous attributes is a necessary step. Although C4.5 can deal with continuous attributes not discretization, which is an important step of system integration. Discretization can not only shorten the time of derivating classifier, but also help to improve the understandability of data, and get higher accuracy of classification rules.

From the angle of optimization discretization methods can be classified into two categories: local and global methods. Local method only disperses one attribute every time, in contrast global method simultaneously make all attributes discretization. Local discretization method is relatively simple, while the global method can often get better results for considering the interaction of attributes, but it's computation cost is very high.

8. Conclusions

According to the research and analysis above, we can draw a conclusion that the key to construct a good decision tree is to choose better logical judgment and attributes. With database and data warehouse technology widespread application and the increasing magnanimity of the data, the efficiency of decision tree technology needs to be improved. It is necessary to improve appropriately data mining method of decision tree on applicability and noise-tolerance, to solve practical problems better.

References

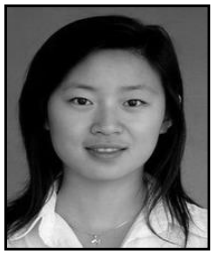
[1] V. W. Aalst, "Exterminating the Dynamic Change Bug: A Concrete Approach to Support Workflow Change", Eindhoven, UK: Eindhoven University of Technology, **(2000)**.
[2] S. Rinderle, M. Reichert and P. Dadam, "Correctness Criteria for Dynamic Changes in Workflow Systems", Data & Knowledge Engineering, vol. 50, no. 1, pp. 9-34.
[3] L. Bo, A. Abbass and B. Mckay, "Classification rule discovery with ant colony optimization", IEEE Computational Intelligence Bulletin, vol. 3, no. 1, **(2004)**, pp. 31- 35
[4] S. Xin, H. Chu and F. Qian, "A Supervised Classification Method Based on Conditional Random Fields with Multi-scale Region Connection Calculus Model for SAR Image", IEEE Geoscience and Remote Sensing Letters, vol. 8, no. 3, **(2011)**, pp. 497-501.
[5] X. Wu, V. Kumar and J. R. Quinlan, "Top 10 algorithms in data mining", Knowledge and Information Systems, vol. 14, no. 1, **(2008)**, pp. 1-37.
[6] M. Kantardzie, "Data mining: concepts, models, methods, and algorithms", Journal Computer Information Science Engineering, vol. 5, no. 4, **(2005)**, pp. 394-395.
[7] N. Bissantz and J. Hagedorn, "Data mining", Business and Information Systems Engineering, vol. 1, **(2009)**, pp. 118-122.

[8] R. S. Parpinelli, H. S. Lopes and A. A. Freitas, "Data Mining with an Ant Colony Optimization Algorithm", IEEE Trans. On Evolutionary Computation, special issue on Ant Colony algorithms, **(2002)**.

Authors

Qing-yun Dai, received the B. Eng. degree in Electrical Engineering from BaoTou Institute of Iron & Steel in BaoTou China and the M. Eng. degree in Computer Engineering from YanShan University in Qinhuangdao china. He is currently a lecturer in Shijiazhuang Vocational and Technology Institute in Shijiazhuang, China. His main research interests are in the areas of computer networks, and data mining.

Chun-ping Zhang, Zhang received the B. Eng. degree in Computer Application from Hebei Institute of Technology in Tangshan China and the M. Eng. degree in Computer Application from Hebei University of Technology in Tianjin China. She is currently a lecturer in Shijiazhuang Vocational and Technology Institute in Shijiazhuang, China. Her main research interests are in the areas of computer software, and embedded system.



Hao Wu, received the B. Eng. degree in Computer Science and Application from Hebei University of Economics and Business in Shijiazhuang China and the M. Eng. degree in Computer Science and Technology from Tianjin Polytechnic University in Tianjin China. She is currently a lecturer in Shijiazhuang Vocational and Technology Institute in Shijiazhuang, China. Her main research interests are in the areas of computer network, and database technology.

