Advances in Economics, Business and Management Research, volume 203
Proceedings of the 2021 3rd International Conference on Economic Management and Cultural Industry (ICEMCI 2021)

Applications of the Decision Tree in Business Field

Zixuan Zhang

[1] *Malvern Qingdao College, Qingdao Shandong province, 266000*
**Corresponding author. Email:1837497482@qq.com*

ABSTRACT
With the advent of the information age and the development of the Internet, the data show explosive growth, machine learning is becoming a more and more popular field. This paper lists some different applications of decision trees in the business fields which are related to personal credit and the stock market, The data of this paper are all collected from the official website. With two examples, two different algorithms are used to establish the models which are the C4.5 algorithm and the C5.0 algorithm respectively. Therefore, two algorithms would be discussed and compared in this paper to illustrate both their advantages and disadvantages. C5.0 is an advanced algorithm based on the invention of C4.5. For some variables with a relatively small number of indicators and simple logical relationships, the prediction accuracy of the decision tree algorithm is high, which makes this algorithm suitable for analyzing the data.

Keywords: *decision tree, machine learning, stock market, personal credit*

1. INTRODUCTION

Along with the computer technology, especially the rising popularity of database technology, people gradually started to attach importance to hidden in the secret behind the big data, more and more attention focused on how to quickly and efficiently find knowledge treasure hidden in the data, to serve his own business, service for the progress of the society. Lots of algorithms start to be invented such as decision trees, which can aim to apply in the business area. This paper mainly introduces decision trees including its classification and establishment. Also, two applications of decision tree in business fields are introduced which are segmentation of customers about the personal credit and forecasting and judging the behaviors of the stock market. Then, this paper discusses three different algorithms of decision trees then compares the advantages and drawbacks of using them, which can illustrate that C5.0 is the latest algorithm. Therefore, this paper aims to emphasize the business field's applications and make comparisons between different algorithms to find out the differences between them, and the situations of using certain suitable models.

2. DECISION TREE

The decision tree is a kind of nonlinear discriminant analysis method, from the field of machine learning has gradually developed a classification function approximation method, is essentially by establishing a series of rules in a tree on the sample classification process, according to the results of the properties and classification of the known samples generated tree classifier rules and using them to classify unknown data and forecast, which is a typical supervised single classifier. In 1966, scholar Hunt developed the first Concept Learning System, which is the basis of many decision tree learning algorithms, and proposed the application of decision tree to concept learning [1].

The technique for the decision tree is started from this framework. To lessen the intricacy of the decision tree, it is important to figure out how to create a decision tree with a less difficult design. Researcher Magidson proposed the CHAID algorithm in 1975. This calculation can just arrangement with input factors of class type, so the calculation requires discretization of info factors.

The decision tree recursively builds the model using the top-down method, deriving classification rules from irregular and disordered input, and finally presenting the tree structure. Every time the decision tree is divided, an attribute value comparison is performed at the node to decide the next branch direction until the leaf node is reached.The application of the decision tree classification technique is split into two parts: tree construction and application. The decision tree algorithm's focus is on obtaining knowledge from empirical data, performing machine learning, building models, or constructing classifiers. It's frequently broken down into decision tree construction [2]. The application is relatively simple, using the established decision tree model to classify or predict the new data. After the main body of the tree is built, the next step is to prune it [3].

3. APPLICATIONS OF DECISION TREE

3.1. Application in personal credit

One of the main areas of the decision tree that can be utilized is personal credit. The personal consumption credit business appeared relatively late, but the development speed is very fast. In recent years, the government has formulated a large number of macroeconomic policies aimed at expanding domestic demand, which has vigorously promoted the development of the personal consumption credit business. The business scope has developed to house purchase, house decoration, car purchase, and so on. Of course, the biggest proportion is the loans related to houses and vehicles. As per the most recent report delivered by The Boston Counseling Gathering (BCG), the remarkable individual utilization advances in China developed at a normal yearly pace of 29% from 2005 to 2010, and the current 7 trillion yuan market is relied upon to develop at a normal yearly pace of 24% throughout the following five years. Also, came to around 21 trillion yuan in 2015[4].

At present, foreign financial institutions generally adopt methods 5C and 1S [5]. This method uses "5C and 1S" related parameters as the most important indicators to measure personal consumption credit and judge the credit status of borrowers. The 5C includes Capacity, Character, Collateral, Capital, Condition, Stability, Stability. Through a long time of practice, the "5C and 1S" credit rating model" is effective and feasible.

There is the data set of a German bank's personal credit customers, with a total of 1000 personal credit records, each information is composed of 21 attributes, the first 20 attributes are used to measure user indicators, including age, occupation, marriage, education, credit history and so on. The last attribute is the category attribute, which indicates the credit level of the customer and includes two categories: "Good customer" and "poor Customer". The so-called "good credit customer" means that the customer has the potential to repay on schedule and the credit agency is willing to provide credit services for the customer [6].

For building a decision tree, firstly, the initial information entropy of the training sample set is calculated. There are 800 samples in the training sample set, that is, n=800 differential customers. The number of good customers is 561, and the number of differential customers is 239. The improved expression of information entropy can be used to calculate the information entropy of sample S: the information gain rate of the remaining 19 attributes can be calculated by the same method. According to the calculated results, the attribute C1 with the maximum information gain rate is selected, and then branches are created respectively according to the four values of C1, to divide the training samples into four subsets, and each branch creates new nodes for its subsets. And then repeat the above steps for each newly generated node. Until finally all nodes meet the following two conditions: (1) the record items of each subset of the training set all belong to the same category or a certain category accounts for the majority; (2) The generated tree node satisfies some terminating split criterion. The resulting spanning tree consists of 105 nodes with a height of 9. After the decision tree is established, the running time for this program is 0.312s, and the correct rate of the training set is 80%, and the correct rate of the test set is 77.97%.

To facilitate the bank employee in determining customer credit situation, can be more intuitive, more convenient according to the judgment of the information provided by the customer, make decisions, based on the German bank real customer data as the sample, using the improved algorithm model, and on this basis made a customer's credit rating to predict the credit status of new customers.

3.2. Application in stock market

In China, the securities exchange is the result of the market economy and has been supported by numerous financial backers since its development. The securities exchange has many capacities. For instance, the recorded organizations can assemble the inactive assets in the general public and put them into social proliferation through giving stocks. Simultaneously, in the areas with created capital business sectors, the change pattern of the securities exchange can mirror the present status of monetary turn of events and the pattern of future financial turn of events. For individual financial backers, the securities exchange is a significant channel to acquire capital appreciation through venture and monetary administration. Since the rise of the securities exchange, financial specialists have advanced numerous examination techniques, wanting to make exact investigations and forecasts of the pattern of the financial exchange. With the advancement of information base innovation, information mining innovation has been delivered and grown quickly.

Among the 200 listed companies in the A-stock market in 2012, 50 companies have the stocks with the best comprehensive performance in the A-stock market, 50 companies have the stocks with the worst performance, and the other 100 companies have the stocks with the average performance randomly selected, among which 50 companies' stocks are listed in Shanghai stock market and 50 are listed in Shenzhen stock market [7]. Taking the comprehensive performance grade of the stock as the output variable, it is marked as "excellent", "general" and "poor" respectively, and the C5.0 decision tree is used to establish the classification prediction model. The classification model is established through the training sample set, and then the test sample set is used to verify the validity and accuracy of the model.

80% of the samples in the sample set are randomly selected as training samples to build a decision tree model, and 20% of the samples are used as test samples. Pruning Severity determined the Pruning degree of the generated decision tree and is set to 80. The Minimum Records per child Branch is used to set the number of branches to be split. The decision tree splitting process will continue only if two or more subsequent branches have at least the minimum number of records. For the minimum number of records, this is set to 2. To prune the decision tree, it uses global pruning. When global pruning is used, the system will treat the decision tree as a whole during pruning. Earnings per share growth rate, return on equity, cash flow to debt ratio, asset-liability ratio, cash to debt ratio, and current ratio are all leaf nodes of this decision tree. First of all, the selected 200 listed companies are divided into three categories according to their comprehensive performance grades. Descriptive statistical analysis is conducted on the financial indicators of each category, and the mean value and standard deviation range of different financial ratio indicators of each type of listed company are preliminarily understood.The wrong prediction rate of 7.55%. It can be seen from the prediction results that the prediction accuracy is high.

In conclusion, the growth rate of earnings per share, the ratio of cash flow to liabilities, the ratio of assets to liabilities, and the ratio of liquidity have a great impact on the comprehensive performance of listed companies, among which the most important financial index is the growth rate of earnings per share. According to the decision tree rule, when the growth rate of earnings per share is between (-135.897%, 66.450%) and the return on equity is >=18.570%, the company's stock performance is excellent. When the return on equity reaches a certain value, the performance of the stock will decrease with the increase of the growth rate of earnings per share, possibly because the growth ability of listed companies in the mature stage is not as good as that of companies in the growth stage, but the performance of listed companies in the mature stage is better.

4. ANALYSIS

ID3 algorithm is an information entropy decision tree learning method proposed by Quinlan et al. in 1986 [8]. But it also has some problems: the calculation of information gain depends too much on the characteristics with more attribute values, but the attributes with more attribute values are not necessarily the best; the ability to resist noise interference is also poor. Because of the defects of the ID3 algorithm, Quinlan then proposed the C4.5 algorithm [9]. It inherits the ID3 algorithm's benefits. This algorithm is not only more accurate but also faster than the ID3 algorithm. Despite the fact that the C4.5 method bypasses several components of the ID3 algorithm's bottleneck, and the classification rules generated are relatively accurate and easy to understand, the core idea of C4.5 algorithm still remains in the category of "information entropy", and generates multi-tree. However, the disadvantages are obvious: the C4.5 algorithm can only process data sets that reside in memory. If the training set is too large and exceeds the memory capacity, the algorithm can do nothing. C4.5 algorithm efficiency is low, because in the process of splitting, looking for continuous attributes of the best discriminant ability measurement, all its division point information gain rates are needed to calculate, which results in a great increase in the computation time of the algorithm. If we can find the right division point, save some unnecessary division point information gain rates calculation, a lot of computational time will be saved, and the operating efficiency will be improved.

The key advantages of the C5.0 approach are as follows: C5.0 model is particularly robust when processing data sets with missing data and many input variables, and it builds decision trees quickly. The C5.0 decision tree model has a high level of accuracy, a more dependable result, and a higher reference value. The prediction accuracy of the C5.0 decision tree method is higher, making it better suited to issues like stock grade classification prediction. The decision tree method is ideal for situations when the number of indicators is not excessively big and the logical relationship between each indication is not excessively complicated. The number of indicators used to classify stock grades is relatively minimal, the logical relationship between the indicator variables is straightforward, and the decision tree algorithm's forecast accuracy is high.

5. CONCLUSION

In the market economy environment, to adapt to the competition situation in the new era, commercial banks must timely develop products that meet the new needs of customers. Therefore, the decision tree is a powerful tool. It is a classification technology that can help the product development department to analyze the product demand of existing customers, for example, predicting the demand for a type of new products, and developing the sales status of a type of new products in the market, improving the success rate of sales.

As a non-statistical machine learning algorithm, a decision tree has many advantages in terms of classification. However, modern technology could not achieve establishing perfect decision trees, there is still a lot of space for improvement of decision trees. Firstly, the classification rules are relatively complex. The local greedy algorithm is a rule often used in decision tree generation. When splitting nodes, only one attribute is selected for analysis at a time, which leads to complex generation rules. Also, there is the possibility of over-fitting. In the process of decision tree generation, sometimes the classification design is too complex, which leads to this problem which only fits one kind of model and cannot predict others preciously. Nevertheless, the decision tree can still have a wide range of applications in diverse fields. This paper only discusses limited applications of decision trees, does not make comparisons with other approaches of machine learning, and has not made improvements of algorithms that can increase the accuracy and save time during the operation. For future researches, other approaches should be adopted to build models, analyze each benefit and compare the suitable circumstances of using them, and introduce a wider range of applications.

AUTHORS' CONTRIBUTIONS

This paper is independently completed by Zixuan Zhang.

ACKNOWLEDGMENTS

First of all, I would like to thank Dr. Steve who is an expert in machine learning and business analysis, for teaching me professional knowledge about establishing models. Also teacher assistant Yuhui provided plenty of useful books for the machine learning searching and gave me suggestions about the topics decision. Finally, I would like to thank our team members and friends. During this period, we discussed and overcame the difficulties in the paper. Their support and encouragement made a great contribution to the completion of the paper.

REFERENCES

[1] Kira, K & Rendell L. The feature selection problem: traditional methods and a new algorithm [C] Proceedings of the 9th National Conference on Artificial Intelligence, 2008:129-134.

[2] Guishu, Chen Peiling, Song Hang. Review on Decision Tree Classification Algorithm [J] Science and Technology Square, 2007: 9-12.

[3] Esposito, F. & Malerba, D. & Semeraro. G. Simplifying decision trees by pruning and grafting: New results [J] In Lavrač, N. & Wrobel, S. (Eds.), Proc of 193 8th European Conf on Machine Learning (pp. 287–290). Heraclion, 1995.

[4] Gao Chen. China's personal consumption credit will reach 21 trillion in 2015. [N] Beijing Times, 2011-09-02 (3).

[5] Wang Xiaofeng. Research on the Risk prevention of China's Commercial Banks' Personal Automobile Consumer Credit [D] Soochow University, 2007.

[6] "UCI Machine Learning Repository" [DB/OL] http://archive.ics.uci.edu, 2015.

[7] Yuyu Tao. Application of Decision Tree And Neural Network in Stock Classification Prediction. Hangzhou Dianzi University,2013. Data center of Sina Finance. http://vip.stock.finance.sina.com.cn/datacenter/hqstat.html

[8] J. R. Quinlan. Induction of decision trees. Machine Learning, 1986, 1(1):81-106.

[9] J. R. Quinlan. C4.5: Programs for Machine Learning [M] Morgan Kaufmann Publisher, SanMa- teo, CA, 1993.