Optimization of Decision Tree Machine Learning Strategy in Data Analysis

Sun Zhifang, Li Yi
Shandong University of Engineering and Vocational Technology, Jinan, China, 250200

Abstract. Aiming at the problem of too coarse matching of machine learning decision tree models in the field of data mining and low prediction accuracy, the corresponding improved optimization strategies are proposed. First, the field matching degree of data is further improved by discretizing continuous attributes in multiple intervals. Then, the method makes the selection of business attributes more reasonable in the downward splitting process of the model by compensating the weight of feature attributes by business sensitivity indicators. Finally, the data classification rule transformation is used to further improve the data prediction accuracy of the model. The experimental results show that the introduction of the tree model generated by the business sensitivity index is more concise. In addition, the business pertinence and data classification capabilities are stronger. The results show that the transformed and upgraded data classification rules can effectively improve the accuracy of data prediction compared with the traditional optimization algorithm.

1. Introduction
Data mining and data analysis based on the decision tree model is an important application direction of current big data [1-5]. Most of the algorithms and improvement strategies in the existing decision tree research theories are limited to specific scenarios, and the scope of application is very narrow and many theories are still immature [6-9]. There are some shortcomings in practical applications. Moreover, in many research theories, the accuracy of data prediction is generally not high, and the ability of data classification is low, which makes data mining unable to truly grasp the essence of things [10-14]. How to find a mature, reliable and superior performance data prediction algorithm is an important research direction in the field of data application.

In discussing the process of discretization of feature attributes, this paper proposes a method to divide continuous attributes into discretization by probabilistic method. In addition, it also divides the continuous attributes into multiple reasonable sub-interval segments to make up for the shortcomings of the discretization algorithm. By analyzing the core degree of feature attributes, a calculation method is proposed to measure the weight of feature attributes through the correlation index.

In the reminder of this paper, we first introduce the data classification and feature attribute discretization and then analyze attribute sensitivity of experimental data. Finally, two typical tests are carried out to validate the performance of the proposed method.

2. Decision tree data classification and feature attribute discretization
The decision tree originated from the computer programming structure. It is an excellent classification model in the field of data analysis. The model can be used for machine learning of native data thus obtaining the laws and characteristics of the data. In this way, it can predict the outcome and trend of
unknown data. Decision process of tree creation represents the analysis and processing process of the problem. In the process of creating a decision tree, the concepts of entropy and information gain in information management are used as the basis for decision tree generation. The essence of the decision tree data classification process goes through the training data set induction, learning and reasoning, and the tree model growth process of classification prediction.

For the problem of whether the decision tree model can truly have data prediction ability, it is necessary to use unknown data to test. Generally speaking, the decision tree model generation process includes training samples and verification samples. The verification samples are mainly used to test the prediction ability of the model. If it is better, the prediction rules must be corrected. A good decision model must not only discover the data rules of the training examples, but also accurately classify the test examples.

3. Attribute sensitivity analysis of experimental data

Using information gain to select attributes and downward splitting tends to select attributes with more values. In many cases, those attributes with more values are often not the core attributes, which will cause the predicted results of the generated decision tree to be different from the actual. In response to this situation, literature [8] describes a method of using information gain rate to select attributes. However, it is not efficient when processing data with large sizes. In this paper, we consider introducing business sensitivity indicators to measure each characteristic attribute.

The business sensitivity index $s$ is a variable in the interval $[0,1]$. The main function of this index is to serve as a correction value to strengthen the weight of important attributes and reduce the weight of non-important attributes. So, there will not be too many factors in the process of using decision tree generation. Value non-important attributes cover up the phenomenon of less important attributes. The value of this variable mainly depends on whether the split attribute is a core attribute. If it is a core attribute, its value is relatively small. And if it is a non-core attribute, its value is relatively large. Hence, the weight of this attribute is inversely proportional to this variable. A concept of association degree can be used to measure whether the split attribute is a core attribute. The greater the association degree, the greater the core degree of the split attribute. The measurement of association degree $r$ is as follows:

$$r = \frac{|x^p_i - x^n_i|}{v_i}$$  \hspace{1cm} (1)

where $x^p_i$ represents the number of samples of positive examples in the attribute; $x^n_i$ represents the number of samples of negative examples in the attribute. The larger the value of $|x^p_i - x^n_i|$, the stronger the relationship between the attribute and the employment. And $v_i$ represents the attribute range.

According to the above correlation calculation formula $r$, we can define the core correlation value of a similar split attribute (k), which is the business sensitivity index $s$ as:

$$s(k) = \frac{1}{n_k} \sum_{i=1}^{n_k} \frac{|x^p_i - x^n_i|}{v_i}$$  \hspace{1cm} (2)

4. Experiments

The decision tree model developed by self-coding is used as a prototype tool to classify and test more sample data of job applicants. In the experiment, different model structures are used for classification testing. There are 3 kinds of experimental models, namely: 1 denotes the original model of optimization without any processing; 2 contains an improved model of business sensitivity indicators; 3 contains the ultimate model of business sensitivity indicators and modified data classification rules. The experimental classification output results of the test samples on the three models are as follows As shown in Fig. 1 and Fig. 2, it can be seen from the output results that when the sample size is still
relatively small, the classification accuracy of the three models fluctuates greatly, which cannot reflect the actual classification ability of the model. As the number of experimental samples increases, the data classification capabilities of the three models are gradually reflected, especially when the sample data volume reaches a certain level, the accuracy of the three models is basically stable within a certain range, and the data classification capabilities of the three models are very clear: the Ultimate model > Improved model > Original model. Therefore, the experimental results prove the effectiveness of this paper on the optimization strategy of the decision tree model based on machine learning.

Fig 1 The experimental results in the first test.

Fig. 2 The experimental results in the second test.
5. Conclusion

This paper improves the data prediction ability of the decision tree from several aspects such as data classification, sensitivity index analysis, and classification rule modification. Experimental analysis has verified the effectiveness of the optimization and improvement strategy. Compared with the traditional decision tree ID3 algorithm, the classification rules are improved and wrong classification rules are discarded. The classification ability is strengthened, so the data classification accuracy rate is higher. Compared with the traditional decision tree C4.5 algorithm, the continuous attributes can only be divided. Accordingly, this optimization scheme can discretize continuous attributes into multiple data intervals. So, its discrete process of continuous attributes is more reasonable and accurate, which can meet broader data processing needs. How to further improve the accuracy of data prediction is the direction of future research in this subject.

References

[1] YIN Xiao, WANG Mingyu. Research on scholarship evaluation system based on decision tree algorithm [J]. Computer Knowledge and Technology, 2015, 11(9): 11-13.

[2] Anna Atramentov, Hector Leiva, Vasant Honavar. Multi-Relational Decision Tree Algorithm - Implementation and Experiments[J]. In Second IEEE and ACM International Symposium on Mixed and Augmented Reality, 2003, 8(2):296 - 297.

[3] SALZBERG, STEVEN. Locating Protein Coding Regions in Human DNA using a Decision Tree Algorithm[J]. Journal of Computational Biology A Journal of Computational Molecular Cell Biology, 1995, 2(3):473.

[4] Chandra B, Varghese P P. Fuzzy SLIQ decision tree algorithm[J]. IEEE Transactions on Systems Man & Cybernetics Part B, 2008, 38(5):1294-1301.

[5] Amuthan, Prabakar, Muniyandi, et al. Network Anomaly Detection by Cascading K-Means Clustering and C4.5 Decision Tree algorithm[J]. Procedia Engineering, 2012.

[6] Liang X , Qu F , Yang Y , et al. An Improved ID3 Decision Tree Algorithm Based on Attribute Weighted[C]// 2nd International Conference on Civil, Materials and Environmental Sciences. 2009.

[7] Lee V J , Lye D C , Sun Y , et al. Decision tree algorithm in deciding hospitalization for adult patients with dengue haemorrhagic fever in Singapore[J]. Tropical Medicine & International Health, 2010, 14(9):1154-1159.

[8] Harvinder Chauhan, Anu Chauhan. Implementation of decision tree algorithm c4.5[J]. American Journal of Sports Medicine, 2013, 39(12):2611-2618.

[9] Ding B, Zheng Y , Zang S . A New Decision Tree Algorithm Based on Rough Set Theory[C]// Conference on Information Processing. IEEE, 2009.

[10] Aboul, Ella, Hassanien. Classification and Feature Selection of Breast Cancer Data based on Decision Tree Algorithm[J]. Studies in Informatics and Control, 2003.

[11] Yi W , Lu M , Liu Z . Multi-valued attribute and multi-labeled data decision tree algorithm[J]. Pattern Recognition & Artificial Intelligence, 2011, 2(2):67-74.

[12] Yi W , Lu M , Liu Z . Multi-valued attribute and multi-labeled data decision tree algorithm. Int J Mach Learn Cybern 2:67-74[J]. International Journal of Machine Learning & Cybernetics, 2011, 2(2):67-74.

[13] Chen X J , Zhang Z G , Tong Y . An Improved ID3 Decision Tree Algorithm[J]. Advanced Materials Research, 2014, 962-965:2842-2847.

[14] Wang Y Y , Li J . Feature-selection ability of the decision-tree algorithm and the impact of feature-selection/extraction on decision-tree results based on hyperspectral data[J]. International Journal of Remote Sensing, 2008, 29(10):2993-3010.