ISSN: 2414-1895 DOI: 10.6919/ICJE.201908_5(9).0023

Decision tree improvement algorithm and its application

Jian Di [a], Yinghui Xu [b]

School of Control and Computer Engineering, North China Electric Power University, Baoding 071000, China.

[a]Dijian6880@163.com, [b]Xuyinghui0925@.163.com

Abstract

Aiming at the problems of low efficiency and excessive fitting in data mining classification processing of decision tree algorithm. Therefore, in the process of data mining, the C4.5 algorithm was deeply studied and an improved algorithm, namely B-C4.5 algorithm, was proposed. The main idea of the proposed algorithm is a branch of the improved C4.5 algorithm and the Pruning strategy measure and adjust the C4.5 algorithm in the attribute information gain rate scope, comparing the information gain and probability is obtained by bayesian classifier, use a simplified CCP (Cost-Complexity Pruning) method and evaluation standard, the procedure of the subtree root node has to generate the decision tree surface five check five gain value, to determine whether to remove the decision tree nodes and branches. Simulation experiments are conducted on the improved C4.5 algorithm and the traditional algorithm. The results showed that the improved C4.5 algorithm has a significant improvement in execution time, which is 8.75% shorter than the traditional algorithm. With the increase of the number of experiments, the accuracy rate of the improved algorithm reaches more than 90%.

Keywords

C4.5; B-C4.5; Pruning strategies; CCP; Decision tree.

1. Introduction

With the continuous popularization of computer and the wide application of network technology and database technology, various industries have accumulated a large number of data, how to extract valuable information from these vast sea of data, has become an urgent problem to be solved. Data mining is the exploration process of big data set and reveals its hidden laws. It integrates many technologies and is an important branch of computer science. Among them, classification analysis is one of the important analysis techniques in data mining.Classification analysis is to discover classification rules according to the characteristics of existing data sample sets, and construct classification functions or classifiers, so as to assign categories to the samples of unknown categories to better assist decision-making.

The decision tree method in data mining has the advantages of easy to understand rules, clear display of important fields and less computation. The application of this method does not require the forecaster to have too much prior knowledge in the professional field.

There are many kinds of prediction methods based on decision tree, ID3 algorithm and C4.5 algorithm are the two main algorithms [1]. Among them, C4.5 algorithm improves ID3 algorithm by means of information gain rate and solves the disadvantages of ID3 algorithm that cannot handle continuous attributes and can easily choose values with more values as splitting criteria, so as to make it have better adaptability. In the application process, since C4.5 algorithm constructs the decision tree in a locally optimal way according to the greedy strategy, the decision tree under this method is not necessarily the global optimal.

This paper deals with decision tree C4.5 The branching and pruning strategies of the algorithm are improved to achieve more efficient, clear and reasonable classification results [2]. In order to obtain a more efficient and accurate decision tree model, this paper combines the boundary theorem and the bayesian classifier to obtain a more accurate segmentation cut point, and uses the simplified CCP(Cost- Complexity Pruning) method and evaluation standard to prune the decision tree to improve the classification efficiency.

2. Related Works

In the past few years, decision trees have been widely studied and applied. The decision tree model is a collection of a class of algorithms. Among the top ten data mining algorithms, Common algorithms include Classification and Regression Tree, ID3, C4.5, Random Forest, etc. At present, decision tree algorithm has been widely used in various fields. For example, financial industry can use decision tree to make loan risk assessment, insurance industry can use decision tree to make insurance species promotion prediction, and medical industry can use decision tree to generate auxiliary diagnosis and disposal model.

In order to improve the computational efficiency of the algorithm, literature [3] proposed to change C4.5 by using the property of equivalent infinitesimal, Entropy, information gain and information gain rate of the algorithm [4]. Although the number of times of calling the logarithm operation function is reduced in the calculation process, the error value increases due to the neglect of constant value calculation, leading to a decline in the accuracy of classification results. Aiming at the problem of missing attribute values led to the decrease of the classification accuracy, literature put forward in the process of decision tree generation, when the branch a subset of unknown attribute values, return to the leaf node, marked as unknown, and then the pruning of more than a third (the ratio of the unknown nodes and leaf nodes) deleting unknown nodes [5]. Compared to the traditional C4.5 Compared with the C4.5 algorithm, the time complexity of this algorithm can be greatly improved when the attribute loss rate is high, but when the data set loss rate is low or even no loss rate. In addition, this algorithm lacks a reasonable method to set the threshold of attribute failure rate. In the literature, CCP(Cost-Complexity Pruning) method in post-pruning method is adopted to simplify its model, and then evaluation criteria are added in the same stock to supplement the single CCP method, so as to avoid over-coarse subtree construction. Through comparison with ordinary decision tree, it is concluded that the coverage of decision tree is better than other classifiers [6].

Therefore, this paper adjusts the value range of the information gain rate of attribute metrics in the C4.5 algorithm, and improves C4.5 branch and pruning strategy of algorithm. The surface error rate gain value and S value are calculated for the sub-root node of the generated decision tree, so as to judge whether to delete the node and branch of the decision tree. And put it into practice.

3. Decision Tree

3.1 Concept

Decision tree method is a process of classifying data by a series of rules. Specifically, the information gain in information theory is used to find the attribute field with the maximum information in the database, establish a node of the decision tree, and then build branches of the tree according to different values of the attribute field. In each branch subset, the process of building the lower nodes and branches of the tree is repeated. It is a typical classification method, which first processes data and generates readable rules and decision trees by using induction algorithm, and then analyzes new data [7]. Essentially, a decision tree is a process of classifying data by a series of rules.

3.2 Common Decision Tree Algorithms

The typical algorithms of decision tree include ID3, C4.5 and CART, etc. The classification model based on decision tree has the following characteristics:

(1) the decision tree method is simple in structure and easy to understand;

(2) the decision tree model has high efficiency and is suitable for the case of large training set;

(3) the decision tree method usually does not need to accept the knowledge outside the training set data;

(4) decision tree method has high classification accuracy.

Compared with other algorithms, the classification rules generated by C4.5 algorithm are easier to understand and have higher classification accuracy. It shows how attributes are divided and illustrates how to build a decision tree node, C4.5 algorithm for a set of input data, first choose one of the most can distinguish between data set instance attributes as the root node, then according to the calculated further build a child node, until the instance of the subclass or no residual properties meet the expected condition to continue.

Assuming that Q represents the current sample set and the current candidate attribute set is represented by A, then the pseudo-code of C4.5 algorithm is as follows. The algorithm flow of C4.5 is as follows [8].

Algorithm: Generate decision tree generates a decision tree from given training data.

Input: training samples; Attributelist for the set of candidate attributes;

Output: a decision tree

The flow chart of the algorithm as follows：.

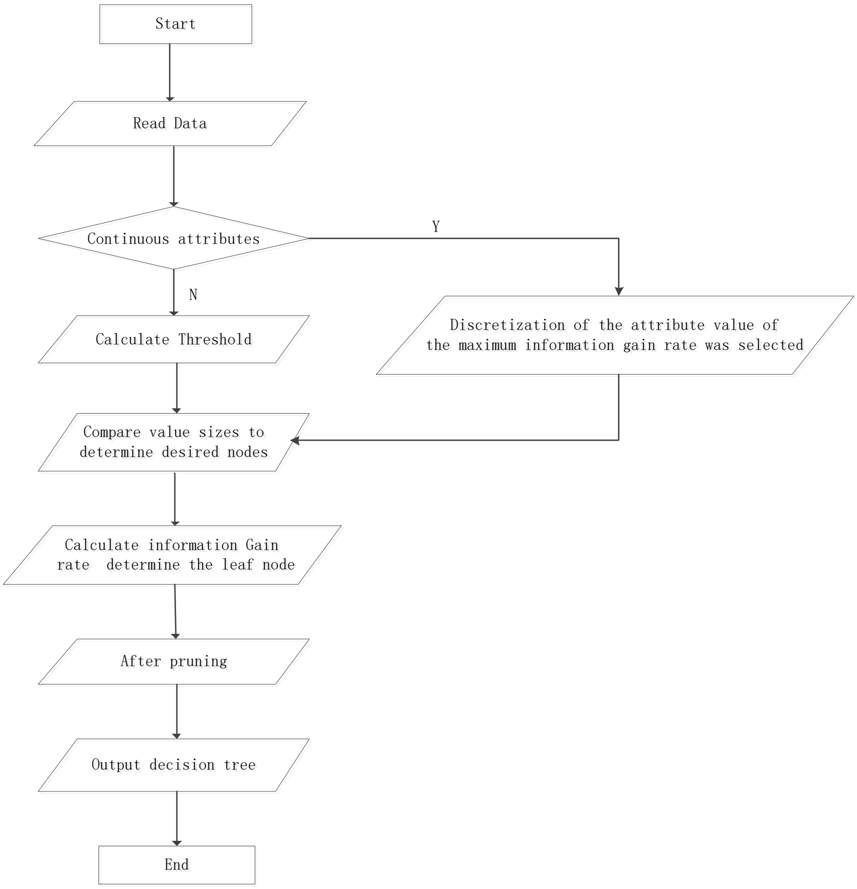
Start
Read Data
Continuous attributes
Y
N
Calculate Threshold
Discretization of the attribute value of the maximum information gain rate was selected
Compare value sizes to determine desired nodes
Calculate information Gain rate determine the leaf node
After pruning
Output decision tree
End

Fig. 1 C4.5 Algorithm process

4. Improved Decision Tree Algorithms

The improvement method is mainly to reduce the information entropy of strong correlation attributes, improve the information entropy of some weak correlation attributes, construct a new decision tree model, and improve the accuracy of decision tree prediction.

4.1 Partition information entropy

Use the attribute V to partition the data in the sample set Q and calculate Entropy of V to Q as Entropy (Q). Attribute V is divided into discrete type and continuity type.

1) V is discrete type. Take N different values, then attribute V divides Q into N subsets {Q1, Q2... , Qm}. After the introduction of the parameter W, The number of samples contained in Qi and Q is | Qi | and | Q |.the information entropy of the property V partition Q is defined as:

$$\text{Entropy}_{\text{v}}\text{(Q)=}\sum_{i=1}^{n}\frac{|Q_i|}{Q}\times Entropy(Q_i)\times\frac{F}{Entropy(Q_i)+1} \tag{1}$$

2) V is continuous. Use the increasing value of attribute V to sort, assume that there are N different values of attribute V in Q, then the ordered value sequence is a1, a2,... , am. Take the mean value of two adjacent values one by one as the segmentation point. The segmentation point divides Q into two subsets, QL and QR. QL is the subset whose value of attribute V is less than the mean value, and QR is the subset whose value of attribute V is greater than the mean value is:

$$\text{Entropy}_{\text{v}}\text{(Q)=}\frac{|Q_L|}{|Q|}Entropy(Q_L)\times\frac{F}{Entropy(Q_L)+1}+\frac{|Q_R|}{|Q|}Entropy(Q_L)\times\frac{F}{Entropy(Q_R)+1} \tag{2}$$

By calculating and comparing the information gain rate of all segmentation points of attribute V, the maximum information gain rate is selected as the information gain rate of attribute V.

4.2 Calculate the information gain rate

Assume that the entropy of the divided sample set Q is G is Entropy (Q), whose value is independent of specific conditional attributes. Then the calculation formula of information gain rate is as follows:

$$Gain(V)=Entropt(Q)-Entropy_V(Q) \tag{3}$$

Through the calculation of the gain rate in the original C4.5 algorithm, the calculation formula of the information gain rate in the new C4.5 algorithm is deduced as follows:

$$GainRatio(V)=\frac{Gain(V)}{SplitE(V)}\times(SplitE(V)+F) \tag{4}$$

and SplitE (V) is the splitting information grouped by attribute V, and the calculation method is the same as C4.5 algorithm [9,10].

In addition, most of the training set attributes in the current medical system are continuous attributes, and the original algorithm will greatly reduce the efficiency of the algorithm.

Therefore, this paper proposes improved C4.5 based on boundary theorem and bayesian classifier algorithm B-C4.5. Improve the segmentation point selection of continuous attributes.

Boundary theorem: no matter how big the data set used for training is, how high the dimension is, or how many classes there are in the data set, the best partition point of continuous attributes is always at the boundary point.

Bayesian classifier: it is the classifier with the lowest probability of classification error among all kinds of classifiers. Its design method is the most basic statistical classification method. The bayesian formula is as follows:

$$P(H_i \mid A) = \frac{P(A \mid H_i)P(H_i)}{\sum_{j=1}^{n} P(A \mid H_j)P(H_j)} \quad (5)$$

and $P(H_i \mid A)$ is the probability of class Hi when the attribute is A, and P (A | Ci) is the probability of A when the class is Hi.

The improvement steps of discretization of continuous attributes are as follows:

(1) The values of each successive attribute of the training set are sorted in ascending order respectively.

(2) Make the minimum attribute value of Ai when the class is H1;

(3) Get the maximum attribute value of Ai when the class is H2;

(4) Calculate the tangent point value based on the maximum and minimum values of each class Hj (j = 1,2);

(5) Calculate the information gain of the tangent point value, and select the maximum value as the candidate segmentation point;

(6) Calculate the tangential value probability, select the largest candidate as the segmentation point;

(7) If the maximum information gain and probability are the same segmentation point, repeat steps (5), step (6), step (7).

The complexity of the decision tree model generated by the algorithm is too large, easy to produce over-fitting phenomenon, the rules of decision tree generation are difficult to understand, and the algorithm efficiency is low, so a variety of pruning methods of decision trees come into being [11].

The common pruning methods are pre-pruning and post-pruning. Pre-pruning refers to stopping the growth of the tree when it reaches a certain height and the node becomes a leaf node or the number of instances reaching a node is less than the set threshold. The problem with pre-pruning is that the nodes become leaf nodes too early, resulting in the possible loss of some properties of the data set. Second pruning, on the other hand, involves cutting out some of the molecules and replacing them with leaves when the decision tree is fully grown. For this reason, the CCP (Cost-Complexity Pruning) method in the post-pruning method is adopted in this paper, which is firstly simplified and then supplemented by adding evaluation criteria to avoid too rough pruning of subtrees [12].

Simplify CCP calculation formula:

$$\alpha = \frac{\mathrm{P}_{\mathrm{j}} - E_{\mathrm{j}}}{C} / (N_{\mathrm{i}} - 1) \quad (6)$$

and P_{j} is the number of error samples of sub-tree root nodes, E_{j} is the sum of error samples of each leaf node, C is the total data amount, and N_{i} is the number of sub-tree leaf nodes.

5. Application of C4.5 algorithm in disease prediction

5.1 Data Preprocessing

The original data used in this paper are electronic medical records of a hospital in Baoding, HeBei province, The experimental system environment is Windows 8.1; Programming tools: PyCharm; Programming language: python. Experimental data came from UCI data set, with a total of 768 records. The data set was divided into 11 groups, among which 10 groups of 600 records were used as the training set. The training set was extracted by random sampling. After the training set is extracted, 168 remaining records in the laboratory data set are taken as the data of the test set.

The experimental data set sample contains 9 attributes, respectively:

(1) number of time pregnant (pregnant times);

ISSN: 2414-1895 DOI: 10.6919/ICJE.201908_5(9).0023

(2) Plasma glucose concentration a 2 hours in an oral glucose tolerance test (The concentration of plasma glucose was 2 hours.

(3) Diastolic blood pressure (mm Hg);

(4) Triceps skin fold thickness (mm) triceps skin fold thickness;

(5) 2---Hour serum insulin (mu U/ml) (2 hours serum insulin);

(6)Body mass index (BMI);

(7) Diabetes pedigree function (glycosuria spectrum);

(8)Today the Age(age);

(9) Class (0or1).

The number of pregnancies between 0 and 1 is marked as Low, while the number between 2 and 5 is marked as Medium, and the number greater than or equal to 6 is marked as High [13]. Attribute BMI = weight(kg)/height(m). According to WHO standards, the grade is: BMI value is less than 18. Let's make it lean. 18.5--24.9 is normal, and 25 or more is overweight. Due to the experimental data sampling scope is too broad, does not represent a country or region of the typical classification basis, therefore on the basis of the experimental data set, attribute information gain rate of BMI by calculation to obtain a node threshold, will be less than the value of the attribute value is marked as Light, is greater than the value of it is Overweight. AGE is treated similarly to BMI, and its values are marked as Young and Elderly respectively. In order to make the generated decision tree concise and convenient for subsequent algorithms, the data set generic name is denoted as NTP, PG, DBP, TSF, HSI, BMI, DPF and AGE.

We get the values of data set attributes to be processed and analyzed in the following table.

Table 1 Data value attribute ID and value

property name	ID	Value
Plasma glucose	PG	continuous
Age	AGE	Continuous(young,elderly)
Diabetes pedigree function	DPF	Continuous
Body mass index	BMI	Continuous
Diastolic blood presure	PG	Continuous
Pregnant Times	PT	Low,Medium,High

5.2 Experimental Results

The information gain and probability values of the tangential points of the attributes obtained by combining the boundary theorem and the bayesian classifier are shown in Table 2.

Table 2 Attribute the information gain value and probability value of each tangent point

property name	point value	gain	probability
DBP	20	0.00076	0.260
	26	0.00089	0.308
	50	0.00085	0.223
	90	0.00087	0.340
PG	80	0.00756	0.160
	90	0.00634	0.201
	200	0.09756	0.380
	190	0.10756	0.121

The tangential point value with the maximum information gain and probability value was selected from Table 2 as the optimal segmentation threshold, and the optimal segmentation threshold was obtained, as shown in Table 3.

Table 3 The optimal threshold of each continuous attribute

ID	Optimal segmentation threshold	unit
BMI	30	Kg/m2
DPF	1.3	
AGE	39	
PG	7.90mmol/l	Mmol/L
DBP	76	mmHg

It can be seen from Table 3 that the optimal segmentation threshold of PG is 7.90 mmol/l), which is 7. 8mmol/l is very close, which reflects that the improvement of discretization of continuous properties is effective. Such as DBP, BMI, DPF and AGE, the optimal cutting threshold values were 76 mmHg and 30kg/m2, 1.3 and 39. In addition, the implementation time between the modified method and the original method is 780 ms and 820 ms respectively, reducing the time by 7.85%. Due to the limited amount of data in the data set, the efficiency is not greatly improved, but it has been significantly improved compared with the original algorithm.

Table4 Accuracy test results

Test Set	Test Number	C4.5 accuracy	B-C4.5 accuracy
T1	50	70.02	76.78
T2	100	70.10	78.90
T3	150	72.70	80.06
T4	200	75.04	82.56
T5	250	82.90	84.30
T6	300	84.56	86.50
T7	350	83.57	90.35
T8	400	86.60	91.23

Table 4 shows C4.5 when the number of instances of test set T_i (I = 1,2,3,4,5,6,7,8) is 50, 100, 150, 200, 250,300, 350,400 respectively. C4.5 algorithm and improved B-C4.5 comparison results of classification accuracy of the algorithm.

We can from the table 4 get that the classification accuracy of B-C4.5 is generally higher than that of the original algorithm under different test sets. As can be seen from the table, when the number of test set instances increases, the accuracy of the improved algorithm presents an upward trend and gradually increases, and the final accuracy reaches 91.23%, which also proves the effectiveness of the B-C4.5 algorithm.

6. Conclusion

In this paper, the boundary theorem is used as the basis to calculate the bayesian probability of nodes. In the process of decision tree generation, the discretization efficiency is improved when the nodes of sub-tree roots are continuous attributes. In pruning after the decision tree, evaluation criteria are added to the original method to ensure that the sub-tree node has less influence on the whole decision tree in the classification process compared with other nodes when the sub-tree node is deleted. By comparison, the performance is significantly better than that of C4.5 algorithm. However, this algorithm also has some shortcomings at present. It does not take into account the influence of noise data and the processing of missing attribute values of data sets in the classification process, which also has some influence on the generated decision tree. Therefore, how to deal with noise data and missing value attribute is the next research direction.

References

[1] J.Sanz, J.Fernandez, H.Bustince, C.Gradin,”A decision tree based approach with sampling techniques to predict the survival status of poly-trauma patients,”IJCIS,vol.10,pp.440–455,2017.

[2] BERGMAN R N, KALABA R E, SPINGARN K. Optimizing Inputs for Diagnosis of Diabetes I. Fitting a Minimal Model to Data[J].Journal of Optimization Theory and Applications.2011, 20(9):317-320.

[3] SONETHUNG D,SRIPANIDKULHALI. Improving type 2 diabetes mellitus risk prediction[C]. International Joint Conference Computer Science and Software Engineering(JCSSE), 2016.

[4] DEWAN MD,FARID. Improve the quality of supervised discretization of continuous valued attributes in Data Mining[C].Proceeding of 14th International Conference on Computer and Information Technology(ICCIT 2011), 2011.

[5] S.Hamali, R.P.NSuci, A.F.Utami, Hanisman and FArga,"Using analytic hierarchy process and Decision Tree for a production decision making,"2016 International Conference on Information Management and Technology (ICIMTech),Bandung,Indonesia,2016, pp.329-332.

[6] Z.Jiang,S.Shekhar,X.Zhou,J.Knight and J. Corcoran,"Focal-Test-Based Spatial Decision Tree Learning," in IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 6, pp. 1547-1559, June 1 2015.

[7] Y. Qian, H. Xu, J. Liang,B. Liu and J. Wang, "Fusing Monotonic Decision Trees," in IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 10, pp. 2717-2728, Oct.1 2015.

[8] P.J.Tan and D.L.Dowe,”Decision Forests with Oblique Decision Trees,”in MICAI,2006,pp.593–603.

[9] ZHANG Heng, JIN Xin, QIN Xiaoqian. Application of constrained KNN regression in noise data [J] . Computer Engineering, 2015,41(12):275-279

[10] J.Shotton,T.Sharp, P.KohliS.Nowozin”Decision Jungles: Compact and Rich Models for Classifification,” in NIPS,2013,pp.234–242.

[11] G. W. Corder and D. I. Foreman, Nonparametric statistics: a step-by step approach.New York, Wiley, 2014.

[12] H.Zhao and X.Li, “A cost sensitive decision tree algorithm based on weighted class distribution with batch deleting attribute mechanism,”Inform.Sciences,vol.378,pp.303-316,2017.

[13] LI Qiyi, WANG Lei, SHI Lei. Estimation of working hours based on decision tree and model tree [J]. Computer integrated manufacturing system, 2017,(12),18-21.