International Journal of Database Theory and Application
Vol.10, No.1 (2017), pp.119-126
http://dx.doi.org/10.14257/ijdta.2017.10.1.11

A Novel Five-Step Data Mining Algorithm

Wang Yiwen

School of Economics, Heilongjiang University of Science and Technology, Harbin, Heilongjiang, 150081, China
E-mail: yscls2013@163.com

Abstract

Based on the traditional data mining algorithm, a novel data mining algorithm is proposed. This algorithm consists of 5 steps: the first step, set the tree set; the second step, set the window third, subtree contribution; decision tree construction; the fourth step test, positive and negative examples set; the fifth step, expand the achievements window. The experimental study on open source data sets. The results showed that the five step proposed data mining method, not only can build a more concise decision tree, data mining and the accuracy is also higher than the traditional decision tree method.

Keywords: *Data mining, positive tree set, inverse tree set, five step method*

1. Introduction

With the rapid development of hardware and software technology, massive data storage has become possible. However, how to store and manage the huge data information become a new task for people [1]. The emergence of data warehouse technology has successfully solved the problem of large data storage and management, and supporting data mining technology for people to extract the most valuable information from the massive data [2].

Data mining is also called data exploration, which is committed to extracting valuable information from existing data information and even forming knowledge. The foundation stone to support the data mining technology is the intelligent algorithm with search, classification, recognition and other functions .At present, data mining techniques that have emerged include: Mining technology based on machine learning, mining technology based on clustering analysis, and mining technology based on association rules and so on [3-5].

Among all kinds of mining techniques based on machine learning, decision tree mining method has been widely used. Banaee constructs a classifier based on decision tree method, which can record the maximum attribute gain of the data on each non leaf node in the decision tree, and as discriminant basis for data classification [6]. Porro selected the Gini coefficient as a new discriminant parameters in the attributes of decision tree test, and through resampling techniques to reduce the error in the construction process of the decision tree, and use the method of minimum cost to reduce the complexity of decision tree construction [7]. Aiming at the problem of data mining in the process of teaching in higher vocational colleges, Boteva set up a decision tree mining method based on goal driven, and achieved satisfactory results [8]. Mayr compared the difference between the ID3 decision tree and the C4.5 decision tree method, which confirmed the advantages of the C4.5 decision tree method in dealing with continuous attributes, incremental learning, dynamic adjustment and so on [9]. Vilares takes the financial system data mining as the research breakthrough point to construct the decision tree - united neural network mining algorithm, effectively promoted the financial fraud investigation rate [10]. Palacios studied the mining of large data sets of healthcare, designed specifically for the mining of tables and histogram data structure and constructed a decision tree based on exchange

ISSN: 2005-4270 IJDTA

policy, pre ordering and breadth first criterion [11]. Elyasigomari takes the cancer data as mining object to construct a high speed scalable decision tree mining method, which also has the supervisory function [12].

Based on the method of ID3 decision tree, this paper constructs an improved decision tree mining method, in order to further improve the process of decision tree construction and improve the efficiency of decision tree method.

2. Classic ID3 Decision Tree Mining Method

Among the existing decision tree mining methods, the ID3 method is one of the most classical methods. The idea of ID3 decision tree mining is that, using the information gain to construct attribute selection basis at all levels of the decision tree nodes, the highest information gain attributes will was eventually identified as the criterion of the nodes, In this way, the minimization of the information of training samples can be achieved, and the construction speed of the decision tree and the structure of simplified the decision tree can be promoted.

$\tilde{M}$ is used to represent a domain, the probability distribution of any one of the arbitrary division $\{M_1, M_2, \cdots, M_n\}$ corresponding to : $p_i = P(M_i)$, Then the information entropy of the source M can be calculated as follows:

$$G(M) = -\sum_{i=1}^{n} p_i \log p_i \tag{1}$$

Let W say another arbitrary division on the domain $\tilde{M}$, W can be expressed as: $W = \begin{Bmatrix} W_1 & W_2 & \cdots & W_n \\ k_1 & k_2 & \cdots & k_n \end{Bmatrix}$, then there are $P(W_j) = k_j$ and $\sum_{j}^{n} k_j = 1$, so that the information source M can be defined on the W condition, as shown in the formula (2):

$$G(W \mid M) = \sum_{i=1}^{n} P(M_i) G(W \mid M_i) \tag{2}$$

For data samples to perform mining processing, we use T and F to represent the positive and the negative example, a decision tree can make the correct classification information entropy as follows:

$$G(T, F) = -\frac{T}{T+F} \log \frac{T}{T+F} - \frac{F}{T+F} \log \frac{F}{T+F} \tag{3}$$

If an attribute A can be used as the root of the decision tree, then the desired entropy is calculated as follows:

$$E(A) = \sum_{i}^{n} \frac{t_i + f_i}{T+F} G(t_i, f_i) \tag{4}$$

At this point, we can calculate the information gain, which is the root of attribute A, and its calculation is as follows:

$$Z(A) = G(T, F) - E(A) \tag{5}$$

For ID3 algorithm, it is to find the maximum information gain in each attribute set as the root node.

The execution flow of the core part of the ID3 decision tree mining algorithm is shown in Figure 1.

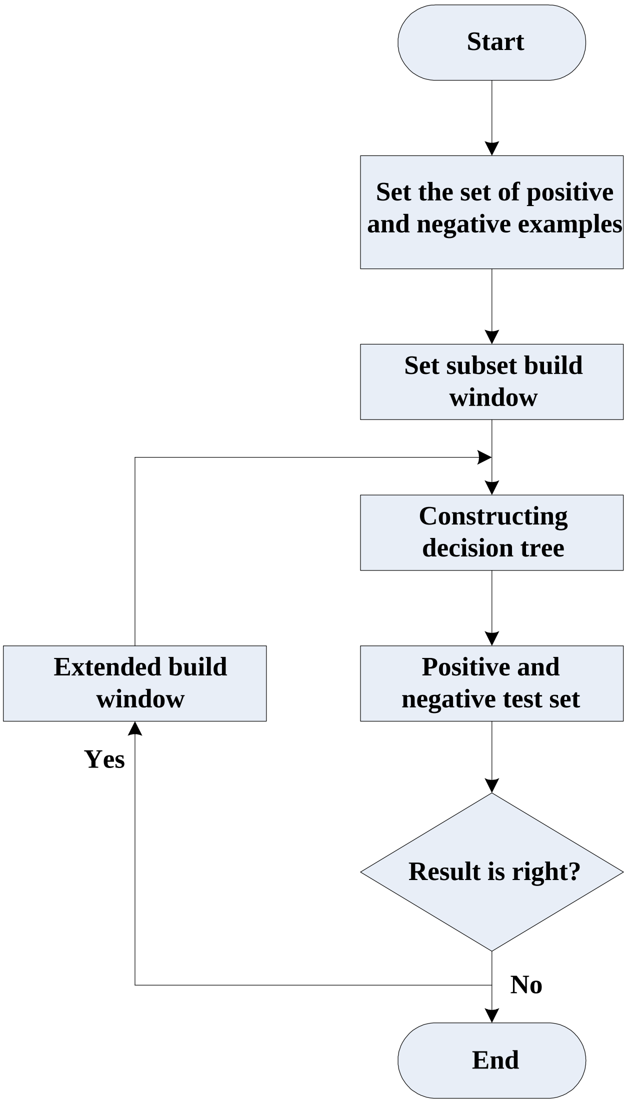
Start
Set the set of positive and negative examples
Set subset build window
Constructing decision tree
Extended build window
Positive and negative test set
Yes
Result is right?
No
End

Figure 1. Algorithm Flow

From the process shown in Figure 1 can be seen, ID3 decision tree mining algorithm is very simple, and it can be roughly divided into the following steps:

The first step, according to the data from the sample data set to construct a set of positive and negative set T and F;

The second step, set the appropriate number of attribute subsets, each of these attributes will be one by one included in the contribution of the window, and calculate the attribute gain;

The third step, with the results of the second step calculation of the attribute gain, the decision tree is built according to the principle of the maximum gain of the attribute as the root node;

The fourth step, according to the third step of constructing decision tree on the sample set of positive and negative examples tested, to observe whether there is an error, if there are errors, then to create the extension decision tree window to rebuild the decision tree, if there is no error, it is proved that the decision tree constructed is reasonable and can be used for the subsequent mining process, and the algorithm is over.

3. Improved Decision Tree Mining Method

ID3 decision tree mining algorithm is not only simple in principle, but also has many outstanding advantages: its search space is complete , with fewer tests , the speed of classification mining is fast , less decision tree nodes, suitable for noise data and discrete data mining.

The biggest limitation of ID3 decision tree algorithm is that its work principle is the attribute gain that calculated based on information entropy theory. The best attribute of this method is a multi valued attribute, but it is not necessarily optimal. This also leads to the result that the mining results tend to be only locally optimal and can not achieve the global optimum.

To this end, this paper improves the classical ID3 decision tree algorithm, and tries to construct a decision tree which can achieve the global optimum.

The calculation of information entropy is the key of ID3 decision tree algorithm to achieve the mining process. Therefore, this paper focused on improving the calculation of entropy, and then on this basis to redesign the implementation process. In the classical ID3 decision tree method, the attribute A would appear multi values in the process of calculation. This paper selects the number of these values as weight coefficient to blend into the entropy calculation of the attribute A.

Assuming that the attribute A has m attributes, the probability of these attributes: are respectively use p_1 、 p_2 、 $\cdots$ 、 p_m to represent. The attribute A as the node corresponding to the m sub nodes can be represented as $\{\theta_1, \theta_2, \cdots, \theta_m\}$. The information entropy information entropy correspond to these attribute values with $G(\theta_1)$ 、 $G(\theta_2)$ 、 $\cdots$、 $G(\theta_m)$. In the improved method in this paper, the final design of the formula to calculate the attribute A is as follows:

$$\overline{G}(A) = m\sum_{i=1}^{m} p_i \times G(\theta_i) \tag{6}$$

Now, we give the implementation steps of the improved decision tree method in this paper:

The first step, for any one of the attributes of A_i, assuming that it has m_i attribute values and you can use $\{\theta_1, \theta_2, \cdots, \theta_{m_i}\}$ to represent the values. The corresponding properties of these values are p_1、 p_2、 $\cdots$、 p_{m_i} , in which the information entropy of each attribute value is calculated as the formula (7):

$$G(\theta_i) = \sum \frac{2t_i f_i}{t_i + f_i} \tag{7}$$

The second step, with the formula (6) to calculate the entropy of attribute A_i;

The third step, according to the first step and the second step method, continue to calculate the corresponding entropy values of attributes of A_{i+1}、 A_{i+2} $\cdots$, and select the attribute corresponding to the minimum entropy as the node from all the entropy;

The fourth step, according to the first to the third step method, continuing each successor node;

The fifth step, when a new round of calculation results to determine the various nodes are leaf nodes, the decision tree construction is completed; otherwise continue to perform the previous step.

4. Experimental Results and Analysis

4.1. Evaluation Index of Performance of Decision Tree

In decision tree learning algorithm, the complexity of decision tree and the classification accuracy are the two most important factors to be considered. The following is the performance evaluation criteria of the decision tree:

/reflected by the mining accuracy.

4.2. Data Set in Experiments

In order to verify the effectiveness of the improved decision tree mining method in this paper, we chose the UCI test data set of University of California to carry out the data mining experiment. There are various types of data in UCI data set, we have selected the 6 kinds of data sets.

(1) Wine data set

Wine data sets can be used to determine the chemical composition of liquor data. The data mainly come from the liquor produced in Italy, these chemical components involved in alcohol, acid, flavonoids and so on. The entire Wine data set contains 178 sample data, 13 integer attributes and 3 categories.

(2) Heart Spect data set

The Heart Spect data set is based on the CT image to determine whether the patient's heart is a normal data set, which is based on the results of the proton emission CT scan. The entire Heart Spect data set contains a total of 267 sample data, 22 attribute characteristics, each attribute is only "-1" or "1" those two kinds of possible expression, 2 categories— normal and abnormal.

(3) Scale Balance data set

Balance scale data set is the data set that according to the psychology experiment to determine whether the balance is balanced. These data according to the weight of the two trays, the distance, the psychology of testers to judge the balance will tilt. The balance scale data set containing a total of 625 sample data, each data sample contains four attributes, and these attributes need to be valued in "1, 2, 3, 4, 5", the three categories are: left tilt, right tilt, balance.

(4) Vehicle Silhouettes data set

Vehicle Silhouettes data set is based on the image of the two-dimensional feature information to determine the formation of the car category data collection. The entire Silhouettes Vehicle data set contains 846 sample data, each sample data contains 18 attributes which need to be valued in the integer space, 4 categories.

(5) Hill Valley data set

Hill Valley data set is used to determine the concave and convex shape of the data set after the connection curve of multi data points. It first connects the 100 data points into a curve, and then determine the concave and convex shape of the curve. The entire Hill Valley data set contains 1212 sample data, each sample data contains 100 attributes, and these attributes are valued in the real space, 2 categories.

(6) Yeast data set

Yeast data set is the data set that uses protein components to predict cell location. The entire Yeast data set contains 1484 sample data, each sample data contains 8 attributes, and these attributes are valued in the real space, 10 categories.

In order to facilitate a clear comparison of the characteristics of the 6 data sets, we will further summarize them as a form of Table 1.

Table 1. Original Data Configuration

Data set	Sample data	Attribute	Classification
Wine	178	13	3
Spect Heart	267	22	2
Balance Scale	625	4	3
Vehicle Silhouettes	846	18	4
Hill Valley	1212	100	2
Yeast	1484	8	10

4.3. Experimental Comparison between ID3 Method and Improved Method

In the 6 types of UCI data set shown in the table, we use the ID3 method and the improved decision tree method in this paper to conduct the mining classification experiment. The comparison results between the number of nodes and the accuracy of the two kinds of methods in decision tree construction are shown in Table 2.

Table 2. Value of Results in First Group

	Number of nodes		Mining accuracy	
	ID3 method	Improved method	ID3 method	Improved method
Wine	340	252	0.81	0.93
Spect Heart	273	221	0.91	0.97
Balance Scale	166	142	0.76	0.85
Vehicle Silhouettes	132	118	0.77	0.88
Hill Valley	96	67	0.72	0.84
Yeast	55	32	0.69	0.82

In order to compare the performance of the two methods in a more intuitive way, we draw the results in Table 2 to Figure 2 and 3 respectively, as shown below:

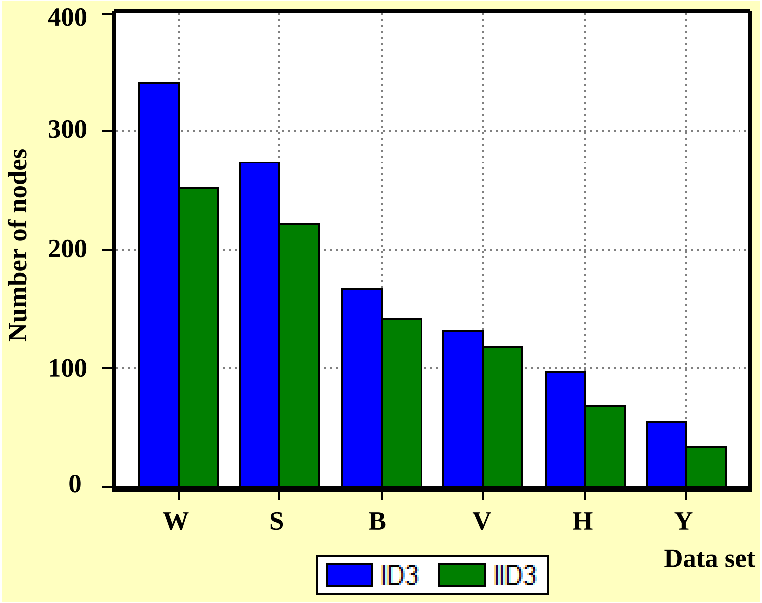

Figure 2. First Experimental Result

In Figure 2, the abscissa axis represents the data sets, for example, W represents the Wine data set; the vertical axis represents the number of nodes in the decision tree constructed by the two methods; IID3 represents the improvement method. As can be seen from the graph, compared to the classical ID3 algorithm, the number of nodes in the decision tree is less. This shows that the decision tree constructed by the improved method is more streamlined.

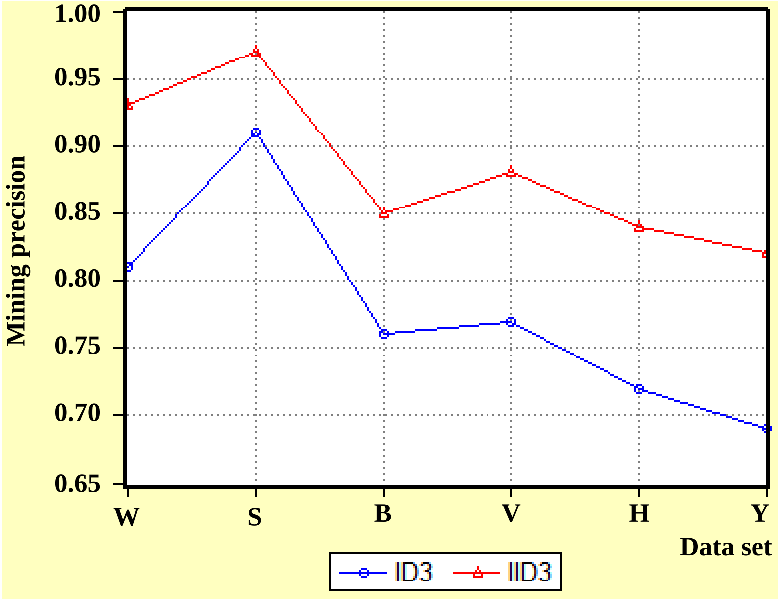

Figure 3. Second Experimental Result

In Figure 3, the abscissa axis represents the data sets, for example, W represents the Wine data set; the vertical axis represents the mining accuracy of the two methods; IID3 represents the improvement method. As can be seen from the graph, compared to the classical ID3 algorithm, there is an obvious improvement in the accuracy of the mining.

5. Conclusion

Aiming at the problem of data mining, the theoretical research and experimental research are carried out based on the classical ID3 decision tree mining method. Firstly, it analyzes the nodes selection process and data mining process of ID3 decision tree mining method; then analyzes the problems exist in ID3 decision tree mining method, and improved the entropy calculation process to obtain a globally optimal mining result. Finally, Aimed at the 6 kinds of data sets of UCI data set, the mining experiment was carried out. The experimental result shows that: the improved mining method is obviously better than the ID3 decision tree mining method in the degree of simplicity and the mining accuracy of the decision tree construction.

References

[1] Settouti Nadera, Aourag Hafid. A comparative study of the physical and mechanical properties of hydrogen using data mining research techniques[J]. JOM, 67(9): 2145-2153. (2015)

[2] Bessas Izaquiel L, C.Padua Flavio L, De Assis Guiherme T. Automatic and online setting of similarity thresholds in content-based visual information retrieval problems[J]. Eirasip Journal on Advances in Signal Processing, 1: 112-119. (2016)

[3] Rahimi Razieh, Shakery Azadeh, King Irwin. Extracting translations from comparable corpora for Cross-Language information retrieval using the language modeling framework[J]. Information Processing and Management, 52(2): 299-318. (2016)

[4] Palacios Ana, Martinez Alvaro, Sanchez Luciano. Sequential pattern mining applied to aeroengine condition monitoring with uncertain health data[J]. Engineering Applications of Artifical Intelligence, 44: 10-24. (2015)

[5] Abdul-Rauf Sadaf, Schwenk Holger, Lambert Patrik, Nawaz Mohammad. Empirical use of information retrieval to build synthetic data for SMT domain adaptation[J]. IEEE/ACM Transactions on Speech and Language Processing, 24(4): 745-754. (2016)

[6] Banaee Hadi, Loutfi Amy. Data driven rule mining and representation of temporal patterns in physiological sensor data[J]. IEEE Journal of Biomedical and Health Informatics, 19(5): 1557-1566. (2015)

[7] Porro Munoz Diana, Olivetti Emanuele, Sharmin Nusrat. Tractome: a visual data mining tool for brain connectivity analysis[J]. Data Mining and Knowledge Discovery, 29(5): 1248-1279. (2015)

[8] Boteva Vera, Gholipour Demian, Sokolov Artern, Riezler Stefan. A full-text learning to rank dataset for medical information retrieval[C]. Advamces in Information Retrieval-38th European Conference on IR Research ECIR Proceedings, 9626: 716-722. (2016)

[9] Mayr Philipp, Frommholz Ingo, Cabanac Guillaume. Editorial for the 3rd bibliometric enhanced information retrieval workshop at ECIR 2016[C]. Proceedings of the 3rd Workshop on Bibliometric-Enhanced Information Retrieval, co-located with the 38th European Conference on Information Retrieval, 1567: 1-4. (2016)

[10] Vilares Jesus, Vilares Manuel, Alonso Migue A, Oakes Michael P. On the feasibility of character n-grams pseudo-translation for Cross-Language information retrieval tasks[J]. Computer Speech and Language, 36, 136-164. (2016)

[11] Palacios Ana, Martinez Alvaro, Sanchez Luciano. Sequential pattern mining applied to aeroengine condition monitoring with uncertain health data[J]. Engineering Applications of Artifical Intelligence, 44: 10-24. (2015)

[12] Elyasigomari V, Mirjafari M.S, Screen H.R.C. Cancer classification using a novel gene selection approach by means of shuffling based on data clustering with optimization[J]. Applied Soft Computing, 35: 43-51. (2015).