Multi-Decision Tree Classification in Master Data Management System

[1]V. Geetha, [2]S. Jessica Saritha
[1,2]Dept. of CSE, J.N.T.U.A.C.E.P, Andhra Pradesh, India

Abstract
To be more conceptual, the basic idea is to provide the simplified decision tree ID3 algorithm. It overcomes the existing bias of ID3 (Iterative Dichotomiser 3) algorithm A multi-decision tree classifier were constituted by combining AdaBoost and improved ID3 algorithm. The master data management system is an application we use, it greatly reduces the manual labor by grouping the redundant data together and also it saves the consumption of human and material resources. The accuracy is better compared to the original decision tree.

Keywords
ADABOOST algorithm, ID3 algorithm, Master data and Master Data Management.

I. Introduction

The data contains valuable and precious information that could be shared among multiple systems also represents the core data is called master data. To manage these kinds of core data has always been important now adays. Such as, knowing who your customers are, what products and services you offer, and what the arrangements or accounts you have with your customers and suppliers is fundamental to the operation of most organizations. Master data is data shared across computer systems in the enterprise. It is core business objects shared by multiple applications across an enterprise and you may have any kind of organizers like a Bank, Retail industry. Master data managment contains a set of tools, processes and technologies to produce and maintain a single clean copy of master data. An Application for creating and maintaining an authoritative view of master data including policies and procedures for access, update, modification, viewing between systems across the enterprise. MDM Systems that focus exclusively on managing information about customers are often called Customer Data Integration systems, and managing the descriptions of products are called Product Information Management (PIM) systems. MDM Systems that enable multiple domains of master data, and that support multiple implementation styles and methods of use, are sometimes also called Multi-Form MDM Systems.

II. Master Data Management System

Establishing a complete view of the data for distributed systems is the main task of master data. There will be a lot of redundant data in the system just because the master data management is to centralized management. For example, when a master data management system is centralized managing for basic information of the patient of all areas of a hospital information system, basic information in centralized management of the patient, there is bound to have a large number of redundant basic information of patients.

Because a patient may have a number of hospital treatment experience in different hospitals and these hospitals will store the patient's basic information in the master data management system which would have resulted in data redundancy. In the actual production environment, data quality has been a problem that troubled people a lot, the quality of data is distinctive in different systems, so we cannot check whether there was duplication between the data by direct, simple character string and we do not want to invest too much human resource and money in this area. We hope to introduce a way to find the data duplication between the data intelligently and accurately, so that create a complete view of master data and make master management more intelligent.

III. Simplified ID3

The decision tree ID3 is used as an algorithm of base classifier in AdaBoost classifier. Decision tree ID3 algorithm is needed to improve, such that the training time is reduced and increases the classification accuracy. Information gain is calculated in the current node of all properties on the list of property, therefore, logarithmic operations involved in the algorithm. It is necessary to simplify logarithmic operations. Through Maclaurin formula we can conclude the following conclusions.

$$f(x) \approx f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \cdots + \frac{f''(0)}{n!}x^n$$

When f (x) = ln(1+ x) , and the value of x is very small, the answer is: ln(1+ x) ≈ x Through the decision tree ID3 algorithm we can get:

$$\text{Gain}(A) = \text{Info}(X) - \text{InfoA}(X) \quad (1)$$

$$\text{Info}(X) = -\sum_{i=1}^{m} p_i \log 2(p_i) \quad (2)$$

$$\text{InfoA}(X) = \sum_{j=1}^{V} \frac{|X_j|}{|X|} \times \text{Info}(X_j) \quad (3)$$

Substitute formula (2), (3) into (1):

$$\text{Gain}(A) = -\sum_{i=1}^{m} p_i \log_2(p_i) - \sum_{j=1}^{V} \frac{|X_j|}{|X|} \times \text{Info}(X_j) \quad (4)$$

Because the total amount of information achieved by all the attributes of any node in the decision tree must be the same, so after the elimination of the first item we can get:

$$\text{Gain}(A)' = -\sum_{j=1}^{V} \frac{|X_j|}{|X|} \times \text{Info}(Xj) = -\text{InfoA}(X) \quad (5)$$

And because the system will classify data into two groups, So m = 2 in equation (2), and we take F ,T to representing the amount of these two data which is input by the current node,

$$F+T=|X|$$

Take f_j, t_j to respectively representing amount of these two data, which belong to the current value of node attribute A and the same value of number j attribute value, so:

$$\text{Info}(X_j) = -\left(\frac{f_j}{f_j+t_j}\log_2\frac{f_j}{f_j+t_j} + \frac{t_j}{f_j+t_j}\log_2\frac{t_j}{f_j+t_j}\right) \quad (6)$$

$$\text{Gain}(A)' = \sum_{j=1}^{V} \frac{f_{j+}t_j}{F+T} \times \text{Info}(X_j) \quad (7)$$

Substitute formula (6) into (7):

$$\text{Gain(A)'}=\sum_{j=1}^{V}\frac{1}{F+T\ln 2}(f_j\ln\frac{f_j}{f_j+t_j}+t_j\ln\frac{t_j}{f_j+t_j}) \quad (8)$$

Because the value of different attributes in the same node same in (F+T)ln 2 of formula (8), so after the elimination of the factor we can get:

$$\text{Gain(A)"}=\sum_{j=1}^{V}(f_j\ln\frac{f_j}{f_j+t_j}+t_j\ln\frac{t_j}{f_j+t_j}) \quad (9)$$

According to the conclusion above, we can get:

$$\ln\frac{f_j}{f_j+t_j}=ln(1-\frac{t_j}{f_j+t_j})\approx-\frac{t_j}{f_j+t_j} \quad (10)$$

$$\ln\frac{f_j}{f_j+t_j}=ln(1-\frac{f_j}{f_j+t_j})\approx-\frac{f_j}{f_j+t_j} \quad (11)$$

Substitute formula (10), (11) into (9):

$$\text{Gain(A)"}\approx-\sum_{j=1}^{V}\frac{2f_jt_j}{f_j+t_j} \quad (12)$$

A. Overcome the Bias in Attribute

we need to improved the previous simplified algorithm ID3 and overcome the bias.

To overcome this bias, we introduceW :

Gain(A) = Info(X)-WInfoA(X)

in order to simplify the complexity of calculation, we make W equal to the number of attribute values. Also, because of Gain(A)" is simplified from formula (5), so we can conclude the expression of Gain(A)"' after the introduction of W :

$$\text{Gain(A)"'=WGain(A)"}\approx-W\sum_{j=1}^{V}\frac{2f_jt_j}{f_j+t_j}$$

We can see that after the introduction of W , the value of information gain of the node of attribute A is dependent on the number of the property value, so take the number of attribute values into account, thus overcome the bias of the original decision tree algorithm ID3.

IV. Judgement Module in the Master Data Management System

Judging the redundant data is to eliminate the investing too much money and human resource in this area.Judgement module of redundant data consistes of three parts:data standardization, data similarity evaluation and data classification. Standardizing the input data of the system and utilizes the format of the data. Match with the same data and calculate the similarity between two data and score the similarity of each of each attribute in these two records. Finally, using classifier to classifying data according to score gained by the similarity degree of each attribute, thereby to determine whether there is duplication between the data in this system. If the two records are repeated through determination of similarity degree of these two records, then assigned a same identity for two records, so as to provide a complete view of the data. In this way, grouped repeated data together automatically and intelligently and avoid a lot of manual labor.

V. Multi- Decision Tree Classifier

Applying of ADABOOST algorithm in Master data management system to constitute the Multi decision tree classification. the system will use the ADABOOST Classifier to classify the data which is based on decision tree and determine the repeatability of the two data.

Structure of ADABOOST:

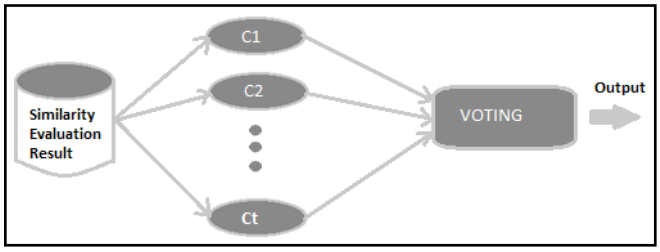

Fig. 1: The Structure of ADABOOST

The Boosting method introduced to improve the performance of weak classifiers.Combines them, to finally output a strong classifier.It is an iterative algorithm.The system designed an ADABOOST classifier which is based on multidecision tree to effectively improve the classification accuracy of using pure decision tree classifier. From figure 1, c1,c2….ct represents a single base classifier, decision tree classifier selected as the base classifier. It takes t iterations to take t no of base classifiers. Each base classifiers has its appropriate weight. By voting machine, the results of t base classifiers can be summarized.

A. Training Process of Multi Decision Tree

The training process of multi decision tree is described as below:

INPUT: $\{x_i, y_i \mid i=1,2,...N\}, x_i \varepsilon X, y_i \varepsilon Y\{-1,+1\}$,iterations T

Initialize: $W_i(i)=\frac{1}{N}, i=1,...N$

For t=1 to 10 do

Rank the training set of X from large to small, 80% of data of X taken as the training set X_t

Through the training of X_t will calculate the base classifier C_t.

In addition to filtering sample Xt used for training of X,Calculate the error rate of the base classifier by the below formula:

$$\varepsilon_i=\frac{1}{N}[\sum_{i=1}^{N}W_t(i)I(h_i(x_i)\neq y_i)]$$

If $\varepsilon_t > 0.5$ then

$W_1(i)=\frac{1}{N}, i=1,...N,$ and then, back to step (4)

end if

Set $\alpha_t=\frac{1}{2}\ln\frac{1-\varepsilon_t}{\varepsilon_t}$

Update sample weights:

$$W_{t+1}(i)=\frac{W_t(i)}{Z_t}\times\begin{cases}e^{-\alpha_t}, h_t(x_i)=y_i\\ e^{\alpha_t}, h_t(x_i)\neq y_i\end{cases}, i=1,...N$$

End for

OUTPUT: $H(x)=sign(\sum_{t=1}^{T}\alpha_t h_t(x))$

VI. Experimental Results

Similarity degrees are evaluated for the 10000 patients as a sample data which are randomly selected, results from computer taken as a data results and divide them into 10 groups to 10 people. First of all we get the accurate rate of ten base classifiers.

Table 1:

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
79.37%	86.51%	89.00%	89.15%	88.45%	91.88%	91.00%	90.13%	87.61%	89.87%

We have taken 10 ID3 classifiers as a base classifier of ADABOOST, the accurate rate of classification was 96.15%. the classification rate has greatly improved after combining the improved Decision tree ID3 with ADABOOST.

The training duration of the ADABOOST classifier of S-ID3 based ADABOOST and the decision tree ID3 algorithm has been tested.

Table 2:

Algorithm	Training duration in min
Decision Tree ID3 based ADABOOST	50
S-ID3 based ADABOOST	35

The duration in finding the redundancy data is greatly improved while compared to the manual processing in our application master data management system.

Table 3:

Function of reducing data	Duration(min/10000 records)
S- ID3 based ADABOOST	5
Manual processing	240-300

Through the test results above we can easily state that the time taken to find the redundant data and its elimination takes less duration while compared to the manual processing.

VII. Conclusion

We can say that the algorithm ADABOOST classifier which is based on improved ID3 can accurate to find the redundancy data in master data management system.It also decreases unnecessary manual processing and also its training duration has been reduced greatly of ensuring the classify accurateness. The improvement is feasible and effective.

References

[1] Nock R, Nielsen F.,"A real generalization of discrete AdaBoost", [J]. Artificial Intellegence, 2007, (171), pp. 25-41.

[2] David Loshin,"Master Data Management", [J]. Artificial Intellegence, 2008, (4), pp. 21–97.

[3] N. Hatami, R. Ebrahimpour,"Combining multiple classifiers: diversity with boosting and combining by stacking", [J]. International Journal of Computer Science and Network Security, 2007, 7, pp. 127–131.

[4] M. Kudo, S. Shirai, H. Tenmoto,"A combination of sample subsets and feature subsets in one-against-other classifier", [J]. Multiple Classifier Systems, 2007, (7), pp. 241–350.

[5] Jiawei Han, Micheline Kamber,"Data mining: concepts and techniques", [J]. Multiple Classifier Systems, 2006, (3), pp. 58-79.

[6] Guishu Ji, Peiling Chen, Hang Song,"Study the survey into the decision tree classification algorithms rule", [J]. Science Mosaic, 2007, (1), pp. 9-12. (In Chinese).

[7] E. Bauer, R. Kohavi,"An empirical comparison of voting classification algorithms: Bagging, boosting, and variants", Machine Learning 36 (1999) 105–139.

[8] C. Domingo, O. Watanabe,"MadaBoost: A modification of AdaBoost", In: Proc. of the 13th International Conference on Computational Learning Theory, 2000.

V. Geetha, Department of CSE, JNTUA College of Engineering, Puliendula, Andhra Pradesh, India.

S. Jessica Saritha, M.Tech., Ph.D, JNTUA College of Engineering, Puliendula, Andhra Pradesh, India.