



RESEARCH ARTICLE

Improved Data Reduction Technique in Data Mining

Pritesh Vora¹, Bhavesh Oza²

¹Information Technology, Gujarat Technological University, L.D. College of Engineering, Ahmedabad, India

²Computer Engineering Department, Gujarat Technological University, L.D. College of Engineering, Ahmedabad, India

¹ pritesh2212@gmail.com; ² bhavesh_oza_2001@yahoo.com

Abstract— *In Data Mining, Data reduction is important issue now a day. Due to huge size of data but maximum of them is irrelevant to objective or some of the data is redundant, which leads to more processing power consumptions and wrong result generation. Data Reduction implies reducing the data but without compromising integrity of it. Decision Tree, attribute subset selections, clustering, data cube aggregation is different techniques basically used for data reduction. Decision tree is a highly effective structure which is gives the possible outcome. In a decision tree in which each branch node represent a choice between alternatives and each node represent the decision or classification. Here we see the generalize algorithm and apply the decision tree technique for reliable outcomes.*

Key Terms: - *Data mining; Decision tree; Data reduction*

I. INTRODUCTION

Today, the development of the computer technology and the degree of the informationization is getting higher and higher, so the people know that the data are needed by them is mass data on the present world. Data mining is the process of extracting important information and knowledge from the large database (mass data)[1]. In these data, information and knowledge are implicit, which people do not know in advance, but potentially useful. At present, the decision tree has important data mining method. Decision tree is commonly used in decision analysis in data mining and machine learning to create knowledge structures that guide the decision making process. Accessing a large amount of data in database which is time consuming process and maintaining large amount of data, is very difficult. In database there are many irrelevant data, noisy data and also duplicate data. Now pre-processing on all this data increase the quality or make the data more feasible to operate. In a database there are many data duplication, irrelevant data and noisy data so to remove them data reduction techniques must be applied.

II. DECISION TREE AND ID3

Decision tree provide the highly effective structure which can give the idea about possible outcomes. In decision tree is a tree in which each branch node represent a choice between a number of alternatives, and each leaf node represents a classification or decision. Every decision tree begins with what is termed a root node, considered to be the parent of every other node. Each node in the tree evaluates an attribute in the data and determines which path it should follow [1]. Typically, the decision test is based on comparing a value against some constant. Classification using a decision tree is performed by routing from the root node until arriving at a leaf node. In more generalize definition of the decision tree written in stepwise form:

- A flow chart like tree structure
- Internal node denotes a test on an attribute
- Branch represents an outcome of the test
- Leaf nodes represent class labels or class distribution

The ID3 algorithm was invented by Ross Quinlan. Quinlan was a computer science researcher in data mining, and decision theory. Received doctorate in computer science at the University of Washington in 1968[2]. ID3 stands for Iterative Dichotomiser 3 Algorithm and it is used to generate a decision tree. Attempt to create a small decision tree.

The basic step of the algorithm is as below:

- In first step the tree is constructed in a top-down recursive divide-and-conquer manner.
- all the training examples are at the root.
- Attributes are categorical
- Examples are partitioned recursively based on selected attributes
- Test attributes are selected on the basis of a statistical measure (e.g., information gain)

In decision tree method information gain (attribute subset selection measure) approach used to obtain suitable property for each node of a generated decision tree. So we select the attribute with the highest information gain as the test attribute for the current node.

Set S is set including s number of data samples whose type attribute can take m potential different values corresponding to m different types of C_i ($i=1,2,3,\dots,m$). Suppose that s_i is the simple number of C_i . So, the required amount of information to classify a given data is

$$I(s_1, s_2, s_3, \dots, s_m) = - \sum_{i=1}^m p_i \log(p_i) \tag{1}$$

Where $P_i = s_i/s$ is the probability that any subset of data samples belonging to categories C_i .

Suppose that A is a property which has v different values. Using the property of A, S can be divided into v number of subsets, in which S_j contains data samples whose attribute A are equal a_j in S set. If property A is selected as the property for test, that is, used to make partitions for current sample set, suppose that S_{ij} is a sample set of type C_i in subset S_i , the required information entropy is

$$E(A) = \sum_{j=1}^v \frac{s_{1j} + s_{2j} + s_{3j} + \dots + s_{mj}}{s} I(s_{1j}, \dots, s_{mj}) \tag{2}$$

Such use of property A on the current branch node corresponding set partitioning samples obtained information gain is

$$Gain(A) = I(s_1, s_2, s_3, \dots, s_m) - E(A) \tag{3}$$

So, the information gain is the metric for selecting a best attribute in the growth of the Decision tree [4].

III. ABOUT DATASET AND TOOL

A. About dataset

The weather dataset obtained from the UCI dataset repository for playing tennis [5]. Which have five attribute like outlook, temperature, humidity, windy, and play. The Weather dataset contain the fourteen instances. Figure1 show the dataset of weather for playing tennis.

	A	B	C	D	E
1	sunny	85	85	FALSE	no
2	sunny	80	90	TRUE	no
3	overcast	83	86	FALSE	yes
4	rainy	70	96	FALSE	yes
5	rainy	68	80	FALSE	yes
6	rainy	65	70	TRUE	no
7	overcast	64	65	TRUE	yes
8	sunny	72	95	FALSE	no
9	sunny	69	70	FALSE	yes
10	rainy	75	80	FALSE	yes
11	sunny	75	70	TRUE	yes
12	overcast	72	90	TRUE	yes
13	overcast	81	75	FALSE	yes
14	rainy	71	91	TRUE	no
15					

Figure1 Weather dataset for playing tennis

B. Data mining Tool

There are a many number of high quality commercial and open source tools for data mining. In this research Weka tool [3] has been used from the perspectives of direct core usage. This serves as a powerful core tool that allows the ability to load, pre-process and visualizes data and also performs standard data mining and machine learning algorithms with sufficient parameterization. So the algorithms can either be applied directly to a dataset or called from custom Java code. In the Weka tool there are many algorithm available and it support for the data mining task like pre-processing, classification, feature selection and clustering here we choose the above Weather dataset and apply the decision tree algorithm.

IV. GENERALIZE ALGORITHM

- First select the initial dataset.
- Find the attribute (features) for the selected dataset.
- Merge those attribute which are correlated and remove those attribute which are not important and irrelevant.
- Apply the dimensionality reduction in dataset.
- Aggregation of attribute and generalization.
- Repeat steps 1 to 5 until further reduction is not possible or information loss

The block diagram of the above generalize algorithm is shown in figure2.

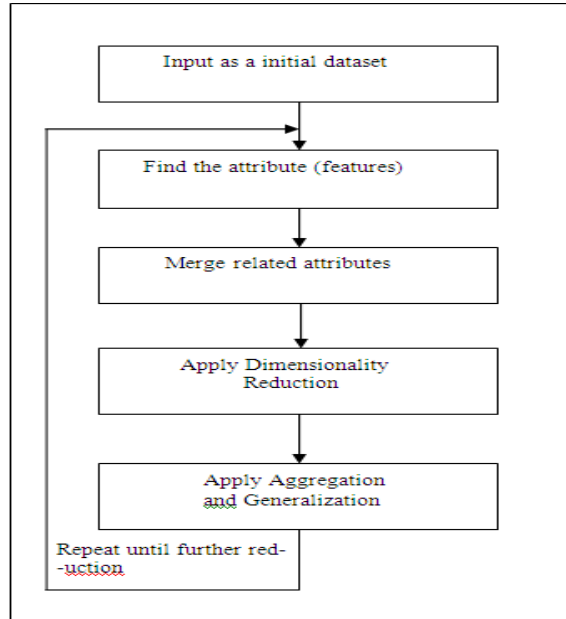


Figure 2 Block diagram of generalize algorithm

V. RESULT COMPARISON

Here we compare the result of the same dataset result in weka tool and java code for the algorithm. The decision tree generated in the weka tool for the weather dataset shown figure 3.

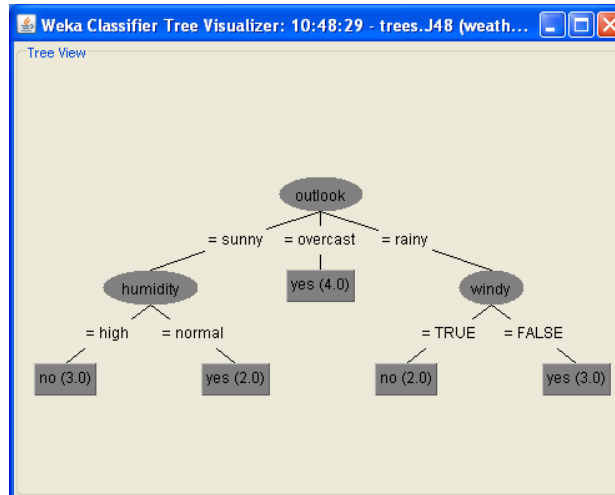


Figure3.Decision tree for the weather dataset in weka tool

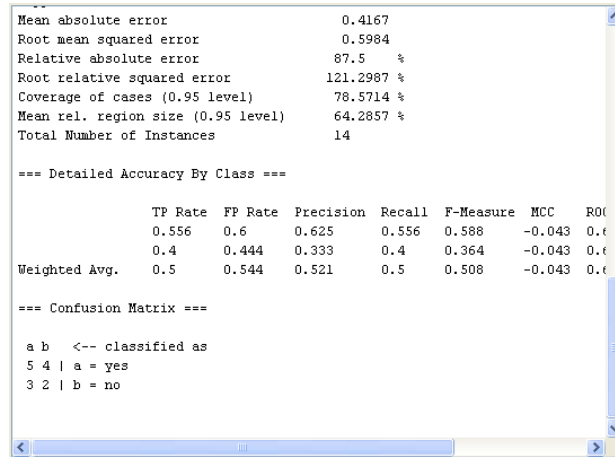


Figure4. Confusion matrix and other values

Figure 3 and figure 4 are the results obtain from the weka data mining tool. Figure3 shown decision tree and figure 4 shown the confusion matrix and other parameters.

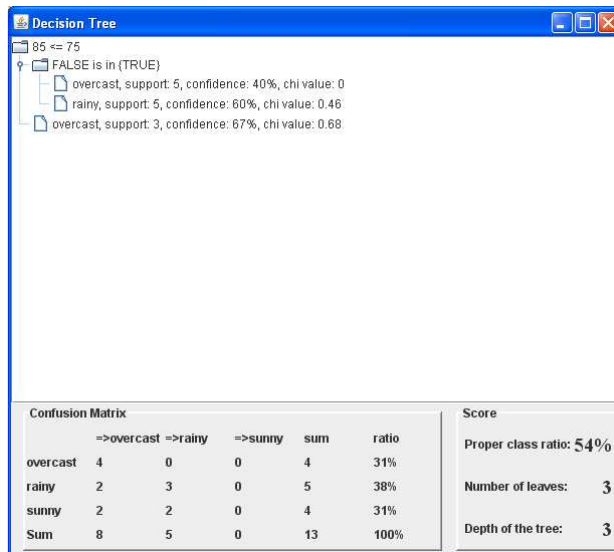


Figure5. Decision tree and confusion matrix

In above figure we see that the confusion matrices contain the information about actual classification done by the classification system. And confusion matrices examined the accuracy produced by the algorithm. Confusion matrices calculate base on following things [5].

- 1) True positive rate (TP) is the proportion of positive cases that were correctly identified, as calculated using the equation[5]:

$$TP = \frac{d}{c+d}$$

- 2) False positive rate (FP) is the proportion of negatives case that were incorrectly classified as positive, as calculated using the equation[5]:

$$FP = \frac{b}{a+b}$$

- 3) Precision (P) is the proportion of the predicted positive cases that were correct, as calculated using the equation[5]:

$$P = \frac{d}{b+d}$$

The confusion matrices have following thing of the context of our study.

- a is the number of correct predictions that an instance is negative,
- b is the number of incorrect predictions that an instance is positive,
- c is the number of incorrect of predictions that an instance negative, and
- d is the number of correct predictions that an instance is positive.

VI. CONCLUSION

In this paper we see that the new propose generalize algorithm and the block diagram of the generalize algorithm. Using the Weather dataset and compare the result of the data mining tool and the improve algorithm using the decision tree. The algorithm selects the highest information gain for the best splitting. Also we see that the different confusion matrices for the same dataset and compare the results.

ACKNOWLEDGEMENT

Author of this paper is thankful to Asst. Prof. Bhavesh Oza, for providing his invaluable time to review the ideas and also answer the queries promptly. Author would also thank to Head of the Department, Prof. D. A. Parikh, their colleagues, friends, classmates, teachers and other guides for providing their immense support and helpful comments to improve this paper.

REFERENCES

- [1] Jiawei Han, Micheline Kamber. Data Mining Concepts and Techniques, 2nd Edition.
- [2] Liu Yuxun, Xie Niuniu, "Improved ID3 Algorithm", pp 465-468 IEEE-2010.
- [3] Margaret H. Dunham, Data Mining Introductory and Advanced Topics, Published by Pearson Education (Singapore) Pte. Ltd. Delhi, India, 2004.
- [4] Chen Jin, Luo De-lin, Mu Fen-xiang. "An improved Id3 decision tree algorithm", page 127-130, IEEE-2009.
- [5] G. Sathyadevi, "APPLICATION OF CART ALGORITHM IN HEPATITIS DISEASE DIAGNOSIS" pp 1283-1287, IEEE-2011
- [6] Imas Sukaesih Sitanggang, Razali Yaakob, Norwati Mustapha, Ahmad Ainuddin B Nuruddin, "An Extended ID3 Decision Tree Algorithm for Spatial Data", page 48-53, IEEE-2011
- [7] J.R. QUINLAN, "Induction of Decision Trees", pp 81-106.