Improving the Prediction Accuracy of Decision Tree Mining with Data Preprocessing

Priyanga Chandrasekar, Kai Qian, Hossain Shahriar
Dept. of Computer Science
Kennesaw State University, Marietta, GA,
{pchandr1, kqian, hshahria}@kennesaw.edu

Prabir Bhattacharya
Dept. of Computer Science
Morgan State University, Baltimore, MD,
prabir.bhattacharya@morgan.edu

Abstract—**A decision tree is an important classification technique in data mining classification. Decision trees have proved to be valuable tools for the classification, description, and generalization of data. J48 is a decision tree algorithm which is used to create classification model. J48 is an open source Java implementation of the C4.5 algorithm in the Weka data mining tool. In this paper, we present the method of improving accuracy for decision tree mining with data preprocessing. We applied the supervised filter discretization on J48 algorithm to construct a decision tree. We compared the results with the J48 without discretization. The results obtained from experiments show that accuracy of J48 after discretization is better than J48 before discretization.**

Keywords-Decision tree; Weka; Preprocessing; Discretization; Entropy; Classification.

I. INTRODUCTION

Data mining is the process of extracting useful information and knowledge from the incomplete, noisy and inconsistent raw data. Data mining extracts information from large dataset and converts it to an understandable form. Data mining is a part of knowledge discovery process.

Classification is a form of data analysis that extracts model describing important data classes. Those models are called classifiers; predict categorical class labels. For example, a classification model can be built to categorize bank loan applications as either safe or risky [1].

Decision tree induction is the process of learning of decision trees from class labeled training tuples. Decision tree is an algorithm which is commonly used to predict model, and also to find out the valuable information through the huge amounts of data classification. A decision tree is a simple flowchart like tree structure, where the topmost node in a tree is the root node [2]. Each leaf node (or terminal node) holds a class label, each internal node (non leaf node) denotes a test on an attribute, and each branch represents an outcome of the test.

J48 is an extension of ID3. The additional features of J48 are accounting for missing values, decision trees pruning, continuous attribute value ranges, derivation of rules, etc.

The remaining of this paper is organized as follows. Section 2 gives a brief note about the related works. Section 3 presents and discusses our research methodology followed by the description of the microarray dataset we have used in our experiments as well as the experimental setup. In Section 4, the evaluation of results is presented. Finally, Section 5 presents conclusions.

II. RELATED WORKS

For surveying the problem of improving decision tree classification algorithm for large data sets, several algorithms have been developed for building DTs of large data sets.

Kohavi & John 1995 [3], searched for parameter settings of C4.5 decision trees that would result in optimal performance on a particular data set. The optimization objective was "optimal performance" of the tree, i.e., the accuracy measured using 10-fold cross-validation. J48, Random Forest, Naive Bayes etc. algorithms [4] are used for disease diagnosis as they led to good accuracy. They were used to make predictions. The dynamic interface can also use the constructed models that mean the application worked properly in each considered case. The classification algorithms [5] Naive Bayes, decision tree (J48), Sequential Minimal Optimization (SMO), and Instance Based for K- Nearest neighbor (IBK) and Multi-Layer Perception are compared by using matrix and classification accuracy. Three different breast cancer databases have been used and classification accuracy is presented on the basis of 10-fold cross validation method. A combination at classification level is accomplished between these classifiers to get the best multi-classifier approach and accuracy for each data set. Diabetes and cardiac diseases [6] are predicted using Decision Tree and Incremental Learning at the early stage.

Liu X.H 1998 [7], proposed a new optimized algorithm of decision trees. On the basis of ID3, this algorithm considered attribute selection in two levels of the decision tree and the classification accuracy of the improved algorithm had been proved higher than ID3. Liu Yuxun & Xie Niuniu 2010 [8], solving the problem of a decision tree algorithm based on attribute importance is proposed. The improved algorithm uses attribute-importance to increase the information gain of attributes which has fewer attributions and compares ID3 with improved ID3 by an example.

0730-3157/17 $31.00 © 2017 IEEE
DOI 10.1109/COMPSAC.2017.146

Experimental analysis of the data shows that the improved ID3 algorithm can get more reasonable and more effective rules. Gaurav & Hitesh 2013 [9], propose C4.5 algorithm which is improved by the use of L"Hospital Rule, this simplifies the calculation process and improves the efficiency of decision making algorithms.

III. RESEARCH METHODOLOGY

Our methodology is to learn about the dataset, apply J48 decision tree classification algorithms and get the accuracy of the algorithm. In preprocessing step, apply the supervised discretization filter on the dataset along with the J48 classification algorithm and find the accuracy. Finally comparing both accuracy and find out which one is better.

A. Leukemia Dataset

In our study we have used a real world leukemia microarray experiment performed by [Golub et al. 1999]. Leukemia is a cancer of bone marrow or blood cells. In general, leukemia's can be grouped into four categories. Myeloid and lymphoid leukemia's can be acute or chronicle whereas myeloid and lymphoid both denote cell types involved. Thus, four main types of leukemia's are: Acute Myeloid Leukemia (AML), Chronic Myeloid Leukemia (CML), Acute Lymphoblastic Leukemia (ALL) and Chronic Lymphoblastic Leukemia (CLL).

In the dataset provided by [Golub et al. 1999], each microarray experiment corresponds to a patient (example); each example consists 7129 genes expression values (features). Each patient has a specific disease (class label), corresponding to two kinds of leukemia (ALL and AML). There are 72 patients (47 ALL and 25 AML). The original study of [Golub et al. 1999] split patients into two disjoint sets2: the training set contains 38 examples (27 ALL and 11 AML) and the test set contains 34 examples (20 ALL and 14 AML). Considering the shortage of examples, it is a common technique in machine learning to use cross-validation or bootstrap [Kohavi 1995, Hastie et al. 2001] rather than isolating training and test sets.

In our study, training dataset participate in the test dataset. Hence our study uses the training set contains 38 examples (27 ALL and 11 AML) and the test set contains 38 examples (27 ALL and 11 AML).

B. WEKA

Weka is open-source software developed at the University of Waikato and the programming language is based on Java. Weka has 4 different applications, Explorer, Experimenter, KnowledgeFlow and Simple CLI. Knowledge Flow is a node and linked based interface and Simple CLI is the command line prompt version where each algorithm is run by hand. In our study, we used Explorer applications of the Weka.

WEKA is an innovatory tool in the history of the data mining and machine learning research communities. By putting efforts since 1994 this tool was developed by WEKA team. WEKA contains many inbuilt algorithms for data mining and machine learning. Weka implements algorithms for data preprocessing, classification, regression, clustering, association rules; it also includes a visualization tools.

C.J48 Classifier

Classification is the process assigning an appropriate class label to an instance (record) in the dataset. Classification is generally used in supervised datasets where there is a class label for each instance. In our study we applied J48 classifier in the dataset. J48 Classifier uses the normalized version of Information Gain which is Gain Ratio for building trees as the splitting criteria. It has both reduced error pruning and normal C4.5 pruning option. In our experiments we have used the algorithm J48 (with default parameters) from Weka [Witten and Frank 2005], a library of several machine learning algorithms. J48 is a Java implementation of the well- known C4.5 algorithms [Quinlan 1993]. J48 uses a modified version of the entropy measure from information theory.

D.Pre-processing

Data usually comes in mixed format: nominal, discrete, and/or continuous. Discrete and continuous data are ordinal data types having orders among values, while nominal values do not possess any order amongst them. Discrete data are spaced out with intervals in a continuous spectrum of values. We used discretization as data preprocessing method.

Discretization: Discretization process will easily interpret numerical attributes turning into nominal (categorical) ones. This process is done by dividing a continuous range into subgroups. Suppose there are 200 people in a group that want to apply for a bank loan and their ages are between 20 and 80. If the bank workers want to categorize them, they have to put them into some groups. For example, one can categorize people between 20 and 40 as young, people between 40 and 65 as middle aged and 65 to 80 as old. So there will be three subgroups, which are; young, middle-aged and old. These subgroups can be increased depending on the choice of the field expert. This makes it easy to understand and easy to standardize.

Discretization of continuous attributes is both a requirement and a way of performance improvement for many machine learning algorithms. The main benefit of discretization is that some classifiers can only work on the nominal attributes, but not numeric attributes. Another advantage is that it will increase the classification accuracy of tree and rule based algorithms that depend on nominal data.

Discretization can be grouped into two categories, Unsupervised Discretization and Supervised Discretization. As the name implies Unsupervised Discretization is generally applied to datasets having no class information. The types of Unsupervised Discretization are: Equal Width Binning, Equal Frequency Binning mainly but more complex ones are based on clustering methods [10]. Supervised Discretization techniques as the name suggests takes the class information into account before making subgroups. Supervised methods are mainly based on Fayyad-Irani [11] or Kononenko [12] algorithms.

Weka uses Fayyad-Irani method as default, so in our study we used Fayyad-Irani Discretization method. Weka has the Discretization algorithm under the preprocessing tab. As shown in Fig. 1, it is embedded right under supervised and attribute options. Fayyad-Irani Discretization method is a supervised hierarchical split method, which will use the class information entropy of candidate partitions to select boundaries for discretization.

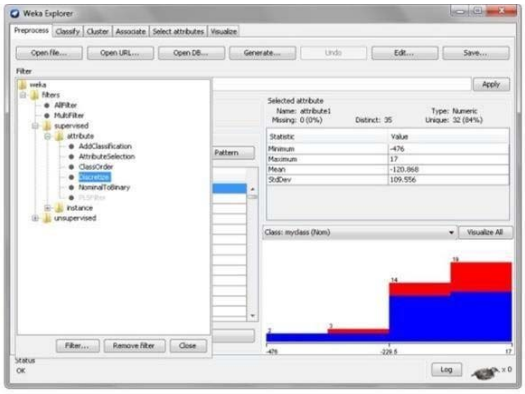

Fig. 1. Selecting Discretization from Preprocess Tab

Class information entropy is a measure of purity and it measures the amount of information which would be needed to specify to which class an instance belongs. It considers one big interval containing all known values of a feature and then recursively partitions this interval into smaller subintervals until an optimal number of intervals are achieved.

One of the supervised discretization methods, introduced by Fayyad and Irani, is called entropy based discretization. The supervised discretization methods handle sorted feature values to determine the potential cut points such that the resulting cut point has the strong majority of one particular class. The cut point for discretization is selected by evaluating the favorite disparity measure (i.e., class entropies) of candidate partitions. In entropy based discretization, the cut- point is selected according to the entropy of the candidate cut- points. Entropy of the one attribute is calculated using the formula

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i \quad (1)$$

The entropy of the two attributes are calculated by using the formula

$$E(T, X) = \sum_{c \in X} P(c) E(c) \quad (2)$$

The Entropy Gain refers to how much entropy you gain by splitting a data set into two bins. Entropy Gain performs splits that maximize the improvement to the information we get from our data. Gain is defined as

$$\text{Gain}(T, X) = \text{Entropy}(T) - \text{Entropy}(T, X) \quad (3)$$

IV. EVALUATION

While Evaluating the J48 classifier, we need to concentrate on false positive and false negative. A false positive means negative instances that are incorrectly assigned to the positive class. A false negative means positive instances that are incorrectly assigned to the negative class. A false positive can have more impact than the false negative. Initial experiment was to investigate the effect of discretization to the learning time and prediction accuracy of the J48 classifier. To figure that out, we need to run the algorithm on the dataset without discretization. Then we need to apply discretization and find out the results and compare the accuracy of the both.

The Confusion matrix for training dataset without/with preprocessing is shown in Table I and the confusion matrix for test dataset without/with preprocessing is shown in Table II. Evaluation was carried out in the test dataset, which consists of 38 examples (27 ALL and 11 AML). The result shows that for the preprocessed dataset, accuracy of the decision tree was increased.

Table I Confusion matrix for training dataset

Confusion Matrix of Training Dataset					
Without Preprocessing			With Preprocessing		
a	b	Classified as	a	b	Classified as
23	4	a= ALL	24	3	a= ALL
2	9	b= AML	2	9	b= AML

Table II Confusion matrix for test dataset

Confusion Matrix of Test Dataset					
Without Preprocessing			With Preprocessing		
a	b	Classified as	a	b	Classified as
25	2	a= ALL	26	1	a= ALL
6	5	b= AML	3	8	b= AML

Performance Analysis: The accuracies obtained by combining J48 Classification without discretization and with discretization were carried in both training and test dataset. The accuracies obtained were charted in Fig. 2 for analysis. The result shows that the Discretization of the numerical attributes increased the performance of J48 by approximately 2.63% for training dataset and 10.53% for test dataset

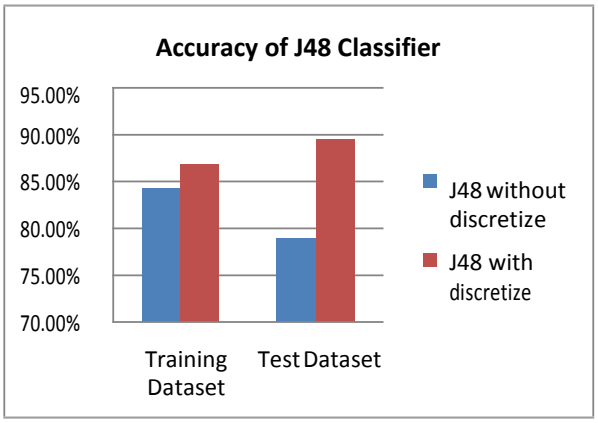

Fig.2. Performance Analysis

V. CONCLUSION

The first step of Data Mining, preprocessing process showed its benefits during the classification accuracy performance tests. In this paper, entropy- based discretization method is used for improving the classification accuracy for datasets including continuous valued features. In the first phase, the continuous valued features of the given dataset are discretized. Second phase, we tested the performance of this approach with the J48 classifier and compared with performance of J48 classifier without discretization.

Discretization of the numerical attributes increased the performance of J48 by approximately 2.63% for training dataset and 10.53% for test dataset. The result proves that the optimal level of discretization improves both the model construction time and prediction accuracy of the J48 classifier. Other benefit of discretization came after the visualization of J48, making the tree easy to interpret, because of the cutting-points it assigned after the discretization of numerical attributes. Thus the test results shows that combining J48 classifier with the proper data pre- processing can improve the prediction accuracy and also proves that the preprocessing phase has a larger impact in the performance of the J48 classifier.

REFERENCES

[1] Mehmed Kantardzic, “Data Mining: Concepts, Models, Methods, and Algorithms”, ISBN: 0471228524, John Wiley & Sons, 2003.

[2] Sushmita Mitra, & Tinku Acharya, “Data Mining Multimedia, Soft Computing, and Bioinformatics”, John Wiley & Sons, Inc, 2003.

[3] Tea Tusar, “Optimizing Accuracy and Size of Decision Trees”, Department of Intelligent Systems, Jozef Stefan Institute, Jamova 39, SI-1000 Ljubljana, Slovenia, 2007

[4] Robu, R.; Hora, C., "Medical data mining with extended WEKA," Intelligent Engineering Systems (INES), 2012 IEEE 16th International Conference on , vol., no., pp.347,350, 13-15 June 2012

[5] Salama, G.I.; Abdelhalim, M.B.; Zeid, M.A., "Experimental comparison of classifiers for breast cancer diagnosis," Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on , vol., no., pp.180,185, 27-29 Nov.,2012.

[6] UM, Ashwinkumar, and Anandakumar KR. "Predicting Early Detection of Cardiac and Diabetes Symptoms using Data Mining Techniques.",IEEE,pp:161-165,2011

[7] Weiguo Yi, Jing Duan, &Mingyu Lu, “Optimization of Decision Tree Based on Variable Precision Rough Set”, International Conference on Artificial Intelligence and Computational Intelligence, 2011.

[8] Liu Yuxun, &XieNiuniu, “Improved ID3 Algorithm”, IEEE, 2010.

[9] Gaurav L. Agrawal, & Prof. Hitesh Gupta, “Optimization of C4.5 Decision Tree Algorithm for Data Mining Application”, International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 3, March 2013.

[10] Joa˜o Gama and Carlos Pinto. Discretization from data streams: applications to histograms and data mining. In Proceedings of the 2006 ACM symposium on Applied computing, SAC ’06, pages 662–667, New York, NY, USA, 2006. ACM.

[11] Usama M. Fayyad and Keki B. Irani. Multi-interval discretization of continu- ousvalued attributes for classification learning. In Thirteenth International Joint Conference on Articial Intelligence, volume 2, pages 1022–1027. Morgan Kauf- mann Publishers, 1993.

[12] Igor Kononenko. On biases in estimating multi-valued attributes. In 14th Inter- national Joint Conference on Articial Intelligence, pages 1034–1040, 1995.