Volume 2, Issue 11, November 2017 ISSN (online): 2456-0006
International Journal of Science Technology Management and Research
Available online at: www.ijstmr.com

Appropriate Attribute Selection To Construct Classifier Based On Decision Tree

Munzzir Sheikh
M. Tech. (CSE) IV Sem
Lord Krishna College of Technology
Indore M.P. India

Vijay Kumar Verma
Asst. Prof. C.S.E. Dept.
Lord Krishna College of Technology
Indore M.P. India

Abstract: ***Classification techniques are capable for handling and processing large amount of data and. Classification techniques are used to predict categorical class labels and classifies data based on training set. Classification procedure is recognized method for repeatedly making such decisions in new situations. Creation of a classification procedure from a set of data for which the exact classes are known in advance is termed supervised learning. Classification examples include assigning individuals to credit status, initial diagnosis of a patient's disease etc. In this paper we proposed attribute selection measure to decide most suitable attribute for construing accurate classifiers for loan approval class.***

Keywords: Classification, Data mining, Predict, Loan, Attribute.

I. INTRODUCTION

Classification is the process of finding a model that describes and distinguishes data classes or concepts. Classification methods can handle both numerical and categorical attributes. Constructing fast and accurate classifiers for large data sets is an important task in data mining and knowledge discovery [7,8].
Classification is two-step process

1. Model construction
2. Model usage

Describing a set of predetermined classes. Each tuple/sample is assumed to belong to a predefined class, as determined by the class label attribute. The set of tuples used for model construction is training set. The model is represented as classification rules, decision trees, or mathematical formulae. For classifying future or unknown objects. Estimate accuracy of the model. The known label of test sample is compared with the classified result from the model. Accuracy rate is the percentage of test set samples that are correctly classified by the model. Test set is independent of training set (otherwise over fitting). If the accuracy is acceptable, use the model to classify data tuples whose class labels are not known.
Basic algorithm (a greedy algorithm) is constructed in a top-down recursive divide-and-conquer manner. At start, all the training examples are at the root. Attributes are categorical (if continuous-valued, they are discretized in advance). Examples are partitioned recursively based on selected attributes. Test attributes are selected on the basis of a heuristic or statistical measure (e.g., information gain).*f* Conditions for stopping partitioning. All samples for a given node belong to the same class. There are no remaining attributes for further partitioning majority voting is employed for classifying the leaf. There are no samples left. Typical applications are Credit, Loan approval, Medical diagnosis, Fraud detection, Web page categorization etc[1,9].

II. CLASSIFICATION TECHNIQUES

Some of the important classification techniques are

1. Decision tree induction
2. Bayesian Classification
3. Support Vector Machines (SVM)
4. K-Nearest Neighbor Classifies
5. Rough Set Approach

Decision tree induction classifier is a simple and widely used classification technique. It applies a strait forward idea to solve the classification problem. Decision Tree Classifier poses a series of carefully crafted attributes of the test record. Each time it receives an answer, a follow-up question is asked until a conclusion about the class label of the record is reached [10.
A Bayesian classifier is based on the idea that the role of a (natural) class is to predict the values of features for members of that class. The idea behind a Bayesian classifier is that, if an agent knows the class, it can predict the values of the other features. If it does not know

the class, Bayes' rule can be used to predict the class given (some of) the feature values. In a Bayesian classifier, the learning agent builds a probabilistic model of the features and uses that model to predict the classification of a new example.
Support Vector Machines is a supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting).
K-Nearest Neighbor Classifies is used to classify unlabeled observations by assigning them to the class of the most similar labeled examples. Characteristics of observations are collected for both training and test dataset. Parameter k which decides how many neighbors will be chosen for kNN algorithm. The appropriate choice of k has significant impact on the diagnostic performance of kNN algorithm. A large k reduces the impact of variance caused by random error, but runs the risk of ignoring small but important pattern.
Rough set is one of the most important classifier used in data mining. Rough set is first described by Polish computer scientist Zdzisław I. Rough set is a formal approximation of a crisp set (i.e., conventional set) in terms of a pair of sets which give the lower and the upper approximation of the original set. In rough set theory the lower- and upper-approximation sets are crisp sets, but in other variations, the approximating sets may be fuzzy sets.

III. LITERATURE REVIEW

In 2012 Akhil jabbar et al. proposed "Heart Disease Prediction System using Associative Classification and Genetic Algorithm". They proposed efficient associative classification algorithm using genetic approach for heart disease prediction. The main advantage of genetic algorithm is the discovery of high level prediction rules is that the discovered rules are highly comprehensible, having high predictive accuracy and of high interestingness values. The proposed method helps in the best prediction of heart disease which even helps doctors in their diagnosis decisions [1].

In 2013 Akhil Jabbar et al proposed "Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection". They proposed a new feature selection method using ANN for heart disease classification. For rank the attributes which contribute more towards classification of heart disease they applied different feature selection methods, and indirectly reduce the no. of diagnosis tests to be taken by a patient. The proposed method eliminates useless and distortive data[2] .

In 2014 N. S. Nithya et al proposed "Gain ratio based fuzzy weighted association rule mining classifier for medical diagnostic interface". They showed that earlier model based on information gain and fuzzy association rule mining algorithm for extracting both association rules and membership functions are not feasible. They used large number of distinct values. They modify gain ratio based fuzzy weighted association rule mining and improve the classifier accuracy [3] .

In 2015 Jaimini Majali et al proposed "Data Mining Techniques for Diagnosis and Prognosis of Cancer". They used data mining techniques for diagnosis and prognosis of cancer. They presented a system for diagnosis and prognosis of cancer using Classification and Association approach in Data Mining. They used FP algorithm in Association Rule Mining to conclude the patterns frequently found in benign and malignant patients [4] .

In 2016 Nikhil N. Salvithal et al proposed "Appraisal Management System using Data mining Classification Technique". The proposed assorted classifier algorithms applied on Talent dataset to spot the talent set so as to judge the performance of the individual. Finally counting on accuracy one best suited classifier is chosen this method has been used to construct classification rules to predict the potential talent that for promotion or not [5].

In 2016 Tanvi Sharma et al proposed "Performance Analysis of Data Mining Classification Techniques on Public Health Care". The proposed study focused on the application of various data mining classification techniques using different machine learning tools such as WEKA and Rapid miner over the public healthcare dataset for analyzing the health care system. The percentage of accuracy of every applied data mining classification technique is used as a standard for performance measure. The best technique for particular data set is chosen based on highest accuracy [6].

IV. PROBLEM STATEMENT

Fundamental problem that are to be consider for a classifiers are

1. Accuracy:- Predicting class label predictor accuracy: guessing value of predicted attributes.
2. Speed:- Time to construct the model (training time) time to use the model (classification/prediction time).
3. Robustness:-Ability to handle noise and missing values
4. Scalability:- Efficiency in disk-resident databases.
5. Interpretability:- understanding and insight provided by the model .

V. PROPOSED APPROACH

Let node N represents or hold the tuple of partition D. The attribute with the highest information gain is chosen as the splitting attribute for the node N. The expected information needed to classify a tuple in D is given by

$$Info(D) = -\sum_{i=1}^{m} pi \log_2 pi$$

Where Pi is the probability that an arbitrary tuple in D belongs to class Ci and is estimated by |Ci,D| / |D|. Info (D) is the average amount of information needed to identify the class label of a tuple in D. Info (D) is also known as the entropy of D. The expected information required to classify a tuple from D, based on the partitioning by attribute A is calculated by

$$Info_A(D) = \sum_{i=1}^{V} \frac{|Dj|}{|D|} \times Info(D_j)$$

Information gain is defined as the difference between the original information requirement (i.e. based on the classes) and the new requirement (i.e. obtained after partitioning on A).

$$Gain(A) = Info(D) - Info_A(D)$$

The Gini Index considers a binary split for each attribute. The Gini Index measures the impurity of D, a data partition or set of training tuples as

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^2$$

Consider a simple dataset for credit card approval

S No	Age	Income	Employee	Credit rating	Credit card approval
1	≤30	High	Govt.	Low	No
2	≤30	High	Govt.	High	No
3	31…40	High	Govt.	Low	Yes
4	>40	Medium	Govt.	Low	Yes
5	>40	Low	Private	Low	Yes
6	>40	Low	Private	High	No
7	31…40	Low	Private	High	Yes
8	≤30	Medium	Govt.	Low	No
9	≤30	Low	Private	Low	Yes
10	>40	Medium	Private	Low	Yes
11	≤30	Medium	Private	High	Yes
12	31…40	Medium	Govt.	High	Yes
13	31…40	High	Private	Low	Yes
14	>40	Medium	Govt.	High	No

The classification of the target is "Credit card approval" which can be Yes or No.
Weather attributes outlook, temperature, humidity and wind speed. They can take the following values:
Age = {≤30, 31...40, >40}
Income = {High, Medium, Low}
Employee = {Govt., Private}
Credit rating = {Low, High}

Info (D) =-9/14*log2 (9/14)-5/14*log2 (5/14) = 0.94.
Calculation for the first attribute
InfoAge(D) = -5/14*Info (D, ≤30) -4/14*Info (D, 31...40) -5/14* Info (D, >40)
=– 5/14*0.9710-4/14*0 – 5/14*0.9710
InfoAge(D) = 0 .694
Gain (D, Income) = Info (D)- Gain Age (D)
Gain (D, Age) = 0 .246

Gain (D, Income) = 0.048
Gain(D, Employee) = 0.0289
Gain (D, Employee) = 0.1515
Here age has the highest gain so we select the root attribute as age.

VI. EXPERIMENTAL ANALYSIS

We implemented our algorithm in VB Dot Net 2013 and SQL server R2 2010 on an Intel Pentium i3 Processor personal computer with 4 GB RAM on Windows 7 having 400 GB Hard disk. We have taken 100,200 and 300 records with 4 and 5 attribute.

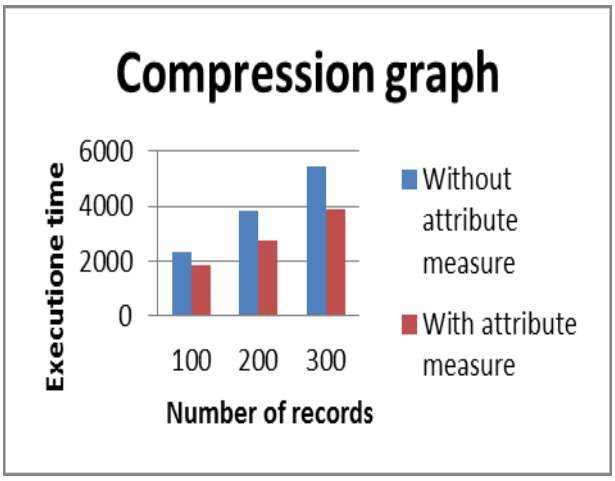

Figure 1. Comparision graph

CONCLUSION

There are several approach have developed to improve the accuracy and time of the decision tree classifiers. Our proposed approaches construct a decision tree by deciding parameters for the attribute. This not only reduces complexity of the decision tree but improve the accuracy and execution time. From the figure 1 it is clear that proposed approach improve execution time as compared to traditional approach.

REFERENCE

1. M. Akhil jabbar & Dr. Priti Chandrab "Heart Disease Prediction System using Associative Classification and Genetic Algorithm" International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies-ICECIT, 2012.
2. M. Akhil Jabbar, B.L Deekshatulu & Priti Chandra "Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection" Global Journal of Computer Science and Technology Neural & Artificial Intelligence Volume 13 Issue 3 Version 1.0 Year 2013 International Research Journal Publisher: Global Journals Inc. (USA)
3. N S Nithya and K Duraiswamy "Gain ratio based fuzzy weighted association rule mining classifier for medical diagnostic interface" Sadhana¯ Vol. 39, Part 1, February 2014, pp. 39–52. Indian Academy of Sciences
4. Jaimini Majali, Rishikesh & Niranjan, Vinamra Phatak "Data Mining Techniques For Diagnosis And Prognosis Of Cancer" International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 3, March 2015
5. Nikhil N. Salvithal "Appraisal Management System using Data mining "International Journal of Computer Applications (0975 – 8887) Volume 135 – No.12, February 2016
6. Tanvi Sharma, Anand Sharma & Vibhakar Mansotra "Performance Analysis of Data Mining Classification Techniques on Public Health Care Data" International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 4, Issue 6, June 2016
7. B Rosiline Jeetha "Efficient Classification Method For Large Dataset By Assigning The Key Value In Clustering" International Journal of Computer Science and Mobile Computing A Monthly Journal of Computer Science and Information Technology ISSN 2320–088X, Vol. 3, Issue. 1, January 2014,
8. Divya Tomar and Sonali Agarwal " A survey on Data Mining approaches for Healthcare" international Journal of Bio-Science and Bio-Technology Vol.5, No.5 (2013), pp. 241-266 http://dx.doi.org/10.14257/ijbsbt.2013.5.5.2
9. V. Krishnaiah , Dr. G.Narsimha, Dr. N. Subhash Chandra "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (1) , 2013.
10. S. Olalekan Akinola, O. Jephthar Oyabugbe Accuracies and Training Times of Data Mining Classification Algorithms: An Empirical Comparative Study" Journal of Software Engineering and Applications, 2015, 8, 470-477 Published Online September 2015 in SciRes.

.