

A Study of Some Data Mining Classification Techniques

Mr. Sudhir M. Gorade¹, Prof. Ankit Deo²,
Prof. Preetesh Purohit³

¹Dept.Of Computer Science and Engineering, SVCE Indore, M.P, India

²Professor Dept. Of Computer Science and Engineering, SVCE Indore, M.P. ,India

³Professor Dept. Of Computer Science and Engineering, SVCE Indore, M.P. ,India

Abstract - An Classification is one the most useful and important techniques. Classification techniques are useful to handle large amount of data. Classification is used to predict categorical class labels. Classification models are used to classifying newly available data into a class label. 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. Classification predicts categorical class labels and classifies data based on the training set. Classification is two steps processes. In this paper we present a study of various data mining classification techniques like Decision Tree, K-Nearest Neighbor, Support Vector Machines, Naive Bayesian Classifiers, and Neural Networks.

Keywords:- classification, prediction ,class label, model, categories.

1. INTRODUCTION

Classification used two steps in the first step a model is constructed based on some training data set, in seconds step the model is used to classify a unknown tuple into a class label.

1.1 Step 1 - Construction of a model

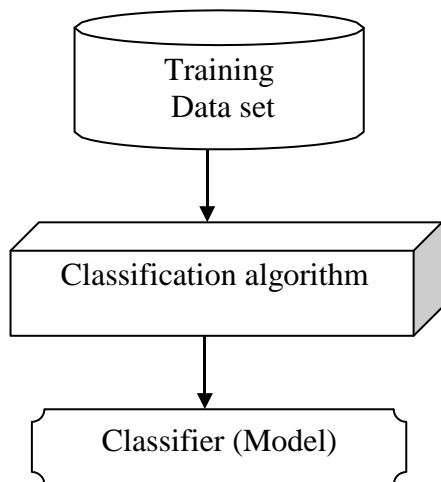


Fig.1 - Model construction step

1.2 Step 2 - Model used for unknown tuple

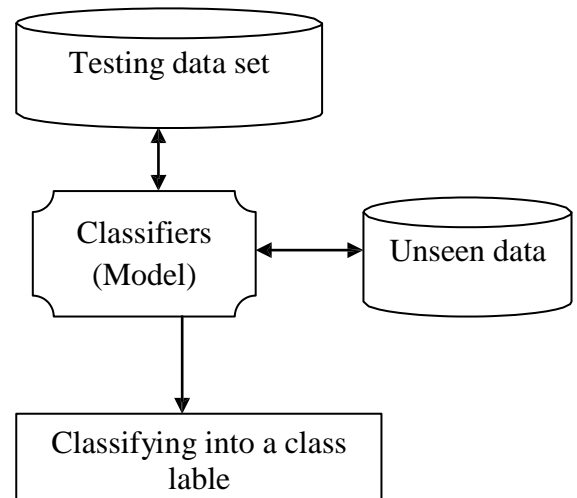


Fig.2 - Use of classifier

2. CHARACTERISTICS OF CLASSIFIERS

Each and every classifier has some quality which differential the classifier form other. The properties are known as characteristics of the classifiers. These characteristics are **Correctness** :- How a classifier classifies tuple accurately is based on these characteristics. To check accuracy there are some numerical values based on number of tuple classify correctly and number of tuple classify wrong.

Time :- How much time is required to construct the model? This also includes the time to use by the model to classify then number of tuple (prediction time). In other word this refers to the computational costs.

Strength :- ability to classify a tuple correctly even tuple has a noise. Noise can be wrong value or missing value.

Data Size :- Classifiers should be independent form the size of the database. Model should be scalable. The performance of the model is not dependent on the size of the database.

Extendibility :- Some new feature can be added whenever required. This feature is difficult to implement.

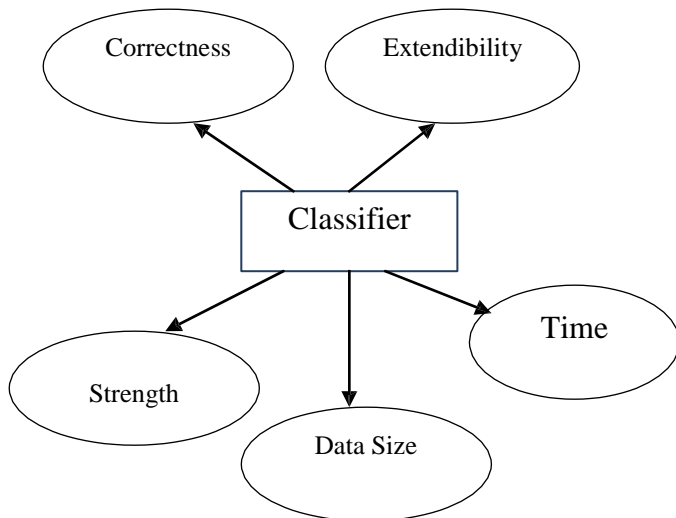


Fig.3 - Characteristics of a Classifier

3. LITERATURE SURVEY

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 S. Olalekan Akinola, O. Jephthar Oyabugbe proposed “Accuracies and Training Times of Data Mining Classification Algorithms: An Empirical Comparative Study”. They proposed study was designed to determine how data mining classification algorithm perform with increase in input data sizes. They used three data mining classification algorithms

Decision Tree, Multi-Layer Perceptron (MLP) Neural Network and Naïve Bayes were subjected to varying simulated data sizes. The time taken by the algorithms for trainings and accuracies of their classifications were analyzed for the different data sizes[4].

In 2015 Jaimini Majali, Rishikesh Niranjana & Vinamra Phatak 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[5].

In 2016 Nikhil N. Salvithal & R.B. Kulkarni 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[6].

In 2016 Tanvi Sharma & Anand Sharma 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[7].

4. VARIOUS CLASSIFICATION MODEL

The main goals of a Classification algorithm are to maximize the predictive accuracy obtained by the classification model. Classification task can be seen as a supervised technique where each instance belongs to a class. There are several model techniques are used for classification some of them are [8,9,10].

- Decision Tree,
- K-Nearest Neighbor,
- Support Vector Machines,
- Naive Bayesian Classifiers,
- Neural Networks.

4.1 Decision Trees -

A decision tree is a classifier and used recursive partition of the instance space. This model consists of nodes and a root. Nodes other than root have exactly one incoming edge.

Intermediate node is test nodes after performing a test they generate outgoing edge. Nodes without outgoing are called leaves (also known as terminal or decision nodes). In a decision tree, each internal node splits the instance space into two or more sub-spaces a certain discrete function of the input attributes values.

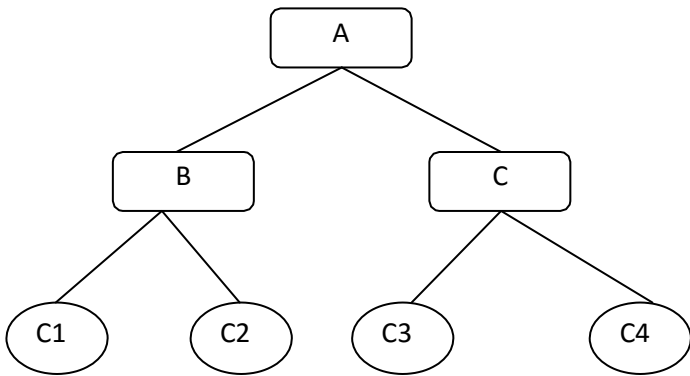


Fig.4- Decision Tree Classifiers

A denotes the root of the tree. B, C are internal nodes denote a test on a particular attribute and C1, C2, C3 and C4.

4.2 K-Nearest neighbor

This classifiers are based on learning by training samples. Each sample represents a point in an n-dimensional space. All training samples are stored in an n-dimensional pattern space. When given an unknown sample, a k-nearest neighbor classifier searches the pattern space for the k training samples that are closest to the unknown sample. "Closeness" is defined in terms of Euclidean distance, where the Euclidean distance, where the Euclidean distance between two points, $X=(x1,x2,\dots,xn)$ and $Y=(y1,y2,\dots,yn)$ is denoted by $d(X, Y)$.

$$d(X, Y) = \sqrt{\sum_{i=1}^n (xi - yi)^2}$$

Nearest neighbor classifiers assign equal weight to each attribute. Nearest neighbor classifiers can also be used for prediction, that is, to return a real-valued prediction for a given unknown sample.

4.3 Bayesian classifiers

Bayesian classifiers are statistical classifiers. They can predict class membership based on probabilities. The Naive Bayes Classifier technique is particularly suited when the dimensionality of the inputs is high. Naive Bayes can often outperform more sophisticated classification methods. Let D be a training set associated class labels. Each tuple is represented by an n-dimensional attributes, $A1, A2,\dots, An$. Suppose that there are m classes, $C1, C2,\dots, Cm$. Given a tuple, X, the classifier will predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naïve Bayesian classifier predicts that tuple x belongs to the class Ci if and only if $P(Ci / X) > P(Cj/X)$ for $1 \leq j \leq m, j$

$\neq i$. Thus we maximize $P(Ci / X)$. The class Ci for which $P(Ci / X)$ is maximized is called the maximum posteriori hypothesis. By Bayes' theorem

$$P(Ci / X) = \frac{P\left(\frac{X}{Ci}\right)P(Ci)}{P(X)}$$

$P(X)$ is constant for all classes, only $P(X/Ci) P(Ci)$ need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C1) = P(C2) = \dots = P(Cm)$, and we would therefore maximize $P(X/Ci)$. Otherwise, we maximize $P(X/Ci)P(Ci)$.

4.4 Neural Networks.

Neural Network used gradient descent method based on biological nervous system having multiple interrelated processing elements. These elements are known as neurons. Rules are extracted from the trained Neural Network to improve interoperability of the learned network. To solve a particular problem NN used neurons which are organized processing elements.

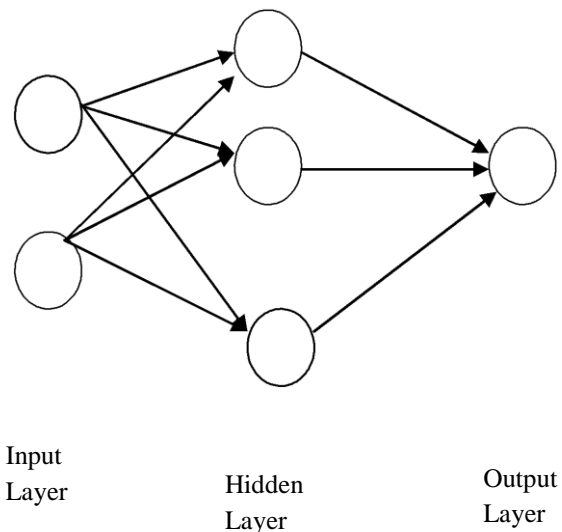


Fig. 5 - Neural networks as a classifier

Neural Network is used for classification and pattern recognition. An NN changes its structure and adjusts its weight in order to minimize the error. Adjustment of weight is based on the information that flows internally and externally through network during learning phase. In NN multiclass, problem may be addressed by using multilayer feed forward technique, in which Neurons have been employed in the output layer rather using one neuron

4.5 Support Vector Machine (SVM)

SVM is a very effective method for regression, classification and general pattern recognition. It is considered a good classifier because of its high generalization performance without the need to add a priori knowledge, even when the dimension of the input space is very high. It is considered a good classifier because of its high generalization performance without the need to add a priori knowledge,

even when the dimension of the input space is very high. For a linearly separable dataset, a linear classification function corresponds to a separating hyper plane $f(x)$ that passes through the middle of the two classes, separating the two. SVMs were initially developed for binary classification but it could be efficiently extended for multiclass problems.

5. ADVANTAGE AND DISADVANTAGE

Each and every model has some advantage and disadvantage. We give some advantage and disadvantage of these methods

Model	Advantage	Disadvantage
Decision Trees	Easy to interpret and explain.	Do not work best for uncorrelated variables.
K-Nearest Neighbor	Effective if training data is large.	Need to determine values of parameter
Support Vector Machines	Useful for non-linearly separable data	
Naive Bayesian Classifiers	Handles real and discrete data.	Assumption is independence of features
Neural Networks	It is a non-parametric method.	Extracting the knowledge (weights in ANN) is very difficult

6. CONCLUSION

There are several classification techniques in data mining and each and every technique has its advantage and disadvantage. Decision tree classifiers, Bayesian classifiers, classification by back propagation, support vector machines, these techniques are eager learners they use training tuples to construct a generalization model.

Some of them are lazy learner like nearest-neighbor classifiers and case-based reasoning. These store training tuples in pattern space and wait until presented with a test tuple before performing generalization.

7. REFERENCES

[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] 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. <http://www.scirp.org/journal/jsea>

[5] 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

[6] Nikhil N. Salvithal "Appraisal Management System using Data mining" International Journal of Computer Applications (0975 - 8887) Volume 135 - No.12, February 2016

[7] 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

[8] 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 IJCSMC, Vol. 3, Issue. 1, January 2014, pg.319 - 324

[9] 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.2013.5.5.25>

[10] 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, 39 - 45