International Journal of Advanced Technology in Engineering and Science
Vol. No. 09, Issue No. 07, July 2021
www.ijates.com
ijates
ISSN 2348 - 7550

Data Mining Classification Techniques Review

[1]Kajal Gupta, [2]Km.Renuka, [3]Amrita Yadav, [4]Vaibhavi Sushil

Computer Science & Engineering,
Buddha Institute of Technology, Gida, Gorakhpur UP (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.*

I. 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.

Step 1 - Construction of a model

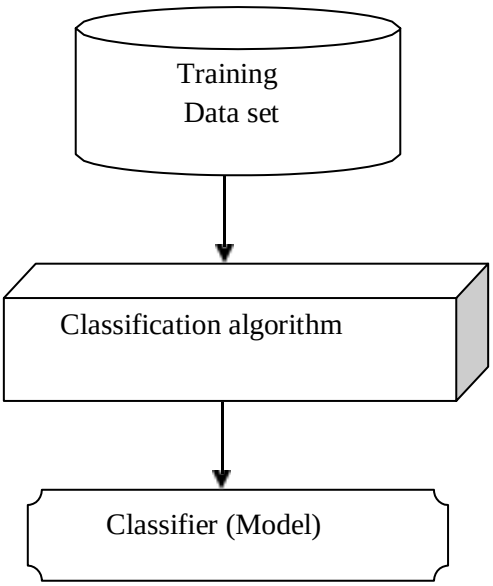

Fig.1-Model construction step

Step 2 - Model used for unknown tuple

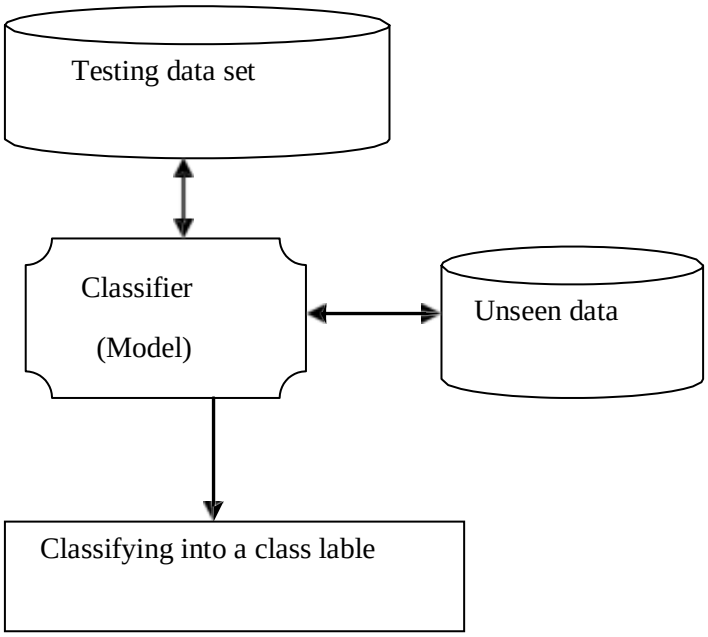

Fig.2 - Use of classifier

II. 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.

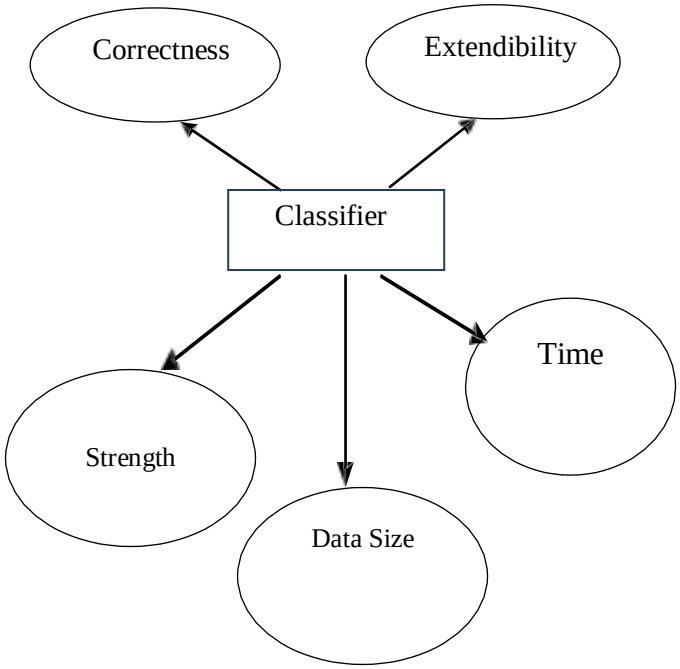

Fig.3 - Characteristis of a Classifier

III. LITERATURE SURVEY

Many investigations have been carried out to demonstrate the importance of the "Data Mining" techniques in education, demonstrating that this is a new concept for the purpose of extracting valid and accurate information about the behavior and effectiveness in the learning process. In the field of education techniques "Data Mining" has also been used to analyze the curriculum and subject of the current research topics, as well as to analyze the students performance.

There have been several investigations made under this proposed study object. For example, Bhardwaj used the Naïve Bayes algorithm to predict student performance based on 13 variables. The results were used to build a model that is used to predefine the students who are at risk of failure and thus activate a guidance and counseling program. Varghese, Tommy and Jacob in their research used the "K means" algorithm to cluster 8000 students based on five variables (input average in the University average scores of the tests / exams, average scores of papers, seminars notes and notes the work by frequency). The results showed a strong relationship between attendance and student performance.

Gulati and Sharma claim that knowledge through analysis by "Data Mining" can improve the education system in orientation, student performance and organizations management. Ayesha Mustafa [16] directed a study on evaluation, taking into account the evolution of learning and analysis of tests at the beginning and end of the courses. Bresfelean conducted a study based on students' results and how ease of these can be provided. Cortez and Silva conducted a research on the education system in Portugal and the results showed that a good and accurate prediction can be published a recent study applied to the entry requirements of the University of Saudi Arabia. They used algorithms and with techniques they have developed and a model that fits the public and the variables that describe it. They took into account input admission to the frequency of notes in previous education, admission notes and even the characteristics that describe the needs of the University. Some studies show the impact of the use of Moodle by applying Data Mining . Sun describes the different data mining techniques that can be applied to promote student learning on digital platforms

Aslam and Ashraf used clustering algorithm to provide a model of student learning. Some investigations discussed how data worked for Data improving the education system and enface knowledge in the

classroom. Vince Kellen in his case study, described the implementation of a structured analysis tool for Data Mining - SAP's HANA at the University of Kentucky, which estimates a value "k-score" for each student. This value will determine the involvement and subsequent guidance for good student performance.

IV. 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:

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

- **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.

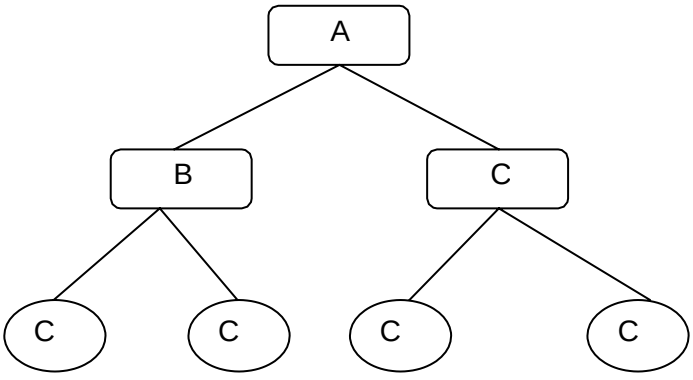

Fig.4- Decision Tree Classifiers

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

- **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,……,xn) and Y=(y1,y2,….,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.

- **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,.., An. . Suppose that there are m classes, C1, C2,…, 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<= j <= m, j≠ 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

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

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) ==P(Cm), and we would
therefore maximize P(X/Ci). Otherwise, we maximize P(X/Ci)P(Ci).

- **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.
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

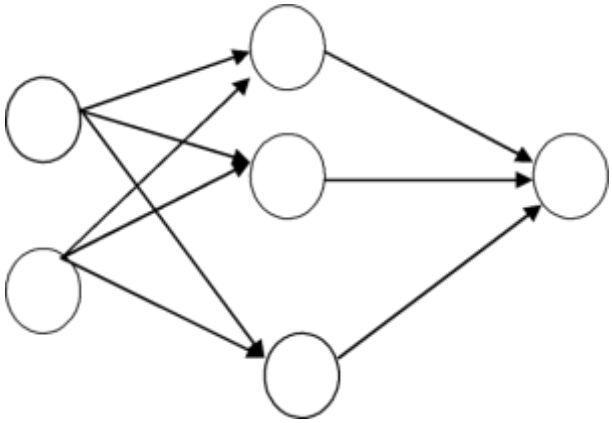

Fig. 5 - Neural networks as a classifier

- **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.

V. 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 Tre	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

VI. 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 than 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.

VII. 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 Nithyaand 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 September2015 inSciRes. 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