# TO STUDY DATA MINING CLASSIFICATION TECHNIQUES.

Saurabh Shewale, Prof. Babitha Kurup

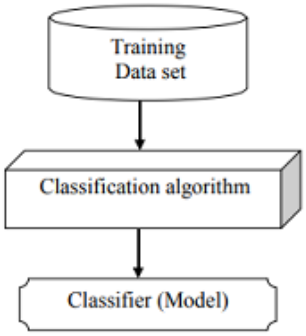
***Abstract-*** A Classification is one of the most convenient and beneficial techniques. Classification techniques are convenient to handle large amount of data. Classification is used to estimate unreserved class labels. Classification models are used to classifying freshly available data into a class label. Classification is the process of detecting a model that describes and differentiate data classes or concepts. Classification methods can handle both scientific and unreserved attributes. Constructing fast and accurate classifiers for large model used for unknown assembling Testing data set data sets is an important task in data mining and knowledge discovery. Classification guess 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 Neighbour, Support Vector Machines, Naive Bayesian Classifiers, and Neural Networks.

**Index Terms**— classification, prediction ,class label, model, categories..

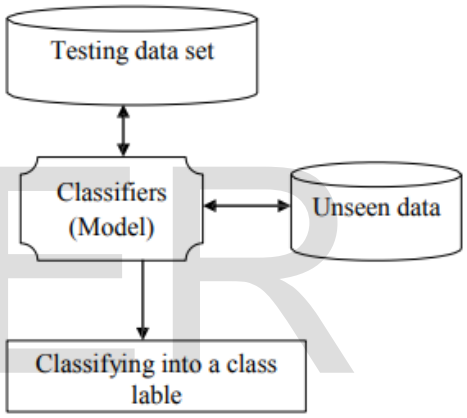
## INTRODUCTION

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 group into a class label.

Step 1 - Construction of a model




Step 2 - Model used for unknown group




ii. CHARACTERISTICS OF CLASSIFIERS:-

Each and every classifier has some quality which distinctives the classifier from other. The belongings or properties are known as characteristics of the classifiers. These characteristics are Correctness :- How a classifier classifies group exactly is based on these characteristics .To check exactly there are some algorithmic values based on number of group classify accurately and number of group classify wrong.

TIME :- HOW MUCH TIME IS NECESSARY TO CONSTRUCT THE MODEL? THIS ALSO INCLUDES THE TIME TO USE BY THE MODEL TO CLASSIFY THEN NUMBER OF GROUP (PREDICTION TIME). IN OTHER WORD THIS REFERS TO THE CALCULATION COSTS.

STRENGTH:- ABILITY TO CLASSIFY A GROUP ACCURATELY EVEN GROUP HAS A NOISE. NOISE CAN BE INACCURATE VALUE OR ABSENT VALUE.

DATA SIZE :- CLASSIFIERS SHOULD BE INDIVIDUALISTIC

FROM 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.

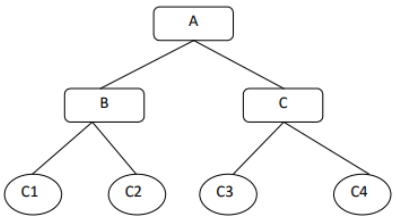
**Methods:-**

The main objective of a Classification algorithm are to maximize the predictive accuracy acquire by the classification model. Classification assignment can be seen as a supervised technique where each example belongs to a class. There are various l model techniques are used for classification some of them are[1 ,2 ,3]

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

.**Decision Tree:-**

A decision tree is a classifier and used as a recursive separation of the instance space. This model contains of nodes and a root. Nodes other than root have approximately one incoming edge. Intermediate node is a test nodes after executing a test they generate outgoing edge. Nodes without outgoing are called leaves (also called as terminal or decision nodes). In a decision tree, every single internal node splits the instance space into two or more than two subspaces at a certain discrete function of the input attributes values.




A INDICATE THE ROOT OF THE TREE. B, C ARE INTERNAL NODES INDICATE A TEST ON A PARTICULAR ATTRIBUTE AND C1, C2, C3 AND C4.

**K-Nearest neighbor :-**

This classifiers are found on learning by training examples. Each example describe a point in an n-dimensional space. All training examples are collected in an n-dimensional pattern space. When given an unknown example, a k-nearest neighbour classifier explore the pattern space for the k training examples that are closest to the unknown example. "Closeness" is also defined in terms of 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 neighbour classifiers assign similar weight to every attribute. Nearest neighbour classifiers can also be useful for prediction, that is, to return a real-valued prediction for a given unknown example.

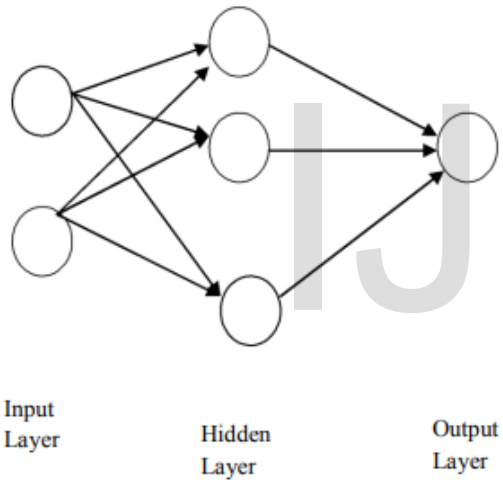
**. Bayesian classifiers:-**

Bayesian classifiers are analytical classifiers. They can estimate class integration based on probabilities. The Naive Bayes Classifier technique is suited when the dimensionality of the inputs is high. Naive Bayes can also outperform more experienced classification methods. Let D be a instruction set associated class labels. Each group is represented by an n-dimensional attributes, A1, A2,.., An. . consider that there are m classes, C1, C2,…, Cm. Given a group, 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 group 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 the P(Ci / X). The class Ci for which P(Ci / X) is maximized is called the maximum posterior hypothesis. By Bayes' theorem P(X) is constant for all classes, only P (X/Ci) P (Ci) needs to be maximized. If the class prior probabilities are not known, then it is generally assumed that the classes are equally likely, that is, P(C1) = P(C2) =……=P(Cm), and we would hence maximize P(X/Ci). Otherwise, we should maximize P(X/Ci)P(Ci).

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

**Neural Networks:-**

Neural Network used incline descent method based on biological nervous system having various interrelated processing elements. These elements are also known as neurons. Rules are released from the trained Neural Network to enhance compatibility of the learned network. To solve a specific problem NN used neurons which are organized processing elements.

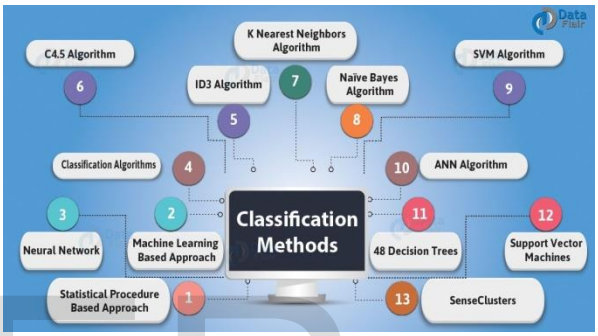



Neural Network is also used for classification and pattern recognition. An NN transform its structure and adjusts its weight in order to minimize the error. Adjustment of weight is based on the data that flows internally and externally through the network during learning phase. In NN multiclass, issues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

**Support Vector Machine (SVM):-**

SVM is a very successful method for regression, classification and general pattern recognition. It is considered a excellent classifier because of its high generalization performance without the need to add a prior knowledge, even when the dimension of the input space is very high. It is still considered as a good classifier because of its high generalization performance without the need to add a prior knowledge, even when the dimension of the input space is very high. For a linearly separable dataset, a linear classification function comparable to a separating hyper plane f(x) that proceed through the middle of the two classes, separating the two. SVMs were start with developed for binary classification but it could be efficiently extended for the multiclass problems.

**Results:-**




**CONCLUSION:-**

There are various 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 impatient learners they use training group to construct a generalization model.

Some of than are slow-moving learner like nearest-neighbour classifiers and case-based reasoning. These store training ples in pattern space and wait until presented with a test group before performing generalization.

**REFERENCES:-**

[1] 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

[2] 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. 241266http://dx.doi.org/10.14257/ijbsbt.2013.5.5.25 [3] 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.