International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 03 Issue: 06 | June -2019 ISSN: 2590-1892

A Study of Some Data Mining Classification Techniques

---***---

Ankita Das
PG Student,
Dept. of MCA
DSCE, Bangalore-78, India

Gulafsha BS
PG Student
Dept. of MCA
DSCE, Bangalore-78, India

Abstract - *An Classification is one the most helpful and significant 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 andaccurate 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 constructedbasedonsometrainingdataset,insecondsstep the model is used to classify a unknown tuple into a class label.

1.1 Step 1 - Construction of amodel

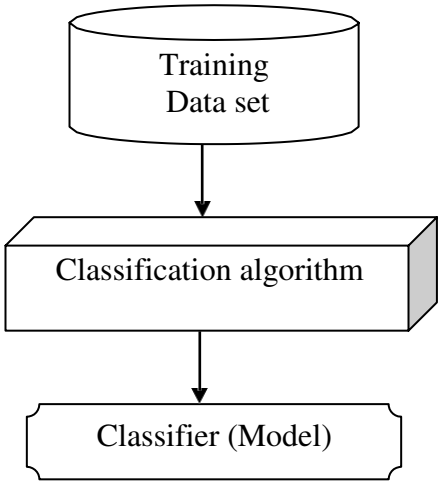

Fig.1 - Model construction step

1.2 Step 2 - Model used for unknowntuple

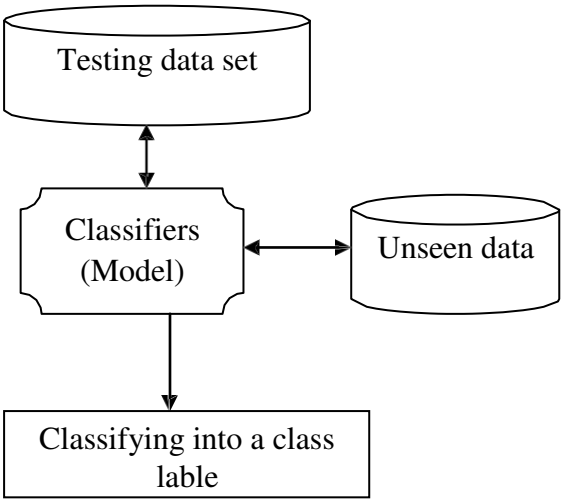

Fig.2 - Use of classifier

2. CHARACTERISTICS OF CLASSIFIERS

Eachandeveryclassifierhassomequalitywhichdifferential the classifier form other. The properties are known as characteristics of the classifiers. These characteristics are **Correctness:-**Howaclassifierclassifiestupleaccuratelyis basedonthesecharacteristics.Tocheckaccuracythereare some numerical values based on number of tuple classify correctlyandnumberoftupleclassifywrong.
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:-abilitytoclassifyatuplecorrectlyeventuplehas anoise.Noisecanbewrongvalueormissingvalue.
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.

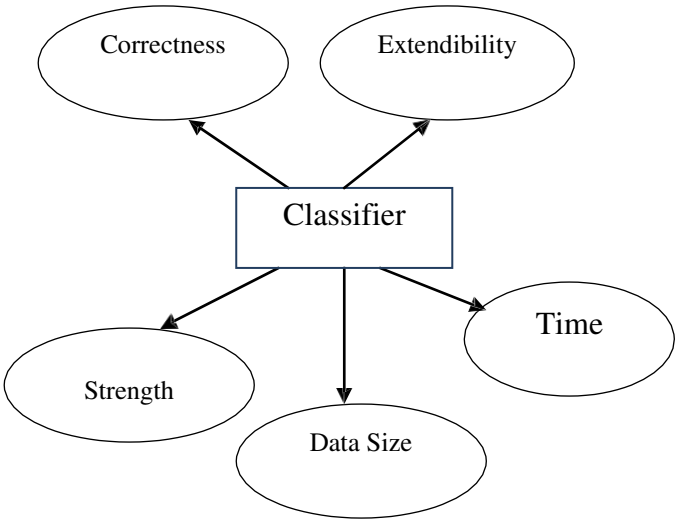

Fig.3 - Characteristis of a Classifier

3. LITERATURESURVEY

In 2012 Akhiljabbar et al. proposed "Heart Disease Prediction System using Associative Classification and Genetic Algorithm". They proposed efficient associative classification algorithm using genetic approach for heart diseaseprediction.Themainadvantageofgeneticalgorithm 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 AkhilJabbar 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 diagnosticinterface".Theyshowedthatearliermodelbased 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].

In2015S.OlalekanAkinola,O.JephtharOyabugbeproposed "AccuraciesandTrainingTimesofDataMiningClassification Algorithms: An Empirical Comparative Study". They proposedstudywasdesignedtodeterminehowdatamining classificationalgorithmperformwithincreaseininputdata sizes.Theyusedthreedataminingclassificationalgorithms 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].

In2015JaiminiMajali,RishikeshNiranjan&VinamraPhatak proposed "Data Mining Techniques for Diagnosis and Prognosis of Cancer". They used data mining techniques for diagnosisandprognosisofcancer.Theypresentedasystem 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 algorithmsappliedonTalentdatasettospotthetalentsetso as to judge the performance of the individual. Finally countingonaccuracyonebestsuitedclassifierischosenthis method has been used to construct classification rules to predictthepotentialtalentthatforpromotionornot[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 healthcaredatasetforanalyzingthehealthcaresystem.The percentage of accuracy of every applied data mining classification technique is used as a standard for performancemeasure.Thebesttechniqueforparticulardata set is chosen based on highestaccuracy[7].

4. VARIOUS CLASSIFICATIONMODEL

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].

- DecisionTree,
- K-NearestNeighbor,
- Support VectorMachines,
- Naive BayesianClassifiers,
- NeuralNetworks.

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.

Intermediatenodeistestnodesafterperformingatestthey 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 attributesvalues.

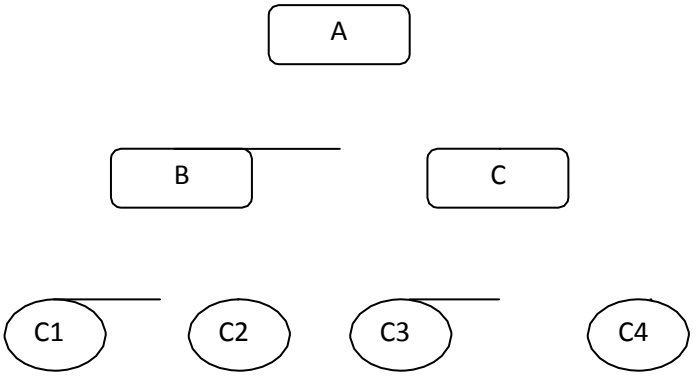

Fig.4- Decision Tree Classifiers

A denotes the root of the tree.B, Care internal nodes denoteatestonaparticularattributeandC1,C2,C3andC4.

4.2 K-Nearestneighbor

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.Whengivenanunknownsample,ak-nearestneighbor classifier searches the pattern space for the k training samplesthatareclosesttotheunknownsample."Closeness" is defined in terms of Euclidean distance, where the Euclidean distance, where the Euclidean distance between twopoints,X=(x1,x2,……,xn)andY=(y1,y2,….,yn)isdenoted 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 predictclassmembershipbasedonprobabilities.TheNaive Bayes Classifier technique is particularly suited when the dimensionality of the inputs is high. Naive Bayes can often outperformmoresophisticatedclassificationmethods.LetD be a training set associated class labels. Each tuple is represented by an n-dimensional attributes, A1, A2,.., An. . Supposethattherearemclasses,C1,C2,…,Cm.Givenatuple, X,theclassifierwillpredictthatXbelongstotheclasshaving the highest posterior probability, conditioned on X. That is, thenaïveBayesianclassifierpredictsthattuplexbelongsto theclassCiifandonlyifP(Ci/X)>P(Cj/X)for1<=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 wewould 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 processingelements.Theseelementsareknownasneurons. 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.

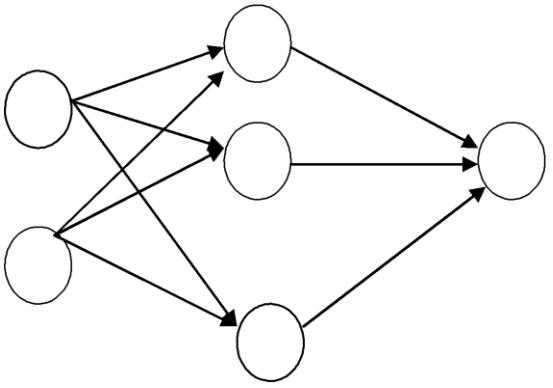

Fig. 5 - Neural networks as a classifier

Neural Network is used for classification and pattern recognition. An NN changes its structure and adjusts its weightinordertominimizetheerror.Adjustmentofweight 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 employedintheoutputlayerratherusingoneneuron

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,

evenwhenthedimensionoftheinputspaceisveryhigh.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. SVMswereinitiallydevelopedforbinaryclassificationbutit could be efficiently extended for multiclassproblems.

5. ADVANTAGE ANDDISADVANTAGE

Each and every model has some advantage and disadvantage. Wegivesomeadvantageanddisadvantageof thesemethods

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, classificationbybackpropagation,supportvectormachines, thesetechniquesareeagerlearnerstheyusetrainingtuples to construct a generalizationmodel.

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.

7. REFERENCES

[1] Data Mining: Concepts and Techniques 3rd Edition by Jain Pei, Jiawei Han, MichelineKamber
[2]]Introduction To Data Mining
[3] Data Mining: Practical Machine Learning Tools and TechniquesBy: Eibe Frank and Ian H. Witten
[4] Mining of Massive Datasets By: AnandRajaraman and Jeffrey Ullman
[5] Handbook of Statistical Analysis and Data Mining Applications By: Gary Miner, John Elder, and Robert Nisbet
[6] Discovering knowledge in dataBy: Daniel T. Larose
[7] Principles of Data Mining By: David J. Hand, HeikkiMannila, and Padhraic Smyth
[8] Data preparation for data mining By: Dorian Pyle
[9] Outlier Analysis By: Charu C. Aggarwal
[10] Practical Data Mining By: Monte Hancock