# A Review of Data Mining Classification Techniques

**Sunidhi Bansal[1] Dr. Kanwal Garg[2]**
[1]Scholar (M. Tech) [2]Supervisor (Assistant Professor)
[1,2]Depatrment of Computer Science and Application
[1,2]Kurukshetra university, Kurukshetra

Abstract— This paper provides a review of different data mining classification techniques such as decision tree, naive bayes, k-nearest neighbor, support vector machine with its favorable, unfavorable features and applications. This paper also describes some of the performance evolution measures that can be used for evaluating the performance of different classifiers.

***Key words:*** Decision Tree, KNN, Naive Bayes, SVM

## I. INTRODUCTION

Now-a-days large amount of data is collected and stored in databases. There is invaluable information and knowledge "hidden" in such databases. Data mining is done on large amount of data to extract the hidden information and then transform it into an understandable structure for further use. Data mining consists of various important techniques and classification is one of them.

Classification is one of the forms of data analysis that can be used to extract models describing important data classes. Such analysis can help provide us with a better understanding of the data at large [17].Classification is a supervised learning which can predict categorical class labels and classifies data based on training set and class labels and hence can be used for classifying newly available data [6]. Classification involves predicting an outcome based on a given input [9]. An example would be assigning a given email into "spam" or "non-spam" classes or identifying loan applicants as "safe" or "risky". Data classification has two steps. In first step, a classifier is built describing a predefined set of data classes or concepts. It is a leaning step. Second step is of using the classifier for classification. Here the test data is used to estimate the accuracy of classification rules. The classification rules can be applied to the new data tuples if the accuracy is considered acceptable.

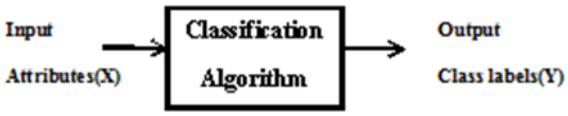


Fig. 1: Classification mapping attributes(X) into class label(Y)

### A. Issues Regarding Classification

Classification algorithms cannot be applied directly to raw data. Data requires some preprocessing for better classification results. Some preprocessing of data that may be required is given below.

### B. Data Cleaning:

This refers to the preprocessing of data. Data may contain noise or missing values this needs to be removed or reduced.

### C. Relevance Analysis:

Many attributes in data may be redundant. Relevance analysis, in the form of correlation analysis and attribute subset selection, can be used to detect the attributes that do not contribute in classification task [9].

### D. Data Transformation and Reduction:

Data can be transformed by normalization or by generalizing it to higher level concepts.

This paper is organized as follows. Section II explains the literature review, Section III describes the classification methods such decision tree, naïve bayes, KNN, SVM, Section 4 describes some of the performance evolution measures that can be used for evaluating the performance of different classifiers and last section concludes this work.

## II. LITERATURE REVIEW

Data classification is an important data analysis technique. Following are some of the researches which explore this field.

[1] Eduardo P. Costa et al. (2007) reviews the main evaluation metrics proposed in the literature to evaluate hierarchical classification models. Cristóbal Romero et al. (2008) [2] compare different data mining methods and techniques for classifying students based on their Moodle usage data and the final marks obtained in their respective courses and also developed a specific mining tool for making the configuration and execution of data mining. Muhammad Naufal Mansor et al.(2009) [3] presents an integrated system for detecting facial changes of patient in a hospital in Intensive Care Unit(ICU) and it also demonstrate that the k-NN can be used to classify the awakeness with the average accuracy of 94%.

Cuiping Leng et al.(2009)[4] applied Naive Bayes Classifiers to Incomplete Data and show that compared with the common methods dealing with missing data, this method is more efficient and reliable. Kaushik H. Raviya and Biren Gajjar (2013)[5] present the comparison between K-nearest neighbor, Bayesian network & Decision tree respectively on super market dataset and analysis shows that Bayesian algorithms have good accuracy over above compared algorithms. Delveen Luqman Abd AL-Nabi and Shereen Shukri Ahmed (2013)[6] present the comparison between the classification techniques which are K- Nearest Neighbor classifier, Decision tree and Bayesian network algorithms. It present the strength and accuracy of each algorithm for classification in term of performance efficiency and time complexity required.

Dr. Md. Ali Hussain et al. (2013)[7] present a new decision tree model based on multivariate statistical method Principal Component analysis on multi-attribute data for reducing dimensionality and to transform traditional decision tree algorithm to form a new algorithmic model. V. Vaithiyanathan et al. (2013) [8] present the comparison between J48, Multilayer Perception, Bayes N/w, Naïve Bayes Updatable on labour, soybean, weather dataset in terms of accuracy and time taken and result shows that

 

Naïve Bayes Updatable performed well with Labor dataset and Multilayer perception with Soybean dataset and weather dataset in terms of accuracy. N. Abirami et al. (2013) [9] present the analysis of several data mining classification techniques using WEKA machine learning tools over the healthcare datasets. The standards used for comparison are percentage of accuracy and error rate.

K. Wisaeng (2013) [10] present the comparison of different classification techniques in open source data mining software which consists of a decision tree methods and machine learning for a set of bank direct marketing dataset. Ahmad Ashari (2013) [11] Performed a performance comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool. S.Archana and Dr. K.Elangovan (2014) [12] provide an inclusive survey of different classification algorithms such as C4.5, k-nearest neighbor classifier, Naive Bayes, SVM, and ID3 and their advantages and disadvantages.

Vaibhav P. Vasani and Rajendra D. Gawali (2014) [13] present classification of the data collected from students of polytechnic institute and also compares the results of Decision Tree and Naïve Bayesian Algorithm with respect to different performance parameters. M. Soundarya, R.Balakrishnan (2014) [14] present the survey of the several classification techniques of classification methods such as decision tree induction, Bayesian networks, k-nearest neighbor classifier and fuzzy logic techniques. T. Revathi and P. Sumathi (2014) [15] analyzed data mining classification techniques Decision Tree and Support Vector Machine (SVM) on Aortic Stenosis disease dataset. Analysis shows that SVM predicts Aortic Stenosis disease with least error rate and highest accuracy. S.Vijayarani et al. (2015) [16] analyses the performance of classification techniques such as BayesNet, Naïve Bayes, IBK, Kstar for hepatitis and thyroid dataset and concluded that Naïve Bayes has better results on hepatitis, BayesNet classifier gives best accuracy for thyroid dataset.

## III. CLASSIFICATION METHODS

There are many classification techniques some of them are described below.

### A. *Decision Tree Algorithm:*

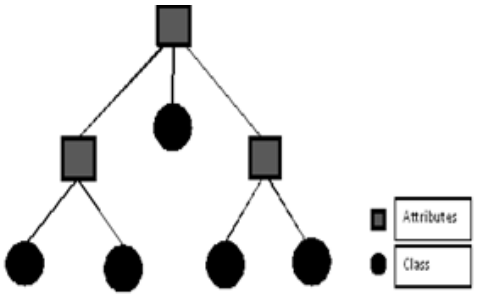


Fig. 2: Decision Tree

Decision trees are one kind of inductive learning algorithms that offer an efficient and practical method for generalizing classification rules from previous concrete examples. Most decision tree classifiers perform classification in two phases: tree-growing (or building) and tree-pruning. The tree building is done in top-down manner [5]. In decision tree induction the entire data in the training set is used as root node for the tree. Then the root node is split into several sub-nodes depending upon some splitting criterion. The process of splitting sub-node continues, till all leaf nodes are generated else if all the instances in the sub-node belong to the same class [7]. A decision tree can easily be converted to a set of classification rules [2].

*1) Constructing Set of Rules from Decision Tree:*

1) Represent the Decision tree in form of IF-THEN rule.
2) One rule is created for each path from the root to a leaf.
3) Leaf represents the class label.

*2) Example:*

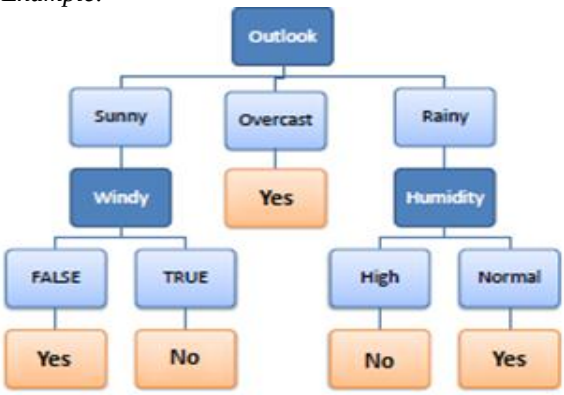


Fig. 3: Decision tree [19] resides on 15 April 2015.

- IF Outlook=”Sunny” AND Windy=”False” THEN Play=”Yes”
- IF Outlook=”Sunny” AND Windy=”True” THEN Play=”No”
- IF Outlook=”Overcast” THEN Play=”Yes”
- IF Outlook=”Rainy” AND Humidity=”High” THEN Play=”No”
- IF Outlook=”Rainy” AND Humidity=”Normal” THEN Play=”Yes”

### B. *Naive Bayes:*

The Naïve Bayes is a simple probabilistic classifier [9].It assumes that the effect of an attribute value on given class is independent of the values of the other attributes. This assumption is called class conditional independence.
It has the star-like structure [4] shown in figure 4

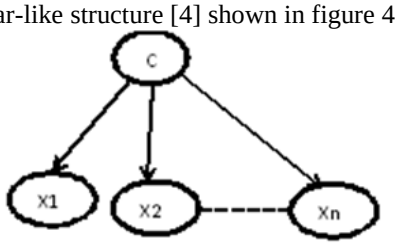


Fig. 4: The structure of Naïve Bayes Classifier

In figure 4 X1…Xn and C denotes attribute variables and class variable respectively.

The probabilities applied in the Naïve Bayes algorithm are calculated using Bayes Rule the probability of hypothesis H can be calculated on the basis of the hypothesis H and evidence about the hypothesis X according to the following formula

$$P(H|X)=\frac{P(X|H)P(H)}{P(X)} \quad (1)$$

 

*1) Example:*

- X: 35 year old customer with an income of $40,000 and fair credit rating.
- H: Hypothesis that customer will buy a computer.

*C. K-Nearest Neighbor Algorithm:*

KNN a non-parametric lazy algorithm called as "Closest Point Search" is a mechanism that is used to identify the unknown data point based on the nearest neighbor whose value is already known by comparing given test tuple with training tuples that are similar to it. When given an unknown tuple, a k-nearest-neighbor classifier searches the pattern space for the k training tuples that are closest to the unknown tuple as shown in figure 5 [19]. These training tuples are the k "nearest neighbor" of the unknown tuple [17].

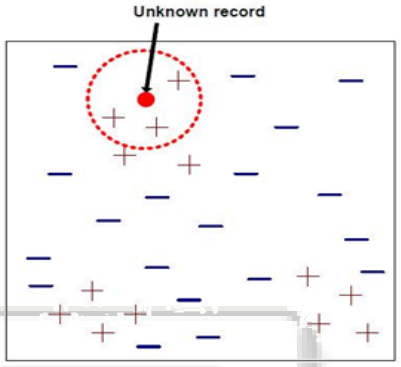


Fig. 5: K-Nearest Neighbor

A Euclidean Distance measure is used to calculate how close each member of the training set is to the test class that is being examined [2]. Say, X=(x1,x2,…..,xn) and Y=(y1,y2,……..yn), is

$$d(x,y) = \sqrt{\sum_{i}^{n} (x_i - y_i)^2} \quad (2)$$

The k-nearest neighbors' algorithm is among the simplest of all machine learning algorithms. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its *k* nearest neighbors. *K* is a positive integer, typically small [14].

*D. Support Vector Machine:*

Support Vector Machines are supervised learning methods used for classification, as well as regression. The support vector machine usually deals with pattern classification that means this algorithm is used mostly for classifying the different types of patterns [12]. The basic support vector machine takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier [10]. The advantage of Support Vector Machines is that they can make use of certain kernels in order to transform the problem, such that we can apply linear classification techniques to non-linear data. Applying the kernel equations arranges the data instances in such a way within the multi-dimensional space, that there is a hyper-plane that separates data instances of one kind from those of another.

The kernel equations may be any function that transforms the linearly non-separable data in one domain into another domain where the instances become linearly separable. Kernel equations may be linear, quadratic, Gaussian, or anything else that achieves this particular purpose.

Once we manage to divide the data into two distinct categories, our aim is to get the best hyper-plane to separate the two types of instances. This hyper-plane is important because it decides the target variable value for future predictions. We should decide upon a hyper-plane that maximizes the margin between the support vectors on either side of the plane. Support vectors are those instances that are either on the separating planes on each side, or a little on the wrong side [18]. The following figure 6 [20] illustrates these definitions, with + indicating data points of type 1, and – indicating data points of type –1.

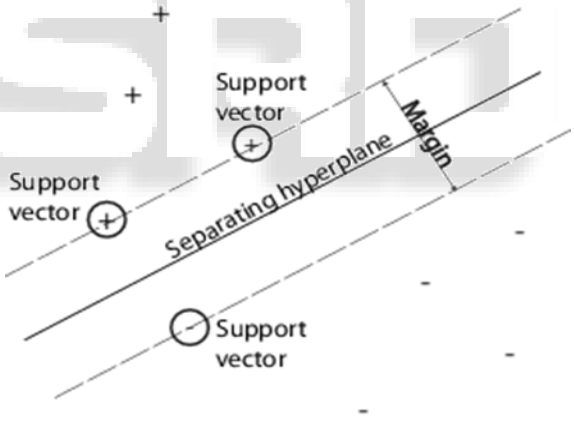


Fig. 6: Support Vector Machine

Advantages, disadvantages and application of various classification techniques is shown in Table 2.

| Classifier | Fundamentals | Working | Favorable features | Unfavorable features | Applications |
|---|---|---|---|---|---|
| Decision Tree | Decision trees are one kind of inductive learning algorithms. It can be converted to a set of classification rules. | It uses greedy recursive algorithm to partition the data until all the data items belong to a particular class is identified. | -Simple to understand and interpret.<br><br>-Can determine worst, best and expected values for different scenarios. | -They easily overfit that means they generally needs pruning.<br><br>-Splitting a lot leads to complex trees. | Operation Research specially in decision analysis, Teaching, Research area etc. |

 

| | | | | | |
|---|---|---|---|---|---|
| Naïve Bayes | Naïve Bayes classifiers are probabilistic classifier based on Bayesian theorem with independence assumptions. | It computes the conditional probability distribution of each feature given label, and then it applies Bayes' theorem to compute the conditional probability distribution of label given an observation and use it for prediction. | -Easy to implement.<br><br>-Require small amount of training data to estimate the parameters.<br><br>-Less pruning required. | -Assumption: Class condition independence, therefore loss of accuracy.<br><br>-Practically dependencies exist among variables. | Medical diagnosis, Document classification, image processing etc. |
| KNN | KNN is a non-parametric lazy algorithm which is best known distance based algorithm. | When given an unknown tuple, a k-nearest-neighbor classifier searches the pattern space for the k training tuples that are closest to the unknown tuple. | -Robust to noisy data.<br><br>-Effective with large training data. | -Value of parameter k (no. of nearest neighbor) needs to be determined.<br>-Computation cost is high. | Pattern Recognition, Internet Marketing, Cluster analysis etc. |
| SVM | Support Vector Machines are based on the concept of decision planes that define decision boundaries. | A support vector machine constructs a hyper plane or set of hyper planes in a high or infinite-dimensional space, which can be used for classification, regression, or other tasks. | -Its prediction accuracy is generally high.<br><br>-It is robust and works well when training examples contain errors. | -It has long training time.<br><br>-Difficult to understand the learned function (weights). | Image classification, Medical science, Text and Hypertext categorization etc. |

Table 1: Comparison of Various Classification Techniques

## IV. Performance Evaluation Measures

Evaluation measures in classification problems are defined from a matrix with the numbers of examples correctly and incorrectly classified for each class, named confusion matrix. The confusion matrix for a binary classification problem (which has only two classes - positive and negative)[1] is shown in Table3

| | **Predicted class** | |
|---|---|---|
| Actual class | Positive | Negative |
| Positive | TP | FN |
| Negative | FP | TN |

Table 3: Confusion Matrix

- True positives (TP): When the outcome is correctly classified as positive when it is positive.
- True negatives (TN): When the outcome is correctly classified as negative when it is negative.
- False positives (FP): When the outcome is incorrectly classified as positive when it is in fact negative.
- False negatives (FN): When the outcome is incorrectly classified as negative when it is positive.

The evaluation measure most used in practice is the accuracy. It evaluates the effectiveness of the classifier by its percentage of correct predictions.

$$Accuracy = (TP+TN)/TP+TN+FP+FN$$

The complement of Accuracy is the error rate, which evaluates a classifier by its percentage of incorrect predictions.

$$Error = 1-Accuracy$$

Recall(R) is the proportion of actual positive cases which are correctly identified.

$$R = TP/TP+FN$$

Specificity(spe) is the proportion of actual negative cases which are correctly identified.

$$Spe = TN/FP+TN$$

Precision(P) is the proportion of positive cases that were correctly identified.

$$P = TP/TP+FP$$

F-measure(FM): F-measure is harmonic mean of precision and recall.

$$FM=2*(precision*recall)/(precison+recall)$$

## V. Conclusion

This paper covers various classification techniques used in data mining. Survey based on previous researches shows that KNN accuracy is better when applied to small dataset but its accuracy decreases when dataset is large. SVM is less time efficient but its accuracy does not depend on dataset size. Decision tree and naïve bayes are more time efficient algorithms as compared to KNN and SVM. As there are many classification techniques so we have to choose most appropriate technique based on needed condition to mine the data.

## References

[1] Eduardo P. Costa, Ana C. Lorena, Andre C. P. L. F. Carvalho and Alex A. Freitas, "A Review of Performance Evaluation Measures for Hierarchical Classifiers", 2007.
[2] Cristóbal Romero, Sebastián Ventura, Pedro G. Espejo and César Hervás," Data Mining Algorithms to Classify Students".

 

[3] Muhammad Naufal Mansor, Sazali Yaacob, R. Nagarajan, Lim Sin Che,M.Hariharan and Muhd Ezanuddin," Detection of Facial Changes for ICU Patients Using KNN Classifier Muhammad", IEEE, 2009, pp.1-5.
[4] Cuiping Leng, Shuangcheng Wang, Hui Wang," Learning Naive Bayes Classifiers with Incomplete Data", IEEE, International Conference on Artificial Intelligence and Computational Intelligence, 2009, pp.351-353
[5] Kaushik H. Raviya and Biren Gajjar ," Performance Evaluation of Different Data Mining Classification Algorithm Using WEKA" Indian Journal of Research, Volume 2,Issue 1 January 2013
[6] Delveen Luqman Abd AL-Nabi and Shereen Shukri Ahmed," Survey on Classification Algorithms for Data Mining:(Comparison and Evaluation)" Vol.4, No.8, 2013,pp.18-24.
[7] Dr. Md. Ali Hussain, Dr. M. Kameswara Rao and Dr. Ali Mirza Mahmood, "An Optimized Approach To Generate Simplified Decision Trees" IEEE International Conference on Computational Intelligence and Computing Research,2013.
[8] V. Vaithiyanathan, K. Rajeswari, Kapil Tajane and Rahul Pitale, " Comparison of Different Classification Techniques using Different Datasets", International Journal of Advances in Engineering & Technology, Vol. 6, Issue 2, May 2013, pp. 764-768.
[9] N. Abirami, T. Kamalakannan and Dr. A. Muthukumaravel ," A Study on Analysis of Various Data mining Classification Techniques on Healthcare Data" International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 7, July 2013,pp.604-607.
[10] K. Wisaeng,"A Comparison of Different Classification Techniques for Bank Direct Marketing", International Journal of Soft Computing and Engineering, Volume-3, Issue-4, September 2013,pp.116-119.
[11] Ahmad Ashari, Iman Paryudi and A Min Tjoa, "Performance Comparison between Naïve Bayes, Decision Tree and k-Nearest Neighbor in Searching Alternative Design in an Energy Simulation Tool", International Journal of Advanced Computer Science and Applications, Vol. 4, No. 11, 2013.
[12] S. Archana and Dr. K. Elangovan, "Survey of Classification Techniques in Data Mining", International Journal of Computer Science and Mobile Applications, Vol.2 Issue. 2, February-2014, pg. 65-71.
[13] Vaibhav P. Vasani and Rajendra D. Gawali, "Classification and performance evaluation using data mining algorithms", International Journal of Innovative Research in Science, Engineering and Technology Vol. 3, Issue 3, March 2014.
[14] M.Soundarya, R.Balakrishnan," Survey on Classification Techniques in Data mining", International Journal of Advanced Research in Computer and Communication vEngineering,Vol. 3, Issue 7, July 2014
[15] T. Revathi and P. Sumathi, "An Overview of Data Mining Classification Methods in Aortic Stenosis Prediction", International Journal of Engineering and Advanced Technology (IJEAT), Volume-3 Issue-6, August 2014.
[16] S.Vijayarani, R.Janani and S.Sharmila," Data Mining Classification Algorithms for Hepatitis and Thyroid Data Set Analysis", International Conference on Computing and Intelligence Systems, Volume: 04, Special Issue: March 2015, pp. 1270-1275.
[17] Han and M. Kamber, "Data Mining; Concepts and Techniques", Morgan Kaufmann Publishers, Second edition, 2009.
[18] "Classification Methods," (Last visited in May 2015). [Online]. Available: http://www.d.umn.edu/~padhy005/Chapter.html
[19] "K-Nearest Neighbors," (Last visited in May 2015). [online]. Available: http://mines.humanoriented.com/classes/2010/fall/csci568/portfolio_exports/lguo/knn.html
[20] "Support Vector Machine," (Last visited in May 2015). [online]. Available: http://in.mathworks.com/help/stats/support-vector-machines-svm.html

 