International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-04, Issue-02, May 2018

Comparative analysis of classifier algorithm in data mining

Aikjot Kaur Narula#, Dr.Raman Maini*

#Student, Department of Computer Engineering, Punjabi university Patiala, India, aikjotnarula@gmail.com

*** Professor, Department of Computer Engineering, Punjabi university Patiala, India, research_raman@yahoo.com**

Abstract - Data mining is a process of extracting data from the large set of data sets. **It is used to choose the data by specifying a pattern to solve the problem**. **D**ata mining is a **technique which** is currently used in almost every field such as science, education, technology**,**and research**,**etc. in order to draw the concrete conclusions. classification is a method of data mining which is used for predicition In this paper, there will be a comperative analysis of classifiers. Moreover, the main focus is on Four **data mining c**lassifires namely, Decision tree,**Naïve Bayes** , kNN and SVM.**from comparative analysis it has been observed that decision tree algorithm is more used to solve the long problems because it splits the attributes of the problem and then finds the threshold value for splitting which is helpful in constructing trees to solve the problems more accurately**. **The kNN algorithm calculates the data more efficiently because it uses the value of nearest attributes**.**Whereas the SVM algorithm is most widely used to design the hyperplane**. **These hyper plane are used to find a value that gave the best result**

Key words :- k nearest neighbour , support vector machine.

I. INTRODUCTION

Data mining is used by the companies to extract the useful information from the raw data. For example, the group collects the similar documents returned by the search engine according to context as Amazon rainforest, Amazon.com. There is various classification algorithms used in data mining. Classification of data is an important section of analysis which is done using any of the data mining techniques. The classification technique can be explained as defining the data that in which category it falls for obtaining the results on category based data classification. The classification goal is to identify the category of a new observation which needs to be compared with the present list of observatory data or training data. [3]

The classification is generally donein order to identify the patterns followed in a set of data to understand the future work or observations which are similar to the data already classified. The classification technique made the entire process of data mining to be very easy. For easy classification of data, the implementation of a classification is needed to be donein a concrete manner,and for completing the classification work, different algorithms are required which are termed as classifiers.

Classification technique apply on discrete value not on continuous value or floating point value classification model may have multiclass types where it have more than two value .In order to find the relationship between the values of target class and values of predictor in multiclass type various algorithm uses various techniques whereas the simple type of classification model is binary classification which have only two values The commonly used methods for data mining classification tasks can be categorized as follows

1. Decision tree
2.K Nearest Neighbour
3.Naïve Bayes
4.K Nearest Neighbour
5. Support vector machines

This paper is distributed into 4 section the foremost part explain introduction. Moving further, the second section and third section provide information about classifiers algorithm and comparative analysis of various classifiers algorithm however, the last segment contain the conclusion of all four classifiers

II. CLASSIFIERS ALGORITHMS

The various classifier algorithm are as follows

2.1 Decision Tree:-A decision tree is like a flowchart used to classify the data. An open source data visualization and analysis tool for data mining using this type of algorithms in their decision tree classifiers. This is also known as supervised learning algorithm as it requires the set of training data in the form of pairs. These pairs include the input data and the required output values. The input data can be referred as an object and the output values as a class. This algorithm provides the complete analysis of the training set and builds the classifier which has the capacity

 DOI : 10.18231/2454-9150.2018.0174

to arrange both the test and training data in the accurate form. The algorithm which is used for generating decision tree areC4.5 and ID3 While generating the decision tree, the algorithm takes help of information gain. The single pass pruning process is used in theC4.5algorithms, which helps in mitigating the overfitting. This algorithm works for both the discrete as well as continuous data. C 4.5 is an extension of Iterative Dichotomiser 3 (ID3). It is a statistical classifier that's why it is used in decision-making tree. [7]

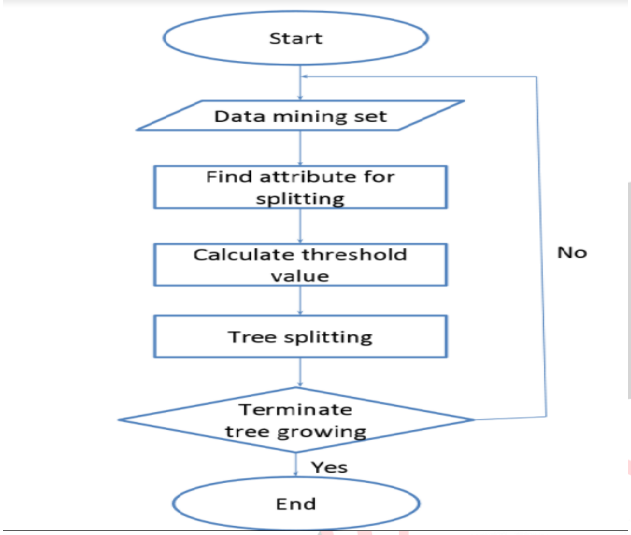

Figure 1: Flowchart of C4.5 Algorithm

It works on the concept of the entropy in machine learning. This classifier is a decision tree that is built from the root to leaves under the recommendations of Occam's razor. It uses the p-dimensional vectors to classify the data sets and samples, and define the attributes of samples. For classifying theunrecognized data, the present dataset is used. The dataset is then further divided into the attributes and instances. Then the data execution enables to take a decision that which action must be performed in any particular condition for getting best outcomes. The Weka windows consist of several classifiers, for example, bays, lazy, functions, and meta. it is a greedy algorithm These classifiers are used to construct the trees using recursive ,top down ,divide and conqure manner

***2.2 Naïve Bayes*:-**Naïve Bayes is from the family of classification that shares a common assumption. In this algorithm, features of data areindependent of all other features given in the class. Two features of data said to be independent when the value of one feature does not impact on the value of another feature. Due to all independent assumption, it is called Naïve, and Bayes comes from the Thomas Bayes, who discovered this algorithm. For example, a fruit may be considered to be an apple if it is red, round. Even if these features depend on each other or upon the existence of other features of a class, a naive Bayes classifier considers all of these properties to independently contribute to the probability that this fruit is an apple. We can make predictions using Naïve Bayes Theorem.[2]

$$P(c \mid x) = (P(x \mid c) * P(c) / P(x) \quad \ldots\ldots..(1)$$

In Which:-

P(c|x) in which hypothesis is c and x is given data . This is called as the posterior probability.

P(c|x) c is the probability of given data that the hypothesis x was true.

P(c) is used for hypothesis prediction p(c) being true (regardless of the data). This is called the prior probability of c.

P(x) is used for data prediction (regardless of the hypothesis).

2.3 K Nearest Neighbour:-*K Nearest Neighbour is also known as* k-NN. It comes under the category of instance-based learning, or lazy learning, where the function is locally approximated and all computation is carried out until classification is done. The k-NN algorithm is amongst the simplest of all machine learning algorithms. It is instance learner because it does not do much in training. It works only when a new unlabeled input looks to classify the type of learner. This algorithm is used in many applications like pattern recognition and statistical estimation. This method is the non-parametric method of classification. [7] .The two types of KNN are

a) *Structureless NN technique*

In this technique, the complete data is classified into two samples of test sample and training sample. Then, the distance is evaluated from the training point to the sample point. The point which has the lowest distanceis called as the nearest neighbor point.

b) *Structure-based NN technique*

The structure-basedtechnologyis based on the various structure of data. The different types of the data structure include the nearest future line, k-d tree, OST (orthogonal structure tree), central line and axis tree.

kNN algorithm can be used in both classifier and regression predictive problems. For evaluating a problem usingkNN, three main aspects are considered, which includes ease to interpret output, calculation time and predictive power. It is better implemented on an unclassified dataset example by looking at k that is already classified as a neighbor and find out the class in which the maximum of parameters lies. After that, it chooses the nearest neighbor to the unknown item. These neighbors are most similar to the item that is placed by mapping the item. There are more chances that all the neighbors may belong to the same class.

K-NN algorithm is also known as supervised learning algorithm.k-NN algorithms are very easy to understand when there are few predictor variables When we are using a k-nearest neighbour algorithm, we have to choose appropriate value K is important because If the value of

 DOI : 10.18231/2454-9150.2018.0174

K is too small it is susceptible to over fitting and not able to classify data correctly.[7]

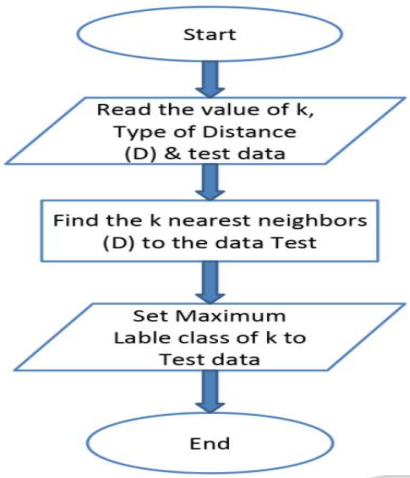

Figure 2: Flowchart of kNN Algorithm

2.4 Support Vector Machine:- support vector machine is abbreviated as SVM. It is another algorithm which is used for the classification.This algorithm is used to classify the data into two classes. SVM almost works similar to that of theC4.5algorithm, but the SVM does not use the decision trees for the classification. The SVM is based on the hyperplane function. This function works same as that of equation of a line,

$$y = mx + c \quad \ldots\ldots\ldots\ldots (2)$$

Support Vector Machine (SVM) is mostly used as the binary sets that are linearly separated to each other. If we want to classify, two variables say A & B to design a hyperplane by using SVM algorithm. For this, we must design the class square and class circles. After that, we classify all the vectors into these classes and make two classes, two designs, and two hyperplanes.

All the elements or instances in the classes are properly classified using the hyperplanes concept to identify the right decision, classification of the unknown information. or optimum sepration of classes

Hyperplane is categorizes in 4 types first if the hyperplane is one dimensional , an hyperplane is called point on the other hand if it two or three dimension it is called line and plane respectively. And the last type is when it is more than three dimensions then it is called as hyperplane.

SVM is a supervised learning algorithm SVM is called as surprised learning because in training phase SVM needs data that has been already classified it is basically used for classification and finding relationship between variable SVM algorithm based upon two things one is statically learning theory and other is structure risk minimization From the last decade, the SVM algorithm is applied in various domains. It is used for real-world problems such as classification of text and image, recognition of hand-writing, and bioinformatics and bio sequence analysis.[7]

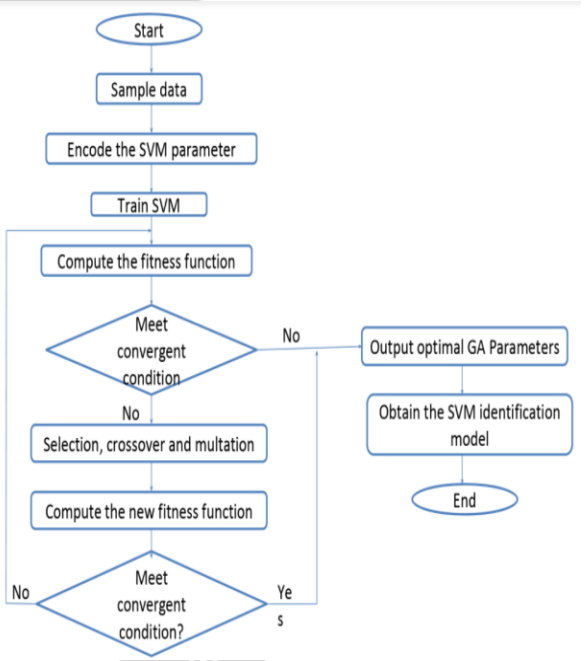

Figure 3: Flowchart of SVM Algorithm

III. COMPARATIVE ANALYSIS OF ALGORITHMS [1-11]

Algorithm	Advantages	Disadvantages	Applications
Decision Tree [14][7]	• It builds easily interpretable models which are easy to implement with noise removal. • This algorithm can be used for both types of values.categorical and continuous. • Reboust and fast • its works easy on non linear data • greedy search	• Variation in data even in minimal amount changes the entire values and it does not work well on small sets. • The smallest amount of variation in the data or information can lead to change in the decision trees. • There may be chance that further sub tree contain some dublicate values and it does not work well when there are many sub classes	• C4.5 is mostly used developed credit card companies as it helps in prediction of users or customers who will most likely take the benefit of life insurance policies. So, In this way they can mail new policies to potential users. • It predicts the performance of students who have chances to get failed in examination.

 DOI : 10.18231/2454-9150.2018.0174

KNN[4] [7]	• This algorithm is very robust with lesser noised training data • KNN is equally effective on large datasets for establishing the results related to data mining. • It is very simple to implement and easily handles multi class case	• Unclear distance-based learning and determination are required for identifying the K value. • The computation cost is high. • kNN is not effective on the learning techniques which are based on distance as it does not understand the distance needed to achieve the best outcomes. • This algorithm works on lazy learner technique	• Classify the reasons of Heart Disease. • Text categorization. • Feature extraction. • Use for prediction of Breast cancer.
SVM [7]	• This algorithm is accurate. • Work better in small datasets. • It is more advantageous as it only uses subset of training points • It works better on when data is noisy and when input is non monotonous and non linearly separable form	• This algorithm is not good when datasets is large as training with SVMs is high. • SVM is Less Effective when datasets has noise with overlapping classes. • Doing computation with svm is very expensive.	• It is used for Recognition of handwriting. • Classification of Images. • It is used for Detection of face. • Bioinformatics.
Naïve Bayes [2][7]	• It can be easily implemented with a small amount of training data with good results in maximum cases. • It is developed on very easy concepts which make the use of an algorithm to be quite easy. • It require small training data set • It easily handles both continuous and discrete value as well as it gave better result when it implemented in practical	• The implementation of this algorithm brings a loss in accuracy with dependency among variables. • This algorithm provides a very low level of accuracy and preciseness when applied toafewdatasets or information. • it faces major problem ie. zero conditional probability problem	• It is mainly used in text classification as it is better than other algorithms and gives more success ratio. • It is mostly used in filtering spam. For instance: Identifying spam e-mails.

Table1 : Comparative analysis of different Classifier algorithms

From the comparative analysis we found that decision tree works well on non linear data and is fast and robust. Naïve's bayes algorithms works well on high input dimensionality and is mainly used in text classification. Svm works on the concept of hyper plane. Svm works better for small sets of data where as its performance decreases for large and noisy data sets. The knn works well for both small and large datasets but it is computational expensive.

IV. CONCLUSION

The aim of these algorithms discussed in this paper is to provide insight into these algorithms so as to know their strength and weaknesses. This research focuses on comparative analysis of naïve bayes , decision tree ,KNN, and SVM algorithms . From the comparative analysis we found that KNN algorithm works effectively for small and large datasets where as it faces problem regarding distance and is not cost effective as it involves higher cost than other algorithms. Decision tree does not require any knowledge of domain and their interpretation is very easy. But the decision tree does not work well if there are even small changes in the data set. The SVM algorithm helps in solving problem related constrained quadratic optimization, making use of set of non linear functions we can map input on high dimension as SVM learns from different depictions like neutral nets, polynomials estimators etc and even provide with us optimal solutions that are unique in nature. Naïve's bayes algorithm implementation is very simple and is computationally efficient but requires large datasets for good results. All these above algorithms help in determining data and arranged or can grouped when new data source are available this paper enlighten us with features of techniques used along with their limitation on the basis of the particular features researcher can select appropriate technique

REFERENCES

[1] D. Kumar, "Performance Analysis of Various Data Mining Algorithms: A Review," Methods, vol. 32, no. 6, pp9–15, 1995.

 DOI : 10.18231/2454-9150.2018.0174

[2] S. Vijayarani and M. Muthulakshmi, "Comparative Analysis of Bayes and Lazy Classification Algorithms," International Journal of Advance Research in Computer and Communication Engineering, vol. 2, no. 8, pp. 3118–3124, 2013.
[3] S. S. Nikam, "A Comparative Study of Classification Techniques in Data Mining Algorithms," Oriental Journal of Computer Science and Technology, vol. 8, no. 1, pp. 13–19, 2015.
[4] A. Joshi and R. Kaur, "A Review: Comparative Study of Various Clustering Techniques in Data Mining," International Journal of Advanced research in Computer Science and Software Engineering, vol. 3, no. 3, pp. 2277–128, 2013.
[5] W. Becari, L. Ruiz, B. G. P. Evaristo, and F. J. Ramirez-Fernandez, "Comparative analysis of classification algorithms on tactile sensors," 2016 IEEE International Symosiump onConsumer Electronics, vol. 9, no. July, pp. 1–2, 2016.
[6] S. Gupta, D. Kumar, and A. Sharma, "Performance Analysis of Various Data Mining Classification Techniques on Healthcare Data," International journal of Computer Science and Technology, vol. 3, no. 4, pp. 155–169, 2011.
[7] G. Kesavaraj and S. Sukumaran, "A study on classification techniques in data mining," 2013 Fourth International Conference on Computer and Communication Network Technology , pp. 1–7, 2013.
[8] M. Gera and S. Goel, "Data Mining - Techniques, Methods and Algorithms: A Review on Tools and their Validity," International Journal of Computer science and Application, vol. 113, no. 18, pp. 22–29, 2015.
[9] K. S. Rawat, "Comparative Analysis of Data Mining Techniques , Tools and Machine Learning Algorithms for Efficient Data Analytics Keshav Singh Rawat," vol. 19, no. 4, pp. 56–60, 2017.
[10] R. Arora and Suman, "Comparative Analysis of Classification Algorithms on Different Datasets using WEKA," International Journal of Computer science and Application, vol. 54, no. 13, pp. 21–25, 2012.
[11] R. Arora and Suman, "Comparative Analysis of Classification Algorithms on Different Datasets using WEKA," International Journal of Computer science and Application, vol. 54, no. 13, pp. 21–25, 2012.
[12] M. Kumar and Rajesh "Predicting Upcoming Students Performance using Mining Technique," International Journal of Modern Trends in Engineering & Research, vol. 4, no. 7, pp. 38–44, 2017
[13] D. Kabakchieva, "Predicting Student Performance by Using Data Mining Methods for Classification," Cybernetics and Information Technologies, vol. 13, no. 1, Jan. 2013
[14] A. B. Raut and A. A. Nichat,"Students Performance Prediction Using Decision Tree Technique," International Journal of Computational Intelligence Research,Vol.13, no.7, pp. 1735-1741, 2017.

 DOI : 10.18231/2454-9150.2018.0174