Volume 4, Issue 4, April 2019 ISSN (online): 2456-0006

International Journal of Science Technology Management and Research

Available online at: www.ijstmr.com

Comparative Analysis Between Bayesian Classifier and Decision Tree Classifier

Sheetal Bhagat
M Tech. (CSE) Semester
Lord Krishna College of Technology
Indore M.P. India

Vijay Kumar Verma
Asst. Professor Dept. C.S.E.
Lord Krishna College of Technology
Indore M.P. India

Abstract-: ***Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Such analysis can help provide us with a better understanding of the data at large. Whereas classification predicts categorical (discrete, unordered) labels, prediction models continuous valued functions. Classification model categorize bank loan applications as either safe or risky, or a prediction model to predict the expenditures in dollars of potential customers on computer equipment given their income and occupation. Many classification and prediction methods have been proposed by researchers in machine learning, pattern recognition, and statistics. Most algorithms are memory resident, typically assuming a small data size. Recent data mining research has built on such work, developing scalable classification and prediction techniques capable of handling large disk-resident data Data classification is a two-step process. In the first step, a classifier is built describing a predetermined set of data classes or concepts. This is the learning step (or training phase), where a classification algorithm builds the classifier by analyzing or "learning from" a training set made up of database tuples and their associated class labels. Decision tree induction is the learning of decision trees from class-labelled training tuples. A decision tree is a flowchart-like tree structure, where each internal node (non leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node. Bayesian classification is based on Bayes theorem. Classification algorithms have found a simple Bayesian classifier known as the naïve Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases. In The proposed work we present a comparative study over Bayesian classifier and Decision tree induction***

Keywords: Classification, Prediction, Bayes, Decision tree, Accuracy, Loan

I. INTRODUCTION

Databases are rich with hidden information that can be used for intelligent decision making. Classification and prediction are two form of data analysis that can be used to extract models describing important data classes or to predict future data trends. Such analysis can help provide us with a better understanding of the data at large. Whereas classification predicts categorical (discrete, unordered) labels, prediction models continuous valued functions. Classification model categorize bank loan applications as either safe or risky, or a prediction model to predict the expenditures in dollars of potential customers on computer equipment given their income and occupation. Many classification and prediction methods have been proposed by researchers in machine learning, pattern recognition, and statistics. Most algorithms are memory resident, typically assuming a small data size. Recent data mining research has built on such work, developing scalable classification and prediction techniques capable of handling large disk-resident data. Data classification, such as how to build decision tree classifiers, Bayesian classifiers, Bayesian belief networks, and rule based classifiers. Back propagation (a neural network technique) is also discussed, in addition to a more recent approach to classification known as support vector machines. Classification based on association rule mining is explored. Other approaches to classification, such as k-nearest-neighbor classifiers, case-based reasoning, genetic algorithms, rough sets, and fuzzy logic techniques, are introduced. Methods for prediction, including linear regression, nonlinear regression, and other regression-based models, are briefly discussed. Where applicable, you will learn about extensions to these techniques for their application to classification and prediction in large databases. Classification and prediction have numerous applications, including fraud detection, target marketing, performance prediction, manufacturing, and medical diagnosis[15].

The accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. The associated class label of each test tuple is compared with the learned classifier's class prediction for that tuple. There several methods have been developed for estimating classifier accuracy. If the accuracy of the classifier is considered acceptable, the classifier can be used to classify future data tuples for which the class label is not known. (Such data are also referred to in the machine learning literature as "unknown" or "previously unseen" data) For example, the classification rules learned from the analysis of data from loan applications can be used to approve or reject new or future loan applicants. Data prediction is a two step process, similar to that of data classification a. However, for prediction, we lose the terminology of "class label attribute" because the attribute for which values are being predicted is continuous-valued (ordered) rather than categorical (discrete-valued and unordered). The attribute can be referred to simply as the predicted attribute[16].

II. CLASSIFICATION TECHNIQUES

Some of the important classification techniques include decision tree classifiers, Bayesian classifiers, Bayesian belief networks, and rule based classifiers. Back propagation (a neural network technique) and support vector machines etc [15,17,18].

Decision Tree Classifiers

Decision tree induction is the learning of decision trees from class-labeled training tuples. A decision tree is a flowchart-like tree structure, where each internal node (non leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node.

A typical decision tree is shown in figure 1. It represents the concept buys computer, that is, it predicts whether a customer at All Electronics is likely to purchase a computer. Internal nodes are denoted by rectangles, and leaf nodes are denoted by ovals. Some decision tree algorithms produce only binary trees (where each internal node branches to exactly two other nodes), whereas others can produce non binary trees

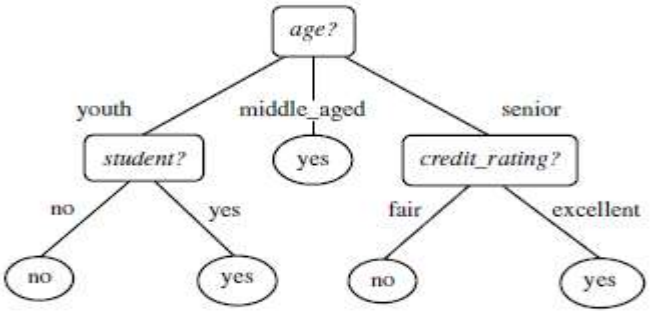


Figure 1 decision tree classifications

The construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data. Their representation of acquired knowledge in tree form is intuitive and generally easy to assimilate by humans. The learning and classification steps of decision tree induction are simple and fast.

Bayesian Classification

Bayesian classification is based on Bayes theorem. Classification algorithms have found a simple Bayesian classifier known as the naïve Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases. Naïve Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computations involved and, in this sense, is considered "naïve" Bayesian belief networks are graphical models, which unlike naïve Bayesian classifiers allow the representation of dependencies among subsets of attributes. Bayesian belief networks can also be used for classification [23, 24].

The naïve Bayesian classifier, or simple Bayesian classifier, works as follows:

1. Let D be a training set of tuples and their associated class labels. As usual, each tuple is represented by an n-dimensional attribute vector, **X** = (X_1, X_2, : : : ,X_n), depicting n measurements made on the tuple from n attributes, respectively, A_1, A_2, : : : , A_n.

2. Suppose that there are m classes, C_1, C_2,.....,C_m. 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

$$P(C_i|X) > P(C_j|X) \quad \text{for } 1 \le j \le m, j \ne i.$$

Thus to maximize P (C_i/X). The class C_i for which P (Ci/X) is maximized is called the maximum posteriori hypothesis. Bayes' theorem

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}.$$

3. As $P(X)$ 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, $P(C_1) = P(C_2) =$…… $= P(C_m)$, and we would therefore maximize $P(X|Ci)$. Otherwise, we maximize $P(X|Ci)P(Ci)$. Note that the class prior probabilities may be estimated by $P(Ci)=|Ci,_D| / |D|$ where $|Ci,_D|$ is the number of training tuples of class Ci in D

Bayesian classifiers have the minimum error rate in comparison to all other classifiers. However, in practice this is not always the case, owing to inaccuracies in the assumptions made for its use, such as class conditional independence, and the lack of available probability data. Bayesian classifiers are also useful in that they provide a theoretical justification for other classifiers that do not explicitly use Bayes' theorem. For example, under certain assumptions, it can be shown that many neural network and curve-fitting algorithms output the maximum posteriori hypothesis, as does the naïve Bayesian classifier.

Bayesian Belief Network

The naïve Bayesian classifier makes the assumption of class conditional independence, that is, given the class label of a tuple, the values of the attributes are assumed to be conditionally independent of one another. This simplifies computation. When the assumption holds true, then the naïve Bayesian classifier is the most accurate in comparison with all other classifiers. In practice, however, dependencies can exist between variables. Bayesian belief networks specify joint conditional probability distributions. They allow class conditional independencies to be defined between subsets of variables. They provide a graphical model of causal relationships, on which learning can be performed. Trained Bayesian belief networks can be used for classification. Bayesian belief networks are also known as belief networks, Bayesian networks, and probabilistic networks.

A belief network is defined by two components a directed acyclic graph and a set of conditional probability tables. Each node in the directed acyclic graph represents a random variable. The variables may be discrete or continuous-valued. They may correspond to actual

attributes given in the data or to "hidden variables" believed to form a relationship (e.g., in the case of medical data, a hidden variable may indicate a syndrome, representing a number of symptoms that, together, characterize a specific disease). Each arc represents a probabilistic dependence.

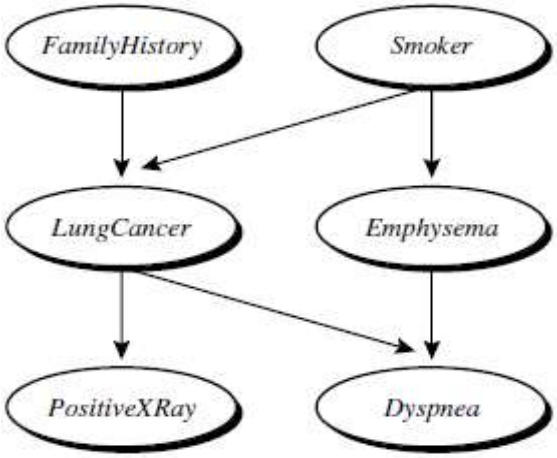


Figure 2 Simple Bayesian belief network

Rule-Based Classification

Rule-based classifiers, where the learned model is represented as a set of IF-THEN rules. We first examine how such rules are used for classification. Rules are a good way of representing information or bits of knowledge. A rule-based classifier uses a set of IF-THEN rules for classification. An IF-THEN rule is an expression of the form

IF *condition* THEN *conclusion*.

An example is rule *R*1,

R1: IF *age* = *youth* AND *student* = *yes* THEN *buys computer* = *yes*.

The "IF"-part (or left-hand side) of a rule is known as the rule antecedent or precondition. The "THEN"-part (or right-hand side) is the rule consequent. In the rule antecedent, the condition consists of one or more *attribute tests* (such as *age* = *youth*, and *student* = *yes*) that are logically ANDed. The rule's consequent contains a class prediction (in this case, we are predicting whether a customer will buy a computer). *R*1 can also be written as

R1: (*age* = *youth*) ^ (*student* = *yes*))(*buys computer* = *yes*).

If the condition (that is, all of the attribute tests) in a rule antecedent holds true for a given tuple, we say that the rule antecedent is satisfied (or simply, that the rule is satisfied) and that the rule covers the tuple.

Classification by Back propagation

Back propagation is a neural network learning algorithm. The field of neural networks was originally kindled by psychologists and neurobiologists who sought to develop and test computational analogues of neurons. Roughly speaking, a neural network is a set of connected input/output units in which each connection has a weight associated with it. During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label of the input tuples. Neural network learning is also referred to as connectionist learning due to the connections between units. Neural networks involve long training times and are therefore more suitable for applications where this is feasible. They require a number of parameters that are typically best determined empirically, such as the network topology or "structure." Neural networks have been criticized for their poor interpretability. For example, it is difficult for humans to interpret the symbolic meaning behind the learned weights and of "hidden units" in the network. These features initially made neural networks less desirable for data mining.

Advantages of neural networks, however, include their high tolerance of noisy data as well as their ability to classify patterns on which they have not been trained. They can be used when you may have little knowledge of the relationships between attributes and classes. They are well-suited for continuous-valued inputs and outputs, unlike most decision tree algorithms. They have been successful on a wide array of real-world data, including handwritten character recognition, pathology and laboratory medicine, and training a computer to pronounce English text. Neural network algorithms are inherently parallel; parallelization techniques can be used to speed up the computation process. In addition, several techniques have recently been developed for the extraction of rules from trained neural networks. These factors contribute toward the usefulness of neural networks for classification and prediction in data mining

K-Nearest Neighbor Classifiers (KNN)

K-Nearest neighbor classifiers are based on learning by analogy. The training samples are described by n dimensional numeric attributes. Each sample represents a point in an n-dimensional space. In this way, all of the 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 unknowns ample. "Closeness" is defined in terms of Euclidean distance, where the Euclidean distance, where the Euclidean distance between two points, $X=(x_1,x_2,\ldots\ldots,x_n)$ and $Y=(y_1,y_2,\ldots,y_n)$ is denoted by $d(X, Y)$.

$$d(X,Y) = \sqrt{\sum_{i=1}^{n} (xi - yi)^2}$$

Unlike decision tree induction and back propagation, nearest neighbor classifiers assign equal weight to each attribute. This may cause confusion when there are many irrelevant attributes in the data. Nearest neighbor classifiers can also be used for prediction, that is, to return areal-valued prediction for a given unknown sample. In this case, the classifier returns the average value of the real-valued associated with the *k* nearest neighbors of the unknown sample. The *k*-nearest neighbors' algorithm is amongst 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. If *k = 1*, then the object is simply assigned to the class of its nearest neighbor. In binary (two class) classification problems, it is helpful to choose *k* to be an odd number as this avoids tied votes. The same method can be used for regression, by simply assigning the property value for the object to be the average of the values of its *k* nearest neighbors. It can be useful to weight the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones.

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. The aim of SVM is to find the best classification function to distinguish between members of the two classes in the training data. The metric for the concept of the "best" classification function can be realized geometrically. 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.

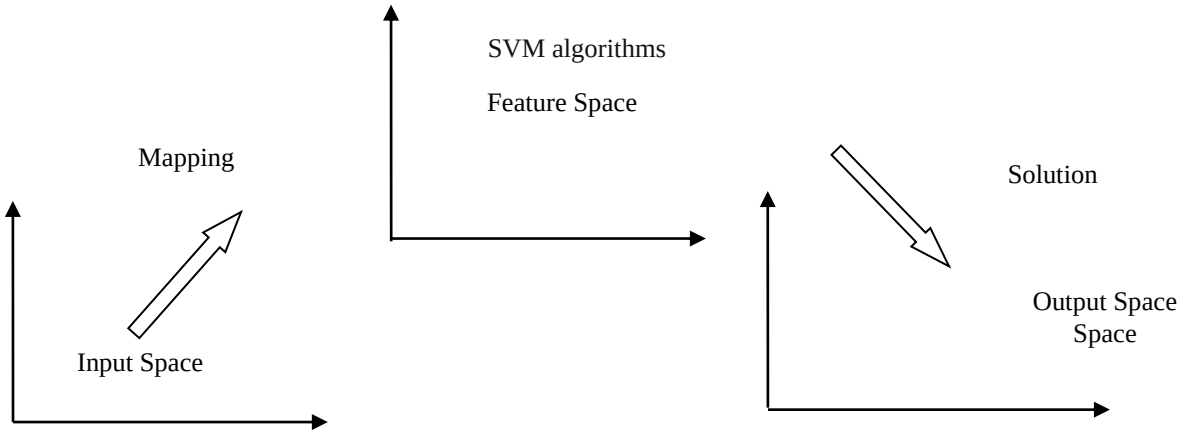


Figure 1.5 Support Vector Machine Classifications

The support vector machine classifier creates a hyper plane or multiple hyper planes in high dimensional space that is useful for classification, regression and other efficient tasks. SVM have many attractive features due to this it is gaining popularity and have promising empirical performance. SVM constructs a hyper plane in original input space to separate the data points. Some time it is difficult to perform separation of data points in original input space, so to make separation easier the original finite dimensional space mapped into new higher dimensional space. Kernel functions are used for non-linear mapping of training samples to high dimensional space. Various kernel function such as polynomial, Gaussian, sigmoid etc., are used for this purpose. SVM works on the principal that data points are classified using a hyper plane which maximizes the separation between data points and the hyper plane is constructed with the help of support vectors. Figure 2.4 shows the working of SVM classification algorithm.

III. LITERATURE SURVEY

In 2011 Mai Shouman, Tim Turner, Rob Stocker proposed "Using Decision Tree for Diagnosing Heart Disease Patients ". Heart disease is the leading cause of death in the world over the past 10 years. They investigates applying a range of techniques to different types of Decision Trees seeking better performance in heart disease diagnosis. A widely used benchmark data set is used in this research. To evaluate the performance of the alternative Decision Trees the sensitivity, specificity, and accuracy are calculated. The research proposes a model that outperforms J4.8 Decision Tree and Bagging algorithm in the diagnosis of heart disease patients. Decision Tree is one of the successful data mining techniques used in the diagnosis of heart disease. Yet its accuracy is not perfect. Most research applies the J4.8Decision Tree that is based on Gain Ratio and binary discretization. This research systematically tested combinations of discretization, decision tree type and voting to identify a more robust, more accurate method [1].

In 2012 M. Akhil Jabbar, Dr. Priti Chandra, Dr. B. L. Deekshatulu proposed **"**Heart Disease Prediction System using Associative Classification and Genetic Algorithm". They proposed an efficient associative classification algorithm using genetic approach for heart disease prediction. The main motivation for using genetic algorithm in the discovery of high level prediction rules is that the discovered rules are highly comprehensible, having high predictive accuracy and of high interestingness values. They proposed a system for heart disease prediction using data mining techniques. In our feature work we plan to reduce no. of attributes and to determine the attribute which contribute towards the diagnosis of disease using genetic algorithm [2].

In 2012 Chaitrali S. Dangareand Sulabha S. Apte proposed "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques ".They analyzed prediction systems for Heart disease using more number of input attributes. The system uses medical terms such as sex, blood pressure, cholesterol like 13 attributes to predict the likelihood of patient getting a Heart disease. Until now, 13 attributes are used for prediction. They added two more attributes i.e. obesity and smoking. The data mining classification techniques, namely Decision Trees, Naive Bayes, and Neural Networks are analyzed on Heart disease database. The performance of these techniques is compared, based on accuracy. As per our results accuracy of Neural Networks, Decision Trees, and Naive Bayes are 100%, 99.62%, and 90.74% respectively [3].

In 2012 Sunita Soni and O.P.Vyas proposed "Fuzzy Weighted Associative Classifier: A Predictive Technique For Health Care Data Mining". They extend the problem of classification using Fuzzy Association Rule Mining and propose the concept of Fuzzy Weighted Associative Classifier (FWAC). Classification based on Association rules is considered to be effective and advantageous in many cases. Associative classifiers are especially fit to applications where the model may assist the domain experts in their decisions. They proposed a new Fuzzy Weighted Associative Classifier (FWAC) that generates classification rules using Fuzzy Weighted Support and Confidence framework. The naïve approach can be used to generating strong rules instead of weak irrelevant rules. Fuzzy logic is used in partitioning the domains. This work presents a new foundational approach to Fuzzy Weighted Associative Classifiers where quantitative attributes are discritized to get transformed binary database. In such data base each record fully belongs to only one fuzzy set. Such database will suffer the crisp boundary problem [4].

In 2013 V. Krishnaiah , Dr. G. Narsimha, Dr. N. Subhash Chandra proposed "Diagnosis of Lung Cancer Prediction System Using Data Mining Classification Techniques". They briefly examine the potential use of classification based data mining techniques such as Rule based, Decision tree, Naïve Bayes and Artificial Neural Network to massive volume of healthcare data. The healthcare industry collects huge amounts of healthcare data which, unfortunately, are not "mined" to discover hidden information. This is an extension of Naïve Bayes to imprecise probabilities that aims at delivering robust classifications also when dealing with small or incomplete datasets. Discovery of hidden patterns and relationships often goes unexploited. They propose a model for early detection and correct diagnosis of the disease which will help the doctor in saving the life of the patient. Prototype lung cancer disease prediction system is developed using data mining classification techniques. The system extracts hidden knowledge from a historical lung cancer disease database[5].

In 2013 Shamsher Bahadur Patel , Pramod Kumar Yadav, Dr. D. P. Shukla proposed **"**Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques". They predict the diagnosis of heart disease with reduced number of attributes. They used fourteen attributes involved in predicting heart disease. But fourteen attributes are reduced to six attributes by using Genetic algorithm. Subsequently three classifiers like Naive Bayes, Classification by Clustering and Decision Tree are used to predict the diagnosis of heart disease after the reduction of number of attributes. They used genetic algorithm to determine the attributes which contribute more towards the diagnosis of heart ailments which indirectly reduces the number of tests which are needed to be taken by a patient. Fourteen attributes are reduced to 6 attributes using genetic search [6].

In 2013 M. Akhil Jabbar, Dr. B.L Deekshatulu, Dr. Priti Chandra proposed "Heart Disease Classification Using Nearest Neighbor Classifier with Feature Subset Selection". They investigate and apply K nearest neighbor with feature subset selection in the diagnosis of heart disease. The experimental results show that applying feature subset selection to KNN will enhance the accuracy in the diagnosis of heart disease for Andhra Pradesh population. India with a population of more than 1 billion accounted for 60%of the world heart diseases. Andhra Pradesh is in risk of more deaths due to heart disease. They employed KNN algorithm with feature subset selection to determine the features which contributes more towards the disease prediction. This method indirectly reduces no. of tests to be taken by patients. This prediction model even helps the doctors in efficient decision making process with fewer attributes to diagnose the heart disease [7].

In 2013 M. Akhil Jabbar, B.L Deekshatulu & Priti Chandra proposed "Classification of Heart Disease using Artificial Neural Network and Feature Subset Selection". They introduced a classification approach which uses ANN and feature subset selection for the classification of heart disease. PCA is used for preprocessing and to reduce no. Of attributes which indirectly reduces the no. of diagnosis tests which are needed to be taken by a patient. Experimental results show that accuracy improved over traditional classification techniques. This system is feasible and faster and more accurate for diagnosis of heart disease. They proposed a new feature selection method for heart disease classification using ANN and various feature selection methods for Andhra Pradesh Population. They applied different feature selection methods to rank the attributes which contribute more towards classification of heart disease, which indirectly reduces the no. of diagnosis tests to be taken by a patient [8].

In 2013 V. Manikantan & S. Latha proposed "Predicting the Analysis of Heart Disease Symptoms Using Medicinal Data Mining Methods ".They used C4.5 algorithm as the training algorithm to show rank of heart attack with the decision tree. Finally, the heart disease database is clustered using the K-means clustering algorithm, which will remove the data applicable to heart attack from the database. The results showed that the medicinal prescription and designed prediction system is capable of prophesying the heart attack successfully. In future work, planned to design and develop an efficient heart attack prediction system with Patient Prescription Support using the web mining and data warehouse techniques[9].

In 2014 Mariammal D., Jayanthi S., Dr. P.S.K. Patra proposed "Major Disease Diagnosis and Treatment Suggestion System Using Data Mining Techniques". They proposed a model to systematically close those gaps to discover if applying single and multiple data mining techniques to all disease treatment data can provide as reliable performance as that achieved in diagnosing disease. Using multiple data mining techniques the accuracy also improved. Disease prediction is a major challenge in the health care industry. Instead of going for a number of tests, predicting the major disease with less number of attributes is a challenging task in Data Mining. Decision Support

in Disease Prediction System is developed using all the five data mining techniques. The Disease diagnosis system extracts hidden knowledge from a historical disease database [10] .

In 2015 Ebenezer Obaloluwa Olaniyi and Oyebade Kayode Oyedotun proposed "Heart Diseases Diagnosis Using Neural Networks Arbitration". They proposed causes of heart diseases, the complications and the remedies for the diseases have been considered. An intelligent system which can diagnose heart diseases has been implemented. This system will prevent misdiagnosis which is the major error that may occur by medical doctors. The dataset of statlog heart disease has been used to carry out this experiment. The dataset comprises attributes of patients diagnosed for heart diseases. The diagnosis was used to confirm whether heart disease is present or absent in the patient. The datasets were obtained from the UCI Machine Learning [11].

In 2016 Isra'a Ahmed Zriqat, Ahmad Mousa Altamimi, Mohammad Azzeh proposed "A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods". They motivate is to develop an effective intelligent medical decision support system based on data mining techniques. They used five data mining classifying algorithms, with large datasets, have been utilized to assess and analyze the risk factors statistically related` to heart diseases in order to compare the performance of the implemented classifiers (e.g., Naïve Bayes, Decision Tree, Discriminate, Random Forest, and Support Vector Machine). Results of the conducted experiments showed that all classification algorithms are predictive and can give relatively correct answer. Although ensemble learning has been proved to produce superior results, but in our case the decision tree has outperformed its ensemble version [12].

In 2017 Sanjay Kumar Sen proposed "Predicting and Diagnosing of Heart Disease Using Machine Learning Algorithms". The main objective of this research is predicting the heart disease of a patient using machine learning algorithms. Comparative study of the various performances of machine learning algorithms is done through graphical representation of the results. They carried out an experiment to find the predictive performance of different classifiers. They select four popular classifiers considering their qualitative performance for the experiment. They also choose one dataset from heart available at UCI machine learning repository. Naïve base classifier is the best in performance. In order to compare the classification performance of four machine learning algorithms, classifiers are applied on same data and results are compared on the basis of misclassification and correct classification rate and according to experimental results it can be concluded that Naïve base classifier is the best as compared to Support Vector Machine, Decision Tree and K-Nearest Neighbour [13].

In 2018 Poornima V, Gladis D proposed "A novel approach for diagnosing heart disease with hybrid classifier". They r proposed an Orthogonal Local Preserving Projection (OLPP) method to reduce the function dimension of the input high-dimensional data. The dimension reduction improves the prediction rate with the help of hybrid classifier i.e. Group Search Optimization Algorithm (GSO) combine with the Levenberg-Marquardt (LM) training algorithm in the neural network. The LM training algorithm is used to solve the optimization problem and it determines the best network parameters such as weights and bias that minimizes the error. The final output of the optimization technique is combined with the performance metrics as accuracy, sensitivity, and specificity. From the result, it is observed that hybrid optimization techniques increase the accuracy of the heart disease prediction system [14].

IV. PROBLEM STATEMENT

Several method have been proposed by various researcher to improve the problem of existing algorithms but evolution of Classification can be done using following criteria:
Accuracy: The accuracy of a classifier refers to the ability of a given classifier to correctly predict the class label of new or previously unseen data. Accuracy can be estimated using one or more test sets that are independent of the training set. Estimation techniques, such as cross-validation and bootstrapping. Because the accuracy computed is only an estimate of how well the classifier swill do on new data tuples, confidence limits can be computed to help gauge this estimate.
Speed: This refers to the computational costs involved in generating and using the given classifier or predictor.
Robustness: This is the ability of the classifier or predictor to make correct predictions given noisy data or data with missing values.
Scalability: This refers to the ability to construct the classifier or predictor efficiently given large amounts of data.
Interpretability: This refers to the level of understanding and insight that is provided by the classifier or predictor. Interpretability is subjective and therefore more difficult to assess.

V. OBJECTIVES

There are several algorithms and methods have been text document clustering. But problem are always arises for finding a new algorithm and process for extracting knowledge for improving accuracy and efficiency. Our major objective are-

1) Apply Euclidean distance measures for distance calculation for one title with another title using terms only.
2) Apply Euclidean distance measures for distance calculation for one title with another title using terms and tokens.
3) find out which distance method is more accurate

VI. COMPARATIVE ANALYSIS

Naive Bayesian Classifiers
The Naïve Bayes Classifier technique is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. Naïve Bayes model identifies the characteristics of patients with heart disease. It shows the probability of each input attribute for the predictable state.
A conditional probability is the likelihood of some conclusion, C, given some evidence/observation, E, where a dependence relationship exists between C and E.This probability is denoted as P(C |E) where

$$P(C|E) = \frac{P(E|C)P(C)}{P(E)}$$

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 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

$$P(Ci|X) = \frac{P(X|Ci)P(Ci)}{P(X)}$$

As 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). Note that the class prior probabilities may be estimated by P(Ci)=|Ci,D|/|D|, where |Ci,D| is the number of training tuples of class Ci in D.

Given data sets with many attributes, it would be extremely computationally expensive to compute P(X|Ci).In order to reduce computation in evaluating P(X|Ci), the naïve assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple (i.e., that there are no dependence relationships among the attributes).

We need to maximize P(X|Ci)P(Ci), for i=1, 2. P(Ci), the prior probability of each class, can be computed based on the training tuples:

P(Loan r=Yes) = 9/14=0.643
P(Loan r=No) = 5/14=0.357

To compute P(X|Ci), for i=1, 2, we compute the following conditional probabilities:

P(age=Youth|Loan=Yes) =2/9=0.222
P(age=Youth|Loan=No) =3/5=0.600
P(income=Medium|Loan=Yes) =4/9=0.444
P(income=Medium|Loan=No) =2/5=0.400
P(Govt_Employee -=Yes|Loan=Yes) =6/9=0.667
P(Govt_Employee =Yes|Loan=No) =1/5=0.200
P(Credit_rating=Fair|Loan =Yes) =6/9=0.667
P(Credit_rating=Fair|Loan=No) =2/5=0.400

Using the above probabilities, we obtain

P(X| Loan =Yes) = P(Age=Youth| Loan=Yes) * P(Income=Medium| Loan=Yes) * P(Govt_Employee=Yes| Loan =Yes) * P(Credit_rating=Fair| Loan =Yes)
=0.222 x 0.444 x 0.667 x 0.667=0.044

Similarly,

P(X| Loan=no) = 0.600 x 0.400 x 0.200 x 0.400 = 0.019.

To find the class, Ci, that maximizes P(X|Ci)P(Ci),
We compute
P(X| Loan =Yes) *P(Loan = Yes)=0.044 x 0.643 = 0.028
P(X| Loan=no) *P(Loan =no) =0.019 x 0.357 = 0.007
Therefore,
The naïve Bayesian classifier predicts Loan = yes for tuple X.

Decision Tree Induction or ID3 Classifier (Iterative Dichotomiser)

Decision tree induction is the learning of decision trees from class-labeled training tuples. A decision tree is a flowchart-like tree structure, where each internal node (non leaf node)denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or terminal node) holds a class label. The topmost node in a tree is the root node.

A typical decision tree contains internal nodes are denoted by rectangles, and leaf nodes are denoted by ovals. Some decision tree algorithms produce only binary trees (where each internal node branches to exactly two other nodes), whereas others can produce non binary trees. How are decision trees used for classification for given tuple, X, for which the associated class label is unknown, the attribute values of the tuple are tested against the decision tree. A path is traced from the root to a leaf node, which holds the class prediction for that tuple. Decision trees can easily be converted to classification rules.

Decision tree classifiers so popular because. The construction of decision tree classifiers does not require any domain knowledge or parameter setting, and therefore is appropriate for exploratory knowledge discovery. Decision trees can handle high dimensional data. Their representation of acquired knowledge in tree form is intuitive and generally easy to assimilate by humans. The learning and classification steps of decision tree induction are simple and fast. In general, decision tree classifiers have good accuracy. However, successful use may depend on the data at hand. Decision tree induction algorithms have been used for classification in many application areas, such as medicine ,manufacturing and production, financial analysis, astronomy, and molecular biology. Decision trees are the basis of several commercial rule induction systems. When decision trees are built, many of the branches may reflect noise or outliers in the training data. Tree pruning attempts to identify and remove such branches, with the goal of improving classification accuracy on unseen data.

The information gain measure is biased toward tests with many outcomes. That is, it prefers to select attributes having a large number of values. For example, consider an attribute that acts as a unique identifier, such as product ID. A split on product ID would result in a large number of partitions (as many as there are values), each one containing just one tuple. Because each partition is pure, the information required to classify data setD based on this partitioning would be Info product ID(D) = 0. Therefore, the information gained by partitioning on this attribute is maximal. Clearly, such a partitioning is useless for classification

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right).$$

Information gain

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

$$Info(D) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.940$$

$$Info_{Age}(D) = \frac{5}{14} \times (-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}) + \frac{4}{14} \times (-\frac{4}{4}\log_2\frac{4}{4} - \frac{0}{4}\log_2\frac{0}{4}) + \frac{5}{14} \times (-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5})$$
$$= 0.694$$

$$Gain(Age) = Info(D) - Info_{Age}(D) = 0.940 - 0.694 = 0.246bits$$

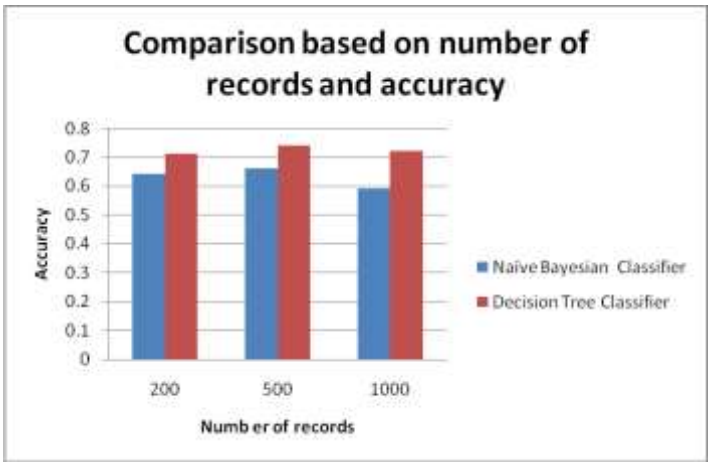
VII. COMPARATIVE ANLYSIS

For comparing the performance Between Bayesian Classifier and Decision Tree Classifier. we use number of records and accuracy of the classifier. In table 1 accuracy of the Naïve Bayesian Classifiersand Decision Tree Classifier with 200, 500 and 1000 records.

Table 1 Number of Record and accuracy in percentages

Number of Records	Naive Bayesian Classifiers	Decision Tree Classifier
200	0.64	0.71
500	0.66	0.74
1000	0.59	0.72

Figure 5.1 Comparisons using number of record and accuracy



VIII. CONCLUSION

There are several algorithms and methods have been developed to solve classification problem. Decision tree induction is easy to understand and explain. It has multiple interesting features those take care various issues like missing values, outlier, identifying most significant dimensions and others. It can also easily handle feature interactions and they're non-parametric. Major disadvantage is over-fitting, but that's where ensemble methods like random forests (or boosted trees) come in. Another one, it does not work well with continuous target variable compare to categorical. Bayesian classifier is type of supervised learning algorithm. It assumes an underlying probabilistic model (Bayes theorem). It is majorly used when more number of classes to predict likes Text Classification, Spam Filtering, Recommendation System and others. We proposed a comparative study of both these techniques. Decision tree induction gives good accuracy as compared to the Bayesian classifier.

REFERENCE

1. Mai Shouman, Tim Turner, Rob Stocker Using Decision Tree for Diagnosing Heart Disease Patients Proceedings of the 9-th Australasian Data Mining Conference (AusDM'11), Ballarat, Australia Copyright © 2011, Australian Computer Society
2. M. Akhil jabbar, Dr B.L Deekshatulu, Dr. Priti Chandra "Heart Disease Classification Using Nearest Neighbor Classifier With Feature Subset Selection Computer Science and Telecommunications 2013|No.3(39) ISSN 1512-1232.

3. Chaitrali S. Dangare Sulabha S. Apte, " Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques" International Journal of Computer Applications (0975 – 888) Volume 47– No.10, June 2012
4. SunitaSoni 1 and O.P.Vyas "FUZZY WEIGHTED ASSOCIATIVE CLASSIFIER: A PREDICTIVE TECHNIQUE FOR HEALTH CARE DATAMINING" International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.1, February 2012
5. V. Krishnaiah, Dr. G. Narsimha, Dr. N. Subhash Chandra "Diagnosis of Lung Cancer Prediction SystemUsing Data Mining Classification Techniques" V. Krishnaiah et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (1) , 2013, 39 - 45
6. Shamsher Bahadur Patel, Pramod Kumar Yadav2, Dr. D. P.Shukla3 " Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques" IOSR Journal of Agriculture and Veterinary Science (IOSR-JAVS) e-ISSN: 2319-2380, p-ISSN: 2319-2372. Volume 4, Issue 2 (Jul. - Aug. 2013), PP 61-64 www.iosrjournals.org
7. M. Akhil jabbar1, Dr B.L Deekshatulu2, DrPriti Chandra3 " Heart Disease Classification Using Nearest Neighbor Classifier With Feature Subset Selection" GESJ: Computer Science and Telecommunications 2013|No.3(39)
8. M. Akhil Jabbar, B.L Deekshatulu&Priti Chandra "Classification of Heart Disease using Artificial Neural Networkand Feature Subset Selection" Global Journal of Computer Science and TechnologyNeural& Artificial IntelligenceVolume 13 Issue 3 Version 1.0 Year 2013Type: Double Blind Peer Reviewed International Research JournalPublisher: Global Journals Inc. (USA)Online ISSN: 0975-4172 & Print ISSN: 0975-4350
9. V. Manikantan& S. Latha Predicting the Analysis of Heart Disease Symptoms Using Medicinal Data Mining Methods International Journal on Advanced Computer Theory and Engineering (IJACTE) ISSN 2319 – 2526, Volume-2, Issue-2, 2013.
10. Mariammal. D, Jayanthi. S, Dr. P. S. K. Patra Major Disease Diagnosis and Treatment Suggestion System Using Data Mining Techniques International Journal of Advanced Research in Computer Science & Technology IJARCST All Rights Reserved 338 Vol. 2 Issue Special 1 Jan-March 2014 ISSN: 2347 - 8446 (Online) ISSN: 2347 - 9817
11. Ebenezer ObaloluwaOlaniyi and OyebadeKayodeOyedotun "Heart Diseases Diagnosis Using Neural Networks Arbitration" I.J. Intelligent Systems and Applications, 2015, 12, 75-82 Published Online November 2015 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijisa.2015.12.08
12. Isra'a Ahmed Zriqat, Ahmad MousaAltamimi, Mohammad Azzeh "A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods" International Journal of Computer Science and Information Security (IJCSIS), Vol. 14, No. 12, December 2016 ttps://sites.google.com/site/ijcsis/ ISSN 1947-5500
13. Sanjay Kumar Sen "Predicting and Diagnosing of Heart Disease Using Machine LearningAlgorith" International Journal Of Engineering And Computer Science ISSN:2319-7242Volume 6 Issue 6 June 2017, Page No. 21623-21631Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i6.14
14. Poornima V, Gladis D A novel approach for diagnosing heart disease with hybrid Biomedical Research 2018; 29 (11): 2274-2280 ISSN 0970-938Xwww.biomedres.info
15. Han, J. and Kamber, M. (2006): Data Mining: Concepts and Techniques. Second ed. The Morgan Kaufmann Series in Data Management Systems Elsevier .
16. Arun k Pujari "Data mining techniques " .
17. Albert Bifet, Geoff Holmes, Richard Kirkby and Bernhard, " Data Stream Mining A Practical Approach" Pfahringer May 2011 .
18. Krishnapuram, B., et al., A Bayesian approach to joint feature selection and classifier design. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2004. 6(9): p. 1105-1111.