IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 19, Issue 1, Ver. IV (Jan.-Feb. 2017), PP 61-65
www.iosrjournals.org

Classification Techniques: A Review

Rajwinder Kaur, Prince Verma
(CSE Department, CTIEMT Shahpur, India)
(CSE Department, CTIEMT Shahpur, India)

Abstract: *Data mining is a process to extract information from a huge amount of data and transform it into an understandable structure. Data mining provides the number of tasks to extract data from large databases such as Classification, Clustering, Regression, Association rule mining. This paper provides the concept of Classification. Classification is an important data mining technique based on machine learning which is used to classify the each item on the bases of features of the item with respect to the predefined set of classes or groups. This paper summarises various techniques that are implemented for the classification such as k-NN, Decision Tree, Naïve Bayes, SVM, ANN and RF. The techniques are analyzed and compared on the basis of their advantages and disadvantages.*

Keywords: *Introduction, classification and techniques, Advantages and Disadvantages.*

I. Introduction

Data mining is defined as discovering the interesting information from the large firm of data stored in database, warehouse or other information repositories [1]. Data mining is the technique to find the necessary knowledge from the hidden pattern from large databases. The pattern that are satisfied and enough to the user's requirements is called knowledge. The output of a program which discovers that useful pattern is called discovered knowledge.

Data mining process: - Data mining is also known as knowledge discovery in database (KDD) [2]. KDD is a Knowledge Discovery in Database process in which the data mining is one of the steps of KDD process. In KDD, the different available data sources are analyzed by using various data mining algorithms [2].

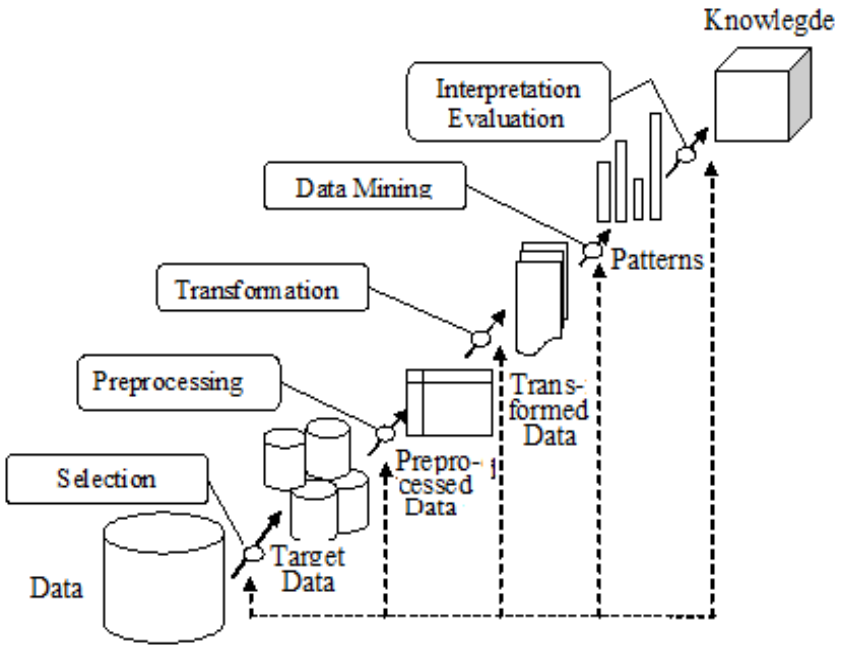

Fig. 1 KDD process

1.1 **Data Mining Task: -** Data mining mainly involve the following tasks to extract data from large database [3]:-

1. **Classification: -** Classification is a technique of data mining to classify each item into predefined set of groups or classes. Classification has number of applications in customer segmentation, business modelling, and drug response modelling and credit analysis [18].

It is the process of finding a model or function that describes & distinguishes data classes or concepts for the aim of being able to use the model to predict the class of object whose class label is unknown.

2. **Regression: -** It is learning a function which maps a data item to a real-valued prediction variable.
3. **Clustering: -** A cluster is a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters.
4. **Association rule mining:-** It is a technique of data mining that identifies strong and interesting co-occurrence of items in database
5. **Summarization: -** It involves methods for detecting a compact description for a subset of data.

II. Data Classification Techniques

2.1 KNN (K Nearest Neighbors Algorithm)

k-NN algorithm also identified as Instance based learning or lazy learning in which function is only approximated locally and all computation is postponed until classification [1]. The k-NN is a nonparametric Method which is used in pattern recognition for classification and regression both. The number of methods such as Euclidean distance, cosine measure etc. are used to measure the same features between two items [4].

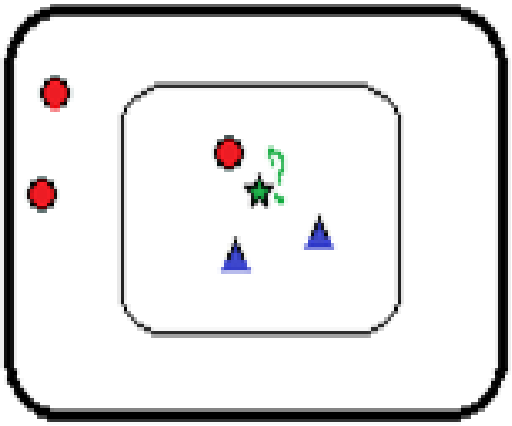

Fig. 2 k-NN Classification

In above fig, the test sample (green star) should be classified either to the first class of blue triangles or to the second class of red circles. If k = 3 (inner square) it is assigned to the first class because there are 2 triangles and only 1 circle inside the inner square. If k = 5 (outer square) it is assigned to the second class (3 circles vs. 2 triangles inside the outer circle).

2.2 Decision tree

Decision tree is a technique which is commonly used in data mining. A decision tree is a tree which consist both internal as well as external nodes connected by branches [17]. An internal node is a decision making unit that take decision which child node to visit next. The external node has no child nodes [5]. In Decision tree, classes are going to be rejected until the correct class is reached. For this purpose the feature space is to split into no. of different parts on the basis of specific threshold. The number of techniques like class assignment rule, splitting criterion and stop-splitting rule are used to design a classification tree. [6] A greedy algorithm is the basic algorithm for decision tree induction [7].

2.3 Support vector machine

Support vector Machine is the type of supervised classification method.SVM algorithm involves two types of versions such as linear and non-linear versions [6]. In the first, linear version, hyperplanes or set of hyperplanes are used for separation of classes. A hyperplane is represented by the following equation

$$WX + b = 0 \qquad \text{Eq no. 1}$$

And in the second, non linear version, classes are not partition i.e. no straight lines can be constructed that separated the classes. With the help of support vectors and margins, SVM finds hyperplane [1]. For text classification, SVM is the most accurate method. It is also widely used in sentiment classification [8]. SVM can be used to learn radial basis function (RBF), polynomial and multi-layer perceptron (MLP) classifiers [7].

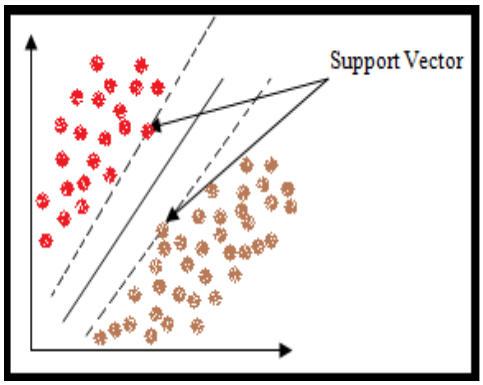

Fig. 3 Linear SVM

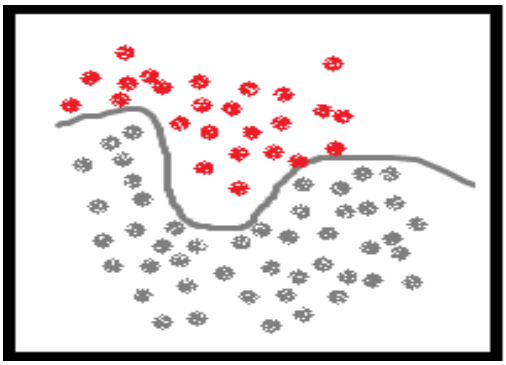

Fig.4 Non linear SVM

2.4 Naive bayes

A naive bayes is the classifier which is used to determine the probability for a given tuples to belong to a specific class. In both throughout traning and classifying steps, Naïve Bayes is used due to its easiness implementation and computation [9]. Pre-processed data is given to train input set as input by classifier by using naïve bayes and that trained model should applied on test data to generate either positive or negative sentiment[10]. The algorithm uses Bayes theorem and it assumes that all attributes value of a one feature is independent of any other features [11, 12].
The bayes theorem is:

$$P(M|N) = P(N|M).\ P(M)/\ P(N) \qquad \text{Eq no. 2}$$

M- Some hypothesis
N – Some Tuples
P (M |N) represents the posterior probability of M conditioned on N
P (M) represents prior probability of M, independent on N
P (N|M) represents the posterior probability of N conditioned on M [13].

2.5 Artificial Neural network

The term, neural network, is a circuit used to compose of a extremely large number of processing elements which are called Neuron [14]. Each element works barely on local information. As such, neural networks are very complex. Furthermore every element operates non-parellerliy, thus no system clock is there [15]. The main idea of a neural network is that to derive attributes from linear combinations of the input data, and then model the output as a non linear function of these attributes. The result is one of the most popular and effective forms of learning system. Neural networks are represented by a network diagram that is composed of nodes connected by directed links. Nodes are aligned in layers and the structure of the most used neural network involves of 3 layers: an input, a hidden and an output layer of nodes. The geo-temporal variations of crime and disorder are predicted. This technique uses to predict for crime incident by concentrate on geographical areas of concern. ANNs are the most widely implemented methods in forecasting building energy consumption [16].

2.6 Random forest

Random Forest (RF) is a classifier that takes the decision tree concept further by producing a large number of decision trees. The approach first takes a random sample of the data and identifies set of features to grow each decision tree. Then these decision trees have their out of bag error determined (error rate of the model) and then compared the collection of decision trees to find the joint set of variables that produce the strongest classification method. Among current classifier algorithms, RF has excellent accuracy. For estimating missing data, RF has an effective method and when a large proportion of the data are missing, it maintains accuracy [5, 7].

2.7 Classification and regression trees (CART)

Classification and regression trees (CART) applied for classification purposes considered as a decision tree method using the diachronic data. CART as a machine learning method is required to be determined number of classes. In order to achieve the best split, the question that separates the data into two homogeneous parts, after splitting the data into training and testing set, built a CART model without pruning on the training set and analyze the performance on the test set. CART uses Gini index to select the attribute which has maximum information [16].

III. Comparative Study of Some Classification Algorithm on The Basis of Their Advantages And Disadvantages

Algorithm	Advantages	Disadvantages
k-NN	• It is robust against raw data which is noisy. • due to process transparency it's easy to implement • It is effective for huge amount of data. • Very little information is needed to make it work.	• It lacks to select the value of N, except by cross validation, which makes very difficult to finding optimal value of N. • When there are many irrelevant attributes in the data, this may cause confusion and thus results into poor accuracy.
Decision tree	• Easy to understand and to generate rules. • It has excellent speed of learning and speed of classification. • Supports multi-classification.	• It may suffer from overfitting. • Does not easily handle non numeric data. • Can be quite large- pruning is necessary.
SVM	• It is less susceptible for over fitting of the feature input from the input items. • Classification accuracy with SVM is quite impressive or high.	• Multiclass items are not perfectly classified as number of items reduce gap of hyperplane.
Naive Bayes	• Work well on numeric as well as textual data. • It is easy to implement and computation are simple comparing with other algorithms. • It can be applied to large data set; no complicated iterative parameter estimation schemes are needed. • Performs well and it is robust.	• The precision of algorithm decreases if the amount of data is less. • Theoretically, when compared with other classifier, naive Bayes classifier have minimum error rate, but practically it is not always true, because of the assumption of class conditional independence.
ANN	• It is easy to use, with few parameters to adjust. • Can be implemented without any problem. • Can be executed in any application. • Applicable to a wide range of problems in real life.	• Requires high processing time if neural network is large. • Difficult to know how many neurons and layers are necessary. • Neural networks need training to operate.
Random forest	• Almost always have lower classification error. • Far easier for humans to understand. • Deal really well with uneven data sets that have missing variables.	• While training set is small, the high classification error rate in comparison with the no. of classes. • Not do well with imbalance data. • For some particular construction algorithm need to discrete data.
CART	• Nonparametric (no probabilistic assumptions). • Automatically performs variable selection. • Handles missing values automatically.	• It splits only by one variable. • Tree structures may be unstable. • Tree is optimal at each split – it may not be globally optimal.

IV. Conclusion

In modeling interactions, Classification methods are typically strong. Each of these methods can be used in different situations as needed where one tends to be useful while the other may not and vice-versa. This paper deals with various classification techniques that has been discussed like k-NN, Naïve bayes, Decision tree, SVM, ANN, Random forest. These classification algorithms can be implemented on various types of data sets according to performances like data of patients, financial data. as given in the paper, each classification technique has its own advantages and disadvantages. Naïve bayes is easy to implement but it does well with data in which the inputs are independent from one another. KNN algorithm is one of the simplest algorithms for classification. Even with such simplicity, it can give highly competitive results. Support vector machine is most widely used classification algorithm for sentiment analysis so it can generate better result. A decision tree is the algorithm that it doesn't require a lot of information about the data to create a tree that could be very accurate and very informative and the representation of the knowledge in decision tree in form of [IF-THEN] rules which is easier for humans understand. Random forest constitute one of the most effective and robust machine learning algorithms for many problems and provide result much accurate than that of other algorithms.

References

[1]. Manisha Kannvdiya,Kailash Patidar and Rishi Singh Kushwaha , "A Survey On: Different Techniques And Features Of Data Classification", *International Journal Of Research In Computer Applications And Robotics, Volume 4 Issue 6, pg 1-6, June 2016.*

[2]. Shital H. Bhojani and Dr. Nirav Bhatt, "Data Mining Techniques and Trends – A Review", *Global Journal For Research Analysis , Volume -5 , Issue -5 ,pg 252-254, May 2016.*

[3]. Supreet Kaur and Amanjot Kaur Grewal, "A Review Paper on Data Mining Classification Techniques for Detection of Lung Cancer", *International Research Journal of Engineering and Technology (IRJET), Volume: 03 Issue: 11, pg 1334-1338, Nov 2016.*

[4]. Navjot Kaur, "Data Mining Techniques used in Crime Analysis: - A Review", *International Research Journal of Engineering and Technology (IRJET), Volume: 03 Issue: 08, pg 1981-1984, Aug-2016.*

[5]. Trilok Chand Sharma and Manoj Jain, "WEKA Approach for Comparative Study of Classification Algorithm", *International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 4,pg 1925-1931, April 2013.*

[6]. Foram P. Shah and Vibha Patel, "A Review on Feature Selection and Feature Extraction for Text Classification", *IEEE WiSPNET 2016 conference, pg 2264-2268.*

[7]. G.Kesavaraj and Dr.S.Sukumaran, "A Study On Classification Techniques in Data Mining", *IEEE 4th ICCCNT, pg 1-7, July 4 - 6, 2013.*

[8]. Rodrigo Moraes, João Francisco Valiati and Wilson P. Gavião Neto, "Expert Systems with Applications", *Elsevier, pg 621-633, 2013.*

[9]. Vimalkumar B. Vaghela and Bhumika M. Jadav, "Analysis of Various Sentiment Classification Techniques", *International Journal of Computer Applications (0975 – 8887), Volume 140 – No.3,pg 22-27, April 2016.*

[10]. Vinod Bharat, Balaji Shelale, K.Khandelwal and Sushant Navsare, "A Review Paper on Data Mining Techniques", *IJESC, Volume 6 Issue No. 5,pg 6268-6271, 2016.*

[11]. Tina R. Patil and Mrs. S. S. Sherekar, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification", *International Journal Of Computer Science And Applications, Vol. 6, No.2, pg 256-261, Apr 2013.*

[12]. Pooja Sharma and Annu Mishra, "Classification Algorithm Using Random Concept On A Very Large Data Set: A Survey", *International Journal of Modern Trends in Engineering and Research, Volume 1, Issue 4, pg 57-64, 2014.*

[13]. Omkar Ardhapure, Gayatri Patil, Disha Udani and Kamlesh Jetha, "Comparative Study Of Classification Algorithm For Text Based Categorization", *International Journal of Research in Engineering and Technology, Volume: 05 Issue: 02, pg 217-220, Feb-2016.*

[14]. Shu-Hsien Liao, Pei-Hui Chu, Pei-Yuan Hsiao, "Data mining techniques and applications – A decade review from 2000 to 2011", *Elsevier, pg 11303-11311, 2012.*

[15]. A.S. Ahmad, M.Y. Hassan, M.P. Abdullah, H.A. Rahman,F. Hussin, H. Abdullah and R. Saidur, "A review on applications of ANN and SVM for building electrical energy consumption forecasting", *Elsevier, pg 102-109, 2014.*

[16]. Hakan Sahin and Abdulhamit Subasi, "Classification of the cardiotocogram data for anticipation of fetal risks using machine learning techniques", *Elsevier, pg 231-238, 2015.*

[17]. Dilip Kumar Choubey and Sanchita Paul, "Classification techniques for diagnosis of diabetes: a review", *Int. J. Biomedical Engineering and Technology, Vol. 21, No. 1, pg 15-39, 2016.*

[18]. Anirban Mukhopadhyay, Ujjwal Maulik, Sanghamitra Bandyopadhyay and Carlos Artemio Coello Coello, "A Survey of Multiobjective Evolutionary Algorithms for Data Mining: Part I", *IEEE Transactions On Evolutionary Computation, Vol. 18, NO. 1, pg 4-19, February 2014.*