International Journal of Scientific Research & Engineering Trends
Volume 4, Issue 6, Nov-Dec-2018, ISSN (Online): 2395-566X

Analysis about Classification Techniques on Categorical Data in Data Mining

Assistant Professor P. Meena
Department of Computer Science
Adhiyaman Arts and Science College for Women
Uthangarai , Krishnagiri, India
mmmeenchat@gmail.com

***Abstract* - In recent years, huge amount of data is stored in database which is increasing at a tremendous speed. This requires need for some new techniques and tool to intelligently analyze large dataset to acquire useful information. This growing need demands for a new research field called Knowledge Discovery in Database (KDD) or data mining. The main objective of the data mining process is to extract information from a large data set and transform it in to some meaning full form for future use. Classification is the one of data mining techniques which is used to classify categorical data item in a set of data in to one of predefined set of classes or groups, in this paper, the goal is to provide a comprehensive analysis of different classification techniques in data mining that incudes decision tree, Bayesian networks, k-nearest neighbor classifier & artificial neural network.**

***Keywords*-Data Mining, Classification techniques, Decision tree, KNN, Support vector machine.**

I.INTRODUCTION

Data mining is exploration and analysis of data from different perspectives, in order to discover meaningful pattern and rules. Data mining extract, transform and load transaction data on to the data warehouse system, it stores and manages the data in a multidimensional database system, provides the data access to business analysts and information technology professionals, it also analyzes the data by application software, and presents the data in a useful format, such as graph or table.

The objective of data mining is to design and work efficiently with large data sets. Data mining provides different techniques to discover hidden patterns from large data set. Data mining is a multistep process which requires accessing and analyzing results and taking appropriate action. The data to be accessed can be stored in one or more operational database. In data mining the data is mined using two learning approaches i.e. supervised learning or unsupervised learning.

Supervised Learning it is also known as directed data mining. In this, the variables under observation is split in to explanatory variables and dependent variables. The main objective is to determine a relationship between these two variables. In directed data mining techniques the values of the dependent variable must be well defined for a sufficiently large part of the data set. Supervised models are neural network and decision trees.

Unsupervised Learning in this, there is no distinction between dependent variables and explanatory variables, both are treated same. The main difference between supervised learning and unsupervised learning is that supervised learning is that supervised learning requires that the value of the target variable should be known and well defined while in unsupervised learning either target variable is unknown or its value is known for small number of cases. The process of data mining is shown in the following diagram.

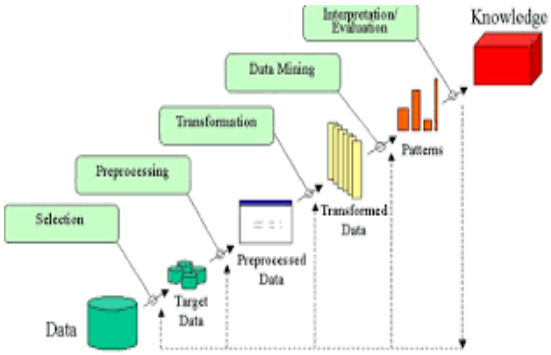


Fig.1 Data Mining Process.

In this paper, we worked on supervised learning based algorithms, performance of various learning algorithms are analyzed on weka tool. Rest of the paper is organized as follows. Section 2 describes various techniques of data mining algorithms while in section 3 different classification techniques have been explained while in section 4 we have shown the experimental setup and in section 5, the results are evaluated than in

section 6 comparison is done and then conclusion is derived which algorithm works better.

II. TECHNIQUES OF DATA MINING

The different data mining techniques used for the specific classes of six activities or tasks are as follows:

- Classification.
- Estimation.
- Prediction.
- Association rules.
- Clustering.
- Description and visualization.

III. CLASSIFICATION

There are two forms for data analysis

1. Classification and prediction

Classification is a machine learning technique which is used to assign each dataset to predefined groups while prediction is used to assign each dataset to pre-defined groups while prediction is used to predict continuous valued function. The main objective of classification is to accurately predict the target category for each item in a given data set. The classification is done in two steps Build the model

2. Use the classifier for classification

The accuracy of classification rules is estimated and if it is found acceptable then applied to other data sets. The simplest classification problem is binary classification which has only two possible values low and high. There are different techniques used for data classification to determine relationships between the values of the predictors and the target value. The commonly used method for data mining classification tasks can be classified in to the following groups.

- Decision tree induction methods
- Rule-based methods
- Memory based learning
- Neural networks
- Bayesian networks
- Support vector machines.

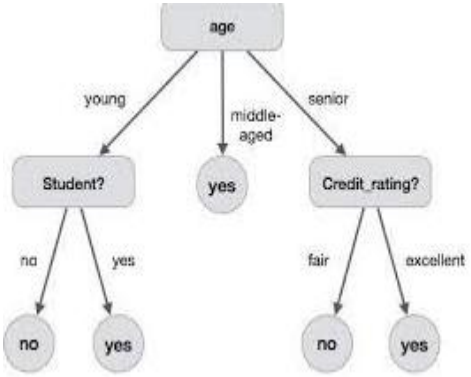
Decision tree is a method commonly used in data mining. It consists of a decision tree generated on the basis of instances. The decision tree is a directed tree with a node called “root” that has no incoming edges while all other nodes have exactly one incoming edge. A node with outgoing edge is called an internal node. All other nodes are called leaves (also known as terminal or decision nodes). Each internal node denotes a test on an attribute and each branch denotes the outcome of a test, and the leaf node holds a class label.

Decision tree induction algorithms function recursively. In this, first an attribute must be selected as the root node. In order to generate the most efficient decision tree, the root node must effectively split the data. Each internal node splits the instance space in to two or more sub-spaces, until they all have the same classification. The best split depends on the statistical property information gain. The basic algorithm for decision tree induction is a greedy algorithm that constructs decision tree in a recursive, top-down, divide-and-conquer manner.

The algorithm [10] is described as follows

- Create a node N;
- If all the samples are of the same class then
- Return N as a leaf node labeled with the class C;
- If attribute-list is empty then
- Return N as a leaf node labeled with the most common class in samples. Select test-attribute, the attribute among attribute-list with the highest information gain.
- Labeled node N with test-attribute;
- For each value ai of test-attribute
- Grow a branch from node N for the condition test attribute=ai;
- Let si be the set of samples for which test attribute=ai;
- If si is empty then
- Attach a leaf labeled with the most common class in samples;
- Else attach the node returned by generate_decision_tree (si-attribute-list-test-attribute).

Decision tree are usually unvaried since they are based on a single feature at each internal node. The diagram below shows the decision tree.



1

Fig. 2 decision tree.

The decision tree is a decision support tool which uses tree like graph model to represent event outcome. Most decision tree algorithms cannot perform well with problems that require diagonal partitioning. Decision tree can be significantly more complex representation for some concepts due to the replication problem a solution is using an algorithm to implement complex features at nodes in order to avoid replication.

To sum up, one of the most useful characteristics of decision tree is their comprehensibility. The assumption

made in the decision tree is that instances belonging to different classes have different values in at least one of their features. Decision trees trend to perform better when dealing with discrete categorical features.

3. k-Nearest neighbor

It is instance based learning or lazy learning which is used in classification and regression. The input taken in both the cases are same only the output determines whether k-NN is used for classification or regression. The output is class membership in k-NN classification while output is property value in object in k-NN regression. In this each attribute is assigned equal weight which may create confusion when there are irrelevant attributes in the data. It is also used for prediction for given unknown sample. It is one of the simplest method of all machine learning algorithms. Euclidean distance is used as distance metric for continuous variables while for discrete variables overlap metric is used. The short coming of k-NN algorithm is that it is that it is sensitive to local data structure. Fig 3 represents the simplest k-nearest neighbor.

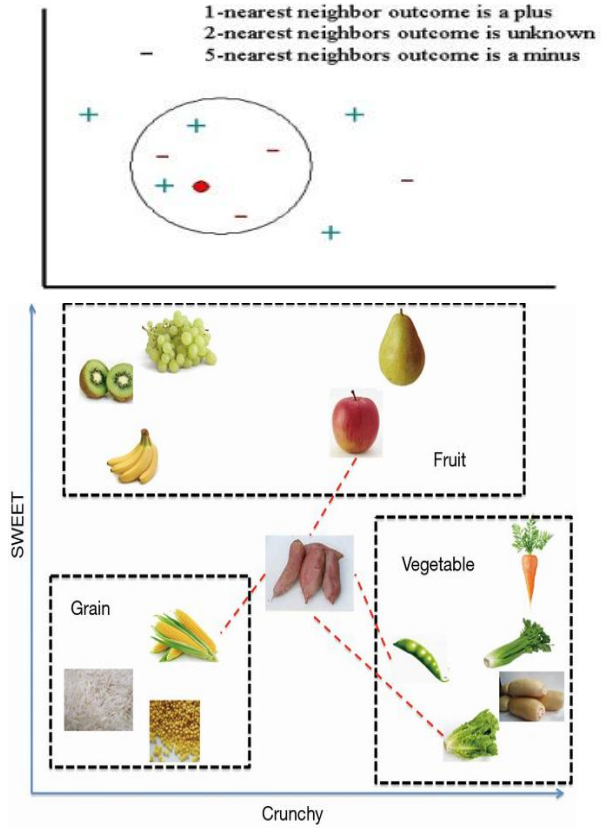


Fig.3 k-Nearest neighbour.

In k- NN classification, the majority vote of neighbors can be used for object classification. If k=1, then objects is assigned to the class of nearest neighbor and in k-NN regression the value is average values of its nearest neighbor. Nearest neighbor classifiers are instance based learners which do not build a classifier until a new sample needs to be classified while in eager learning methods such as decision tree, constructs a model before it receives new samples to be classified.

The k-nearest neighbor classification performance can be improved by supervised or metric learning. Transforming input data in to sets of features is called feature extraction which is done on data before applying k-NN algorithms and for high dimensional data dimension reduction is done prior to k-NN algorithms.

4. Bayesian networks

It is probabilistic graphical directed acyclic model that represents the variables and their dependencies through graph. The nodes represent the random variables which may be any unknown parameter or observable quantity while the edges represent the conditional dependencies. The information about each node is given through probability function which takes particular set of value as input and gives probability distribution of variables as output. The Bayesian network which models sequences of variables are known as dynamic Bayesian network. There are three main inference tasks of Bayesian network inferring unobserved variables, parameter and structure learning. The density of the arcs is the measure of the complexity. Simple model is represented by sparse bayesnet while complex models by dense bayesnet. Thus, it provides a flexible method for probabilistic modeling.

5. Neural Network

An artificial neural network (ANN), often just called a "neural network", is a computational model based on biological neural networks. It is a non-linear statistical data modeling tool. It consist of an interconnected group of modeling tools. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. The neurons within the network work together, in parallel, to produce an output function. They can be used to model complex relationships between inputs and outputs or to find patterns in data. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be either humans or other computer techniques.

Neural networks use a set of processing elements (or nodes) analogous to neurons in the brain. These processing elements are interconnected in a network that can then identify patterns in data once it is exposed to the data. This distinguishes neural networks from traditional computing programs that simply follow instructions in fixed sequential order neural networks can often produce very accurate predictions. However,

one of their greatest criticisms is the fact that they represent a "black-box" approach to research. They do not provide any insight in to the underlying nature of the phenomena. In most cases ANN is an adaptive system that changes its structure based on external or internal information that follows through the network during the learning phase.

IV. EXPERIMENTAL SETUP

We have used the popular, open-source data mining tool weka (version 3.6.8) for this analysis. The categorical data sets have been used and the performances of a comprehensive set of classification algorithms (classifiers) have been analyzed. The analysis has been performed on window 7 enterprise system with intel R core TM i5 CPU, 2.30 GHz processor and 3.00 GB RAM.

V. RESULTS AND EVALUATION

We have conducted or experiments on voting data, accuracy of different classification algorithms are measured on cross-validation method. There are435 instances of voting data are mined with various algorithms and results are shown graphically.

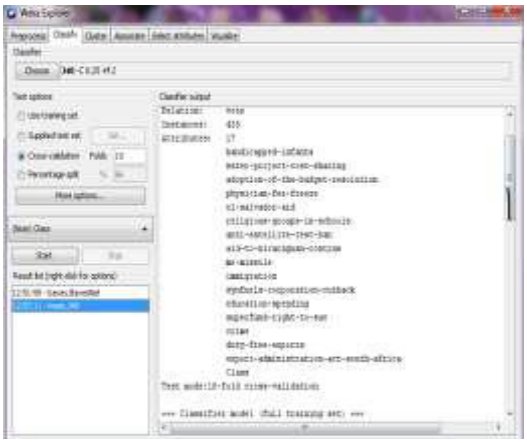
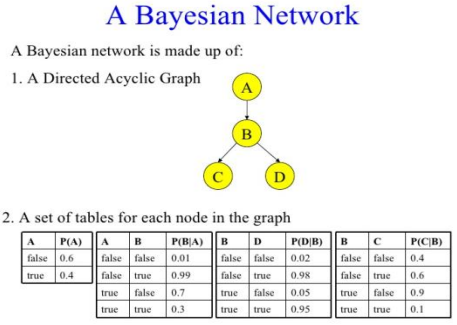


Fig. 4 Attributes of Voting Data**.**



A	P(A)
false	0.6
true	0.4

A	B	P(B|A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3

B	D	P(D|B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

Fig.5 Bayesian Network.

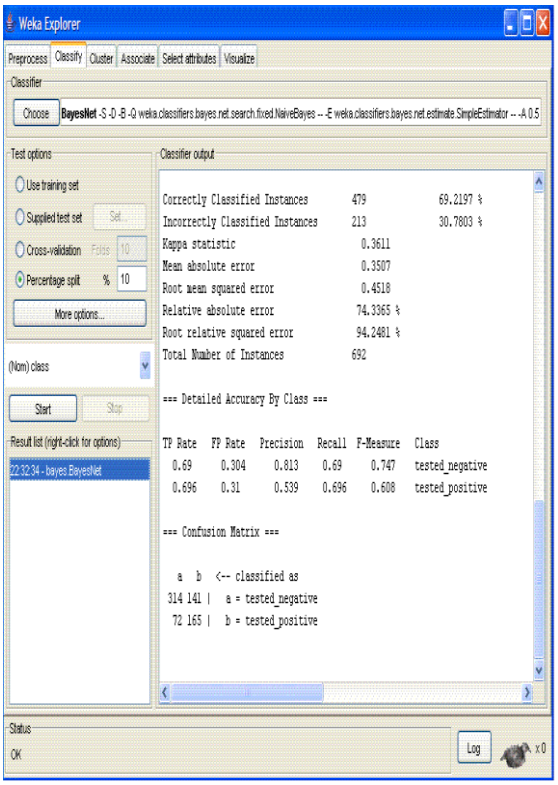


Fig. 6 Bayes Net.

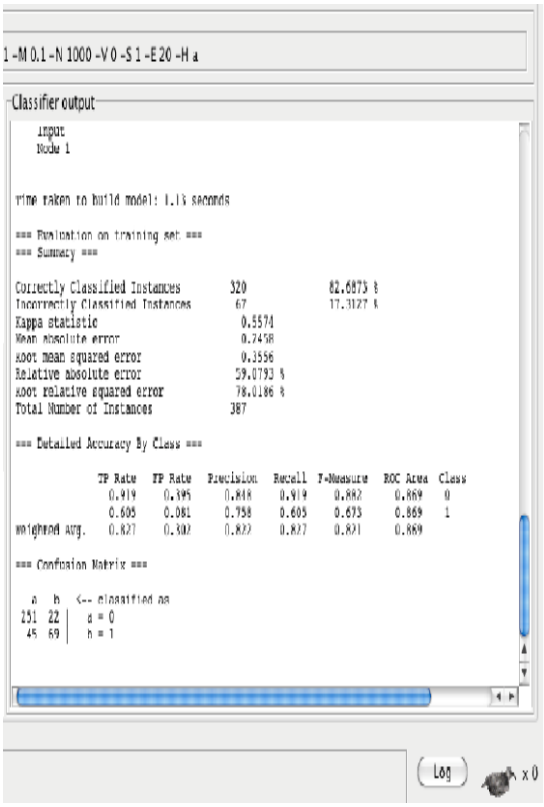


Fig. 7 Artificial Neural Network.

International Journal of Scientific Research & Engineering Trends
Volume 4, Issue 6, Nov-Dec-2018, ISSN (Online): 2395-566X

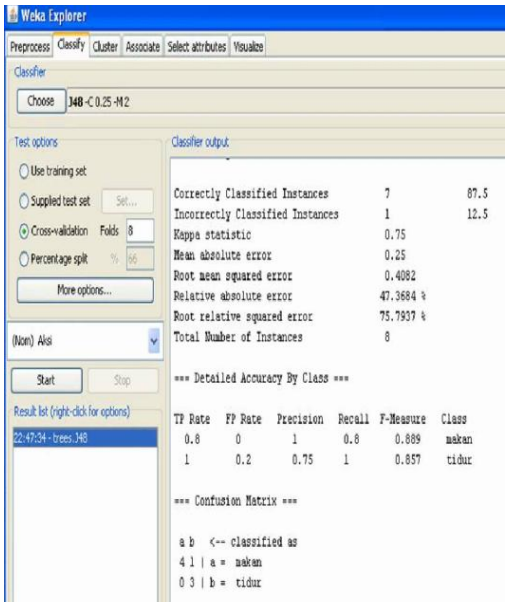


Fig. 8 Decision Tree.

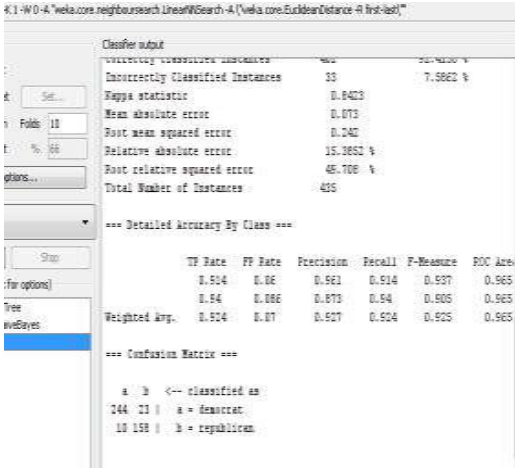


Fig. 9 k-Nearest Neighbor.

VI. DETAILS OF APPLIED CLASSIFIER

Table 1 Details of Applied Classifier

Classification techniques	Classifier
Decision tree	J48
Lazy learner	K star
Naïve bayes	Naïve
Rule based	Artificial Neural Network

Table 2 Comparative analysis.

Model	correctly classified data instances	Incorrectly classified data (in %)	Relative absolute error	Absolute mean Error	TP rate	FP rate
Decision tree	96.3218	3.6782	12.887	0.0611	0.97	0.048
Bayes Net	90.1149	9.8851	21.199	0.1005	0.891	0.083
k-nearest neighbor	92.4138	7.5862	15.3852	0.073	0.914	0.06
Artificial neural network	94.4828	5.5172	19.901	0.0944	0.94	0.048

The below graph shows classifiers accuracy values (in %)

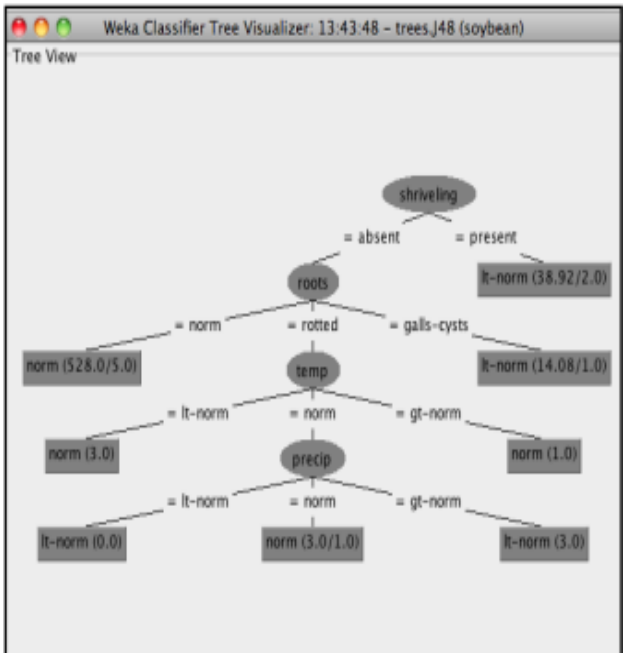


Fig.10 Decision tree form experiment.

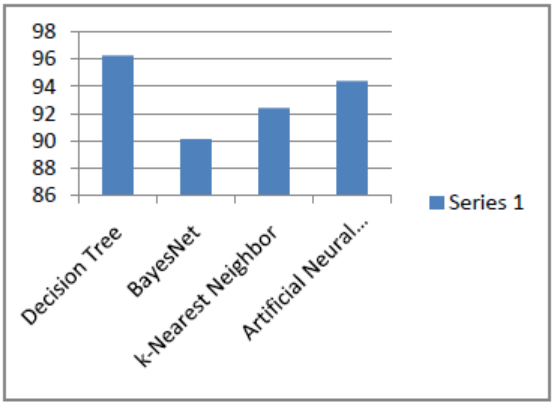


Fig. 11 Classifier Accuracy Values.

VII. CONCLUSION AND FUTURE WORK

This paper covers with various classification techniques used in data mining. Thus in this paper, performance of these algorithms are analyzed and evaluated by accuracy, ability to handle corrupted data and speed. Accuracy can be estimated by calculating error rate between the predicted value and the actual value. Accuracy of decision tree is better than other algorithms while using voting data set, cross validation method. The result of different classification algorithm are compared and to find which algorithm generates the effective results and graphically displays the results.

REFERENCES

[1]. International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169, Volume: 3 Issue: 8 5142 – 5147. A Comparative Analysis of Classification Techniques on Categorical Data in Data Mining, Sakshi, Department Of Computer Science And Engineering, United College of Engineering & Research, Naini Allahabad, sakshikashyap09@gmail.com, Prof. Sunil Khare,Department Of Computer Science and Engineering, United College Of Engineering and Research, Naini Allahabad, khare.sunil75@gmail.com.

[2]. Fabricio Voznika Leonardo Viana "DATA MINING CLASSIFICATION" Springer, 2001

[3]. S.Neelamegam, Dr.E.Ramaraj "Classification algorithm in Data mining: An Overview" International Journal of P2P Network Trends and Technology (IJPTT) – Volume 4 Issue 8- Sep 2013

[4]. SagarS. Nikam "A Comparative Study of Classification Techniques in Data Mining Algorithms" ISSN: 0974-6471 April 2015, Vol. 8, No. (1)

[5]. Anand V. Saurkar, Vaibhav Bhujade, Priti Bhagat, Amit Khaparde "A Review Paper on Various Data Mining Techniques" International Journal of Advanced Research in Computer Science and Software Engineering Volume 4,Issue 4, April 2014

[6]. W. D. M. Primitives, "Data Mining: Concepts and Techniques."

[7]. P. Harrington, Machine learning in action: Manning, 2012.