International Journal of Advanced Trends in Computer Applications
www.ijatca.com

Detailed Analysis of Classification Techniques in Data Mining

[1]Dr Rashmi Agrawal
[1]Associate Professor,
Manav Rachna International University

Abstract: *In classification, no prior information is required for predicting the class label. A classification technique is an organized approach for building classification model from a given input dataset. The learning algorithm of each technique is employed to build a model used to find the relationship between attribute set and class label of the given input data. Various classification techniques used are Decision Tree, Naïve Bayes, and Nearest Neighbour. k- Nearest Neighbour is one of simple and well known classification technique in which distance is measured between input point and all other records of the dataset. The class label of the k-Nearest Neighbour is the class label for input point. The objective of this paper is to understand and analyze various classification techniques used in data mining.*

Keywords: Classification, Decision Tree, Naïve Bayes, Neural Networks, Clustering, Association Rule**.**

I. INTRODUCTION

Data Mining is the extraction of interesting non-trivial, implicit, previously unknown and potentially useful patterns from huge amount of data. Data mining may be thought of as a consequence of the natural progression of information technology because traditional techniques may be unsuitable due to enormity, high dimensionality and heterogeneous distributed nature of data. Data can now be stored in many dissimilar databases and information repositories. One such data repository architecture that has emerged is the data warehouse. Data warehouse technology includes data cleaning, data integration, and on-line analytical processing (OLAP). Data mining is an essential step in the knowledge discovery from data (KDD) process [5]. This process consists of three steps as shown in fig 1:

- **Preprocessing**– In the first step, called Preprocessing, the data is cleaned, integrated, transformed and reduced.
- **Data Mining** – In this main step of KDD, interesting information or patterns are obtained from the data.
- **Post-processing-** In this step, discovered knowledge (information) is presented in the user required manner.

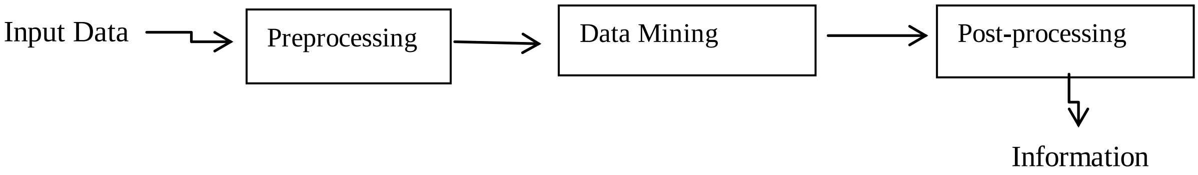


Fig 1: Process of Knowledge Discovery in Data (KDD)

II. DATA MINING TECHNIQUES

Data mining functionalities are used to specify the kind of patterns in data mining tasks [18]. To be more specific, data mining tasks can be classified into two broad categories, descriptive and predictive. Descriptive task are used to describe the properties of the data whereas predictive tasks are used to draw inference from the data. However, the present work is based on the predictive data mining used to predict the properties of unknown data. The following sub-section deal with the data mining techniques.

2.1 Association Rule Analysis

Association Rule Analysis is a method for discovering interesting relation between item sets in large databases [1]. Some items in a dataset frequently follow along with other items. Finding such patterns in a database is known as Frequent Pattern mining. There are two types of frequent patterns in a dataset, frequent sets (also called frequent item-sets) and frequent sequences. Frequent sets are unordered collection of items and frequent sequences are ordered collection of items. Association rule exclusively depends on the finding of frequent item-sets from the data. Algorithms are implemented to determine the set of frequent item-sets from the given transaction database. The algorithms attempt to eliminate the infrequent item-sets in the subsequent passes of finding the frequent item-sets. Once frequent item-sets are determined, then the associations between these items are analyzed by developing the association rules. In real life, the number of frequent item-sets is large in number which results in a large number of association rules for a transactional database. Hence discovery of frequent item-sets with item constraints is an important problem in association rule analysis. Association rule analysis has applications in cross-marketing, attached mailing, catalogue-designing, add-on sales, store layout, computer-aided detection systems, agriculture and recommendation systems.

2.2 Classification

A classification technique is an organized approach for building classification model from the given input dataset. Some of the well-known classification techniques are Neural Networks, Rule-based Classifiers, K-Nearest NeighbourClassifier, Decision Tree and Naïve Bayes Classifier. The learning algorithm of each technique is employed to build a model used to find the relationship between the attribute set and class label of the given input data. The key objective of the technique is to build a model with best generalization capability. Decision tress and rule-based classifiers are examples of eager learners or active learners because their model maps the input attributes to the class labels immediately in the training data. On the other hand, some classifiers delay this process until it is needed to classify. Such classification algorithms are known as lazy learners. Classification techniques are widely used in credit scoring, search engines, handwriting recognition, document categorization, speech recognition and medical image analysis and diagnosis.

2.3 Clustering

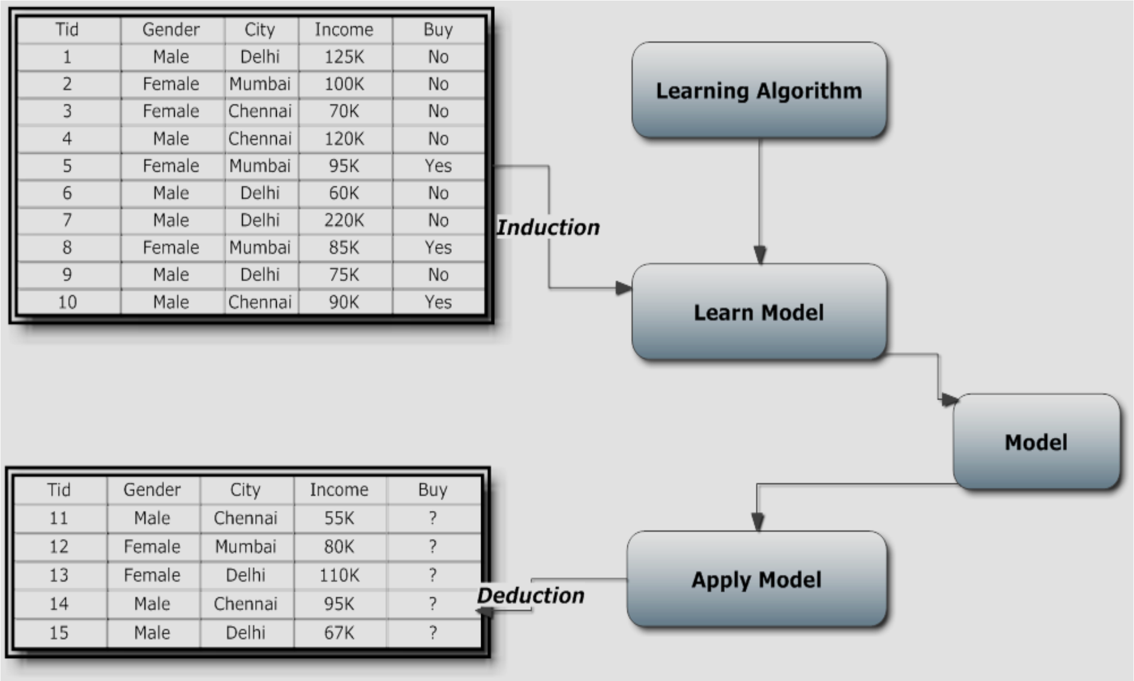
Clustering is an unsupervised learning technique in which class labels are not provided in test samples [8].It is the process of grouping a set of objects into classes of similar objects. Collection of similar data objects is known as a cluster. Similar objects are grouped into one cluster. Clustering is also named as data segmentation because it partitions large datasets into groups as per the similarity of records. Clustering based processes are adaptable to changes and help to draw features that are used to distinguish the different groups. To determine the similarity between two data objects, a metric from the datasets (distance function) is used by the clustering technique. This distance function takes two objects as input and returns the distance between these two objects as output in the form of a real number. Smaller value of this real number represents that two objects are more similar as compared to the larger value. Various commonly used clustering techniques are Partitioning Method, Hierarchical Method, Density-based Method and Grid-based Method. The major applications of clustering are targeting similar people and deciding on the locations for an activity like exam centres, locations for a business chain and planning a political strategy.

III. CLASSIFICATION TECHNIQUES

A classification technique is an organized approach for building classification model from given input dataset. The learning algorithm of each technique is employed to build a model used to find the relationship between attribute set and class label of the given input data. Classification is a two-step process. In the first step, known as learning step, a classifier is built describing a prearranged set of classes. The classification algorithm builds the classifier by learning from the training set. A tuple is represented by multiple attributes and a predefined class, called class label attribute. Generally, this label is categorical in nature and serves as a category or a class. These tuples are known as training tuples and dataset of training tuples is known as training dataset. In the second step, the classifier is applied on the testing dataset to perform the classification process.

Classification is also called as supervised learning because the class label of each training tuple is provided which contrasts with unsupervised learning or clustering in which the class labels of training tuples are unknown to the learning algorithm in advance. The key objective of the learning algorithm is to build models with good generalization capability. General approach for building a classification model is shown in figure2.

Evaluation of the performance of a classifier is based on the number of test records predicted correctly or incorrectly by the classification model. Confusion matrix is a popular matrix which shows the percentage of records of a class that were confused with other classes. Table 1 and 2 depicts respectively the confusion matrix for a binary classification problem and three class problems.



Tid	Gender	City	Income	Buy
1	Male	Delhi	125K	No
2	Female	Mumbai	100K	No
3	Female	Chennai	70K	No
4	Male	Chennai	120K	No
5	Female	Mumbai	95K	Yes
6	Male	Delhi	60K	No
7	Male	Delhi	220K	No
8	Female	Mumbai	85K	Yes
9	Male	Delhi	75K	No
10	Male	Chennai	90K	Yes

Tid	Gender	City	Income	Buy
11	Male	Chennai	55K	?
12	Female	Mumbai	80K	?
13	Female	Delhi	110K	?
14	Male	Chennai	95K	?
15	Male	Delhi	67K	?

Fig 2: General approach for classification

Table 1: Confusion Matrix for a binary classification problem

Predicted Class **Actual Class**	1	0
1	C_{11}	C_{10}
0	C_{01}	C_{00}

Table 2: Confusion Matrix for a three class problem

Predicted Class **Actual Class**	C1	C2	C3
C1	C_{11}	C_{12}	C_{13}
C2	C_{21}	C_{22}	C_{23}
C3	C_{31}	C_{32}	C_{33}

Each entry of the table 1 and2 depicts the number of records from class i predicted to be of class j. for example, C_{10} is the number of records from class 1 incorrectly predicted as class 0. On the other hand C_{00} represents the number of records from class 0 correctly predicted as class 0. From the confusion matrix we can find total number of correct predictions made by the classification model as (C_{11} + C_{00}) and total number of incorrect predictions as (C_{10} + C_{01}).

For a good classification model it is expected to have more number of records in cells of C11 and C00 and less number of records must lie in C01 and C10. The most popular performance metric to evaluate the merit of a classifier is the accuracy defined by-

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions}$$

$$= \frac{C_{11} + C_{00}}{C_{11} + C_{00} + C_{10} + C_{01}}$$

Similarly, to find the error rate of the classification model, we use the following equation-

$$Error\ Rate = \frac{C_{01} + C_{10}}{C_{11} + C_{00} + C_{10} + C_{01}}$$

The key objective of a classification model is to find the highest accuracy and lowest error rate.

In next sub sections, we briefly present an overview of various classification techniques.

3.1 Decision Tree

A decision tree is a simple recursive structure for expressing a sequential classification process in which a case described by a set of attributes is assigned to one of a disjoint set of classes [20]. Various efficient algorithms have been developed to construct a decision tree in a judicious amount of time. Generally, these algorithms (ID3, C4.5, CART, and SPRINT) follow the greedy approach by taking the local optimum decision. Usually information gain, entropy, gini index and other similar measures are used to take the decision of selecting the attribute in split point.

In case if there are two classes, P and N and the set S of examples contain p elements of class P and n elements of class N, then the information gain I is calculated as

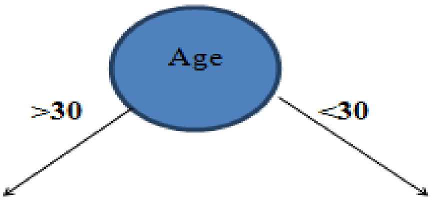
$$I(p,n) = -\frac{p}{p+n}log_2\frac{p}{p+n} - \frac{n}{p+n}log_2\frac{n}{p+n}$$

The classifier selects the attribute with the highest information gain and a decision tree is build. Table 3 shows a sample dataset with two attributes and one class label. First attribute contains the age and second attribute store the information related to car type (S= Small car, Su= SUV car, L= Luxury car). The goal of classification model is to predict whether a person will buy car perfume or not. The following example illustrates the classification using decision tree-

Table 3: Training Database for decision tree

Attribute	**Attribute**	**Class label**
Age	**Car Type**	**Buy**
20	Su	Yes
30	Su	Yes
25	S	Yes
40	S	No
20	L	Yes
40	Su	Yes
60	L	No
55	S	No
30	S	No

If age attribute is selected at the root node then the following decision tree is generated-



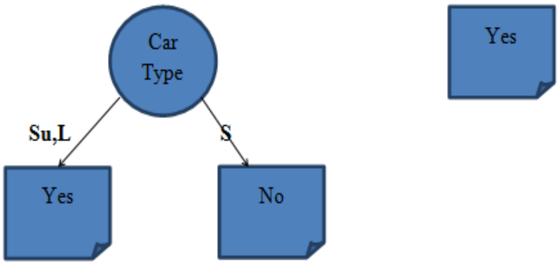


Fig 3: Decision tree of car sample dataset

3.2 Naïve Bayes

A Naïve Bayes classifier is a simple probabilistic classifier and works well for many applications, especially in text classification. Naïve bayes is a supervised and statistical learning method [19].This classifier is based on the Bayes' Theorem with naïve independence assumptions. A naïve baye's classifier assumes that the presence or absence of a feature of a class is not related to any other feature of the class. This method is popular due to many reasons. It is very easy to construct and does not need any complicated iterative parameter estimation schemes. Also, it may be applied to large data sets. It is easy to interpret and sometimes the outcomes of this classifier are surprisingly well.The basis for all Bayesian Learning Algorithms is the Bayes Rule given below-

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

Where:
P(h) = prior probability of hypothesis h
P(D)= prior probability of training data D
P(h|D) = probability of h given D
P(D|h) = probability of D given h

3.3 Neural Networks

A neural network is a computational model based on the biological neuron [6].It is abstracted as a directed graph, where the neurons represent the nodes and connections between them are edges. The weight on each edge represents the inhibiting or stimulating type and the strength of interaction between the neurons. The neural network is trained to respond with certain inputs to produce the desired output. Back propagation network and other neural network architectures are used in machine learning and pattern recognition applications.
Perceptron is the simplest type of neural network which consists of a single neuron with several real-valued or binary inputs and a binary output. Input in a neuron

comes through weighted edges and net input to the neuron is the sum of all the weighted inputs. Thus,

Net input= $w_1x_1 + w_2x_2 + \ldots\ldots + w_nx_n$,

where x_n represents input and w_n represents weight.
If this net input exceeds a threshold, then the neuron will be triggered to produce an output as 1 otherwise 0. Figure 4 shows the model of a perceptron.

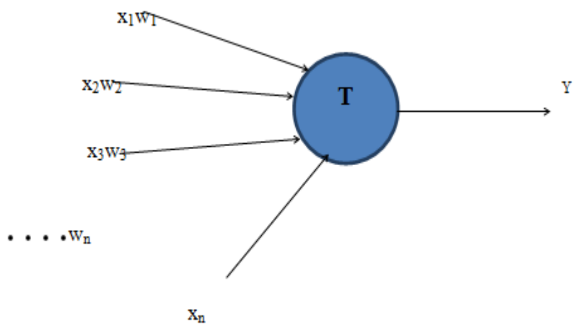


Fig 4: A Perceptron

3.4 Rule-Based Classifier

Classification of records is also done using a collection of "if….then" rules. Such classifiers are known as rule-based classifier [14].Each classification rule can be expressed as -

rule r_i := (condition$_i$) → C_i

The left hand side of a rule is called the rule precondition or rule antecedent and right hand side of the rule is called rule consequent which contains the predicted class C_i. Each antecedent contains a condition which is a conjunction of attribute tests: Each attribute test is known as a conjunct.Table 4 shows the sample dataset for computer purchase.

Table 4: Sample dataset for computer purchase

Age	Qualification	Purchase computer (Y/N)
22	Graduation	Y
30	Metric	N
40	Post-Graduation	Y
32	Post-Graduation	Y
60	Graduation	N

From this dataset rule-based classifier generates the rules as follows-

r_1: (Age between 30-40) ∧(qual = Post- Graduation)→ Y
r_2 : (Age between 20-60) ∧ (qual = Graduation)→ Y
r_3 : (Age =30) ∧(qual = Metric) → N

Quality of a rule is evaluated by using measures such as coverage and accuracy.

For a given dataset 'DS' and a classification rule r (A → C), we can define coverage of a rule as fraction of records in 'DS' that trigger the rule r.Accuracy, also known as confidence factor is defined as the fraction of records triggered by r whose class labels are C. Thus

$$Coverage(r) = \frac{|A|}{|DS|}$$

and

$$Accuracy(r) = \frac{|A \cap C|}{|A|}$$

where |A| is the number of records which satisfies the rule antecedent. |A∩C| is the number of records which satisfies both antecedent and consequent and |DS| is the number of records in a dataset 'DS'.
Generally rule based classifiers are used to produce descriptive models and the expressiveness of a rule almost correspond to decision tree.

3.5 k-Nearest Neighbour (k-NN)

k- Nearest Neighbour classifier (k-NN) is a simple classification technique which delays the process of modeling the training data until it is required to classify the examples. Therefore k-NN is also known as lazy learner. The k-NN algorithm follows non-parametric technique as it does not make any assumptions on data distribution passed to the algorithm [22, 23, and 25]. Despite its simplicity, the algorithm performs very well with the following prerequisites:

a) Set of stored records
b) A distance metric to compute the distance between records
c) Value of k i.e. number of nearest Neighbours

To classify an unknown record, the distance between other records are computed and based on this distance k-nearest Neighbours are selected to determine the class label of the unknown record. The figure 5 represents the 1-nearest, 2-nearest and 3-nearest Neighbour.

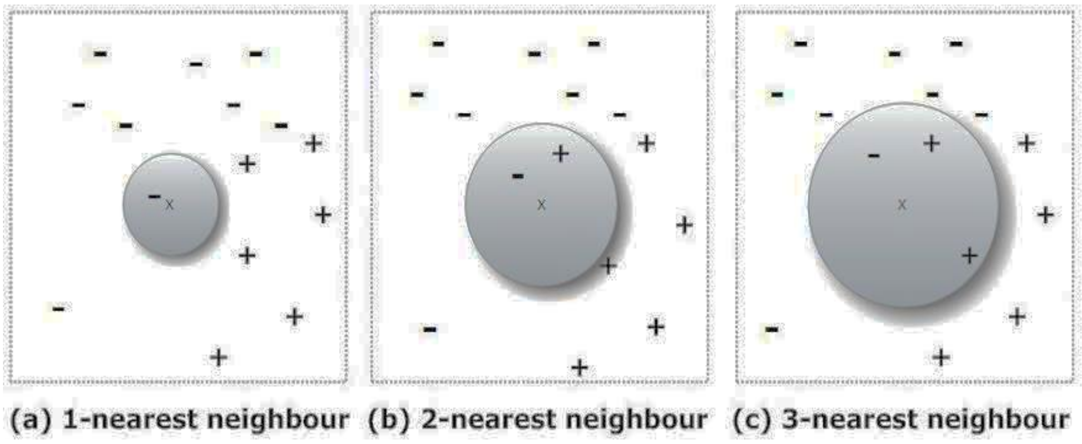


Fig 5: 1, 2 and 3 - Nearest Neighbour

The most common distance function to compute the distance between two records p and q is Euclidean distance d(p,q) defined by

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \quad ,$$

where p=(p_1, p_2,….p_n) and q= (q_1, q_2,………q_n).

The other popular distance functions which are also used by k-NN are Manhattan, Minkowski and cosine distance measures.
Consider the dataset of table5 for height and weight of the student studying in class 6 and 7. To determine the class of student with Roll no 106 in the table 5, the k-nearest Neighbour method will find the nearest Neighbours of this student.

Table 5: Sample Record of Student's Height and Weight

S No	**Roll No.**	**Height**(In Feet)	**Weight**	**Class**
1	101	4.7	50 kg	7
2	102	4.9	52 kg	7
3	103	4.5	40 kg	6
4	104	4.6	42 kg	6
5	105	4.9	50 kg	7
6	106	4.5	43 kg	?

We apply Euclidean function here and compute the distance of record number 6 with all other data records as follows:

$$d(1,6) = \sqrt{(h_1 - h_6)^2 + (w_1 - w_6)^2}$$
$$= \sqrt{(4.7 - 4.5)^2 + (50 - 43)^2}$$
= 7.02

Similarly,
d(2,6)= 9.00,
d(3,6)= 3,
d(4,6)= 1.00
and d(5,6)= 7.01

If we take k as 1, the nearest Neighbour is record number 4 with Rollno 104 having smallest distance hence the class label of unknown record will be determined as class 6.

REFERENCES

[1]. Agrawal, R and Srikant, R, 1994, September. 'Fast algorithms for mining association rules', In Proc. 20th int. conf. very large data bases, VLDB , vol. 1215, pp. 487-499.

[2]. Bhatia, N, 2010,'Survey of nearest Neighbour techniques', arXiv preprint arXiv:1007.0085.

[3]. Dudani, Sahibsingh A, 1976,'The distance-weighted k-nearest-Neighbour rule.' Systems, Man and Cybernetics, IEEE Transactions, pp 325-327.

[4]. Fix and Hodges, 1951,'Discriminatory analysis, nonparametric discrim- 15 ination: Consistency properties', Technical Report 4, U.S. Air Force, School of Aviation Medicine, Randolph Field, TX.

[5]. Han, Kamber, Data Mining: Concepts and Techniques, 2001, London: Academic Press

[6]. Haykin and Network, 2004, '*A comprehensive foundation. Neural Networks*', neural networks, vol 2, pp 696

[7]. Gou et al, 2012, 'A Local Mean-Based k-Nearest Centroid Neighbour Classifier',The Computer Journal, vol. 55, no. 9, pp. 1058-1071.

[8]. Jain, A and Dubes, R.C., 1988, 'Algorithms for clustering data', Prentice-Hall, Inc.

[9]. Keller et al, 1985,'A fuzzy k-nearest Neighbour algorithm' Systems, Man and Cybernetics, IEEE Transactions, pp.580-585.

[10]. Kira, Kenji, and Larry Rendell, 1992,'A practical approach to feature selection.', Proceedings of the ninth international workshop on Machine learning.

[11]. Langley Pat, 1994,'Selection of relevant features in machine learning', Proceedings AAAI Fall Symposium on Relevance, New Orleans, LA , pp. 140–144.

[12]. Leung et al, 2008,'A tree-based approach for frequent pattern mining from uncertain data', Proceedings of the 12th Pacific-Asia Conference on Knowledge Discovery and

Data Mining (PAKDD),Osaka, Japan, pp. 653–661..

[13]. Liu, W, Chawla, S, 2011,‘Class Confidence Weighted KNN Algorithms for Imbalanced Data Sets’,Advances in Knowledge Discovery and Data Mining Lecture Notes in Computer Science, vol6635, no 11, pp. 345-356.

[14]. Liu, B et al, 2000,‘Improving an association rule based classifier’, In Principles of data mining and knowledge discovery, Springer Berlin Heidelberg., pp. 504-509.

[15]. MSudha and Kumaravel, 2014, ‘Performance Comparison based on Attribute Selection Tools for Data Mining’, Indian Journal of Science and Technology, Nov, Vol 7 no 7.

[16]. Marchiori Elena, 2013,‘Class dependent feature weighting and k-nearest Neighbour classification’, Pattern Recognition in Bioinformatics. Springer Berlin Heidelberg, pp 69-78.

[17]. Pawlak, Zdzislaw, 1998,‘Rough set theory and its applications to data analysis’ Cybernetics& Systems, vol 29, no 7, pp 661-688.

[18]. Pujari, Data mining techniques, 2001, Universities Press, Hyderabad.

[19]. Rish, I.,August 2001,. ‘An empirical study of the naive Bayes classifier’, InIJCAI 2001 workshop on empirical methods in artificial intelligence,IBM New York, vol. 3, No. 22, pp. 41-46.

[20]. Safavian, S.R and Landgrebe, D., 1991,‘A survey of decision tree classifier methodology’, IEEE transactions on systems, man, and cybernetics,vol21, no 3, pp.660-674.

[21]. Sarkar, Manish., 2007, ‘Fuzzy-rough nearest Neighbour algorithms in classification’ Fuzzy Sets and Systems, vol 158, no 19, pp 2134-2152.

[22]. Suguna, N and Thanushkodi, K., 2010,‘An improved K-nearest Neighbour classification using Genetic Algorithm’, International Journal of Computer Science Issues, vol7, no 2, pp.18-21.

[23]. Weinberger et al, 2005,‘Distance metric learning for large margin nearest Neighbour classification’, In Advances in Neural Information Processing Systems , pp. 1473-1480.

[24]. Lian and Chen, 2009, 'Efficient processing of probabilistic reverse nearest Neighbour queries over uncertain data', The VLDB Journal, vol. 18, no. 3, pp. 787-808.

[25]. Zhang, M.L and Zhou, Z.H., July 2005, ‘A k-nearest Neighbour based algorithm for multi-label classification’, In Granular Computing, IEEE International Conference, Vol. 2, pp. 718-721.