International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No. (02), January 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

Performance evaluation ofDecision Tree Techniques for Predictive Model in Healthcare Sector

Sanjeev Kumar Sahoo[1], Subhendu Kumar Pani[2], Narayana Patra[3]

[1]*Assistant Professor,PJCMT BPUT, Odisha (India)*

[2]*Associate Professor, Dept. of CSE, OEC, BPUT, Odisha (India)*

[3]*Assistant Professor,Dept. of CSE, SOA University (India)*

ABSTRACT

Machine learning algorithms are techniques that automatically build models describing the structure at the heart of a set of data. Ideally, such models can be used to predict properties of future data points and people can use them to analyze the domain from which the data originates. Decision trees are accurate classifiers that are easy to understand. However, in some domains their comprehensibility suffers from a problem known as sub tree replication. When sub tree replication occurs, identical sub trees can be found at several different places in the same tree structure. The main objective of this research is to prove a set of simple decision tree algorithms that should be useful in practical data mining applications. In this paper, effort has made to compare between few decision tree algorithms such as: Random Forest, Random Tree and C 4.5 etc using Lung cancer datasets. Our main aim to show the comparison of the different decision tree algorithms and find out which algorithm will be most suitable for the Cancer data.

Keywords: Random Forest, Random Tree, C 4.5, Decision trees

I. DATA MINING: AN OVERVIEW

Data mining is the process of using variety of data analysis tools to discover patterns and relationships in data that maybe used to make valid predictions. Kumer and Zaki [1,3,4] define it as *"the iterative and interactive process of discovering valid, novel, useful, and understandable patterns or models in massive databases"*. Data mining concept has been around for some time now. The techniques have existed for years or decades as academic algorithms in the fields of statistics and machine learning. These fields have been working on problems related to pattern recognition and classification, which are tasks now embedded into data mining.

There are various data mining techniques available in carrying out knowledge extractions from large databases. These could be classified into two main categories: "Descriptive" and "Predictive" [1]. The descriptive is concerned with explanatory models that summarize data for the purpose of inference. Summarization and Visualization of databases are the main applications of descriptive data mining. The usefulness of this concept is that it enables one to generalize the data set from multiple levels of abstraction, which facilitates the examination of the general behavior of the data, since it is impossible to deduce that from a large database. The

predictive, on the other hand, is concerned with the creation of models that are capable of producing prediction results when applied to unseen, future cases. Classification is the most frequent type of task that is applied in data mining. Data mining has been used in various areas like health care, business intelligence, financial trade analysis, network intrusion detection etc. [2,12,14].

Data Mining is an iterative process consists of the following list of stages:

i. Data cleaning
ii. Data integration
iii. Data selection
iv. Data transformation
v. Data mining
vi. Pattern evaluation
vii. Knowledge presentation

II. DATA CLASSIFICATION AND PREDICTION

Following are the examples of cases where the data analysis task is Classification −

- A bank loan officer wants to analyze the data in order to know which customer (loan applicant) is risky or which are safe.
- A marketing manager at a company needs to analyze a customer with a given profile, who will buy a new computer.

In both of the above examples, a model or classifier is constructed to predict the categorical labels [13,15,17]. These labels are risky or safe for loan application data and yes or no for marketing data. With the help of the bank loan application that we have discussed above, let us understand the working of classification. The Data Classification process includes two steps −Building the Classifier or Model, Using Classifier for Classification[18,19,20].

Following are the examples of cases where the data analysis task is Prediction[21,22,23,24,25]

−Suppose the marketing manager needs to predict how much a given customer will spend during a sale at his company. In this example we are bothered to predict a numeric value. Therefore the data analysis task is an example of numeric prediction. In this case, a model or a predictor will be constructed that predicts a continuous-valued-function or ordered value.

- This step is the learning step or the learning phase.
- In this step the classification algorithms build the classifier.
- The classifier is built from the training set made up of database tuples and their associated class labels.

Each tuple that constitutes the training set is referred to as a category or class. These tuples can also be referred to as sample, object or data points.

III. DECISION TREE

A decision tree is a structure that includes a root node, branches, and leaf nodes as show in figure. Each

internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label [26, 27, 28, 29,30]. The topmost node in the tree is the root node. Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy). Leaf node (e.g., Play) represents a classification or decision. The topmost decision node in a tree, which corresponds to the best predictor, called root node. Decision trees can handle both categorical and numerical data [31, 32].

Decision Tree Advantage: • Decision trees implicitly perform variable screening or feature selection. • Decision trees require relatively little effort from users for data preparation. • Nonlinear relationships between parameters do not affect tree performance. • The best feature of using trees for analytics - easy to interpret and implement.

Decision Tree Algorithm

Decision tree algorithm recursively partitions a data set of records using depth-fist greedy approach [5] or breadth-first approach, until all the data items belong to a particular class are identified. A decision tree structure is made of root, internal and leaf nodes. Most decision tree classifiers perform classification in two phases: tree-growing (or building) and tree-pruning. The tree building is done in top-down manner. During this phase, the tree is recursively partitioned till all the data items belong to the same class label. In the tree pruning phase the full grown tree is cut back to prevent over fitting and improve the accuracy of the tree [10] in bottom up fashion. It is used to improve the prediction and classification accuracy of the algorithm by minimizing the over-fitting(noise or much data in training data set).Decision tree algorithm structure is given in two phases as under BuildTree (data set S) if all records in S belong to the same class; return; for each attribute Ai; evaluate splits on attribute Ai; use best split found to partitionS into S1 and S2 BuildTree (S1); BuildTree (S2); EndBuildTree

Algorithm for decision tree growth phase

PruneTree
(node t) if t is
leaf
return C(S) +1
/* C(S) is the cost of encoding the classes for therecords in set S */
minCost1:= PruneTree (t1);
minCost2:= PruneTree (t2);
/* t1, t2 are t'schildren*/
minCost t:= min{ C(S)+1, Csplit (t)+1+minCost 1+minCost2};
return minCostt;
/* C split: cost of encodinga split
*/ EndPruneTree

ID3

ID3(Iterative Dichotomized) algorithm is based on the Concept Learning System (CLS) algorithm.CLS algorithm is the basic algorithm for decision tree learning. The tree growth phase of CLS is the matter of choosing attribute to test at each node is by the trainer. ID3 improves CLS by adding a heuristic for attribute selection. ID3 is based on Hunt's algorithm and is implemented in serially[7,8]. This algorithm recursively partitions the training dataset till the record sets belong to the class label using depth first greedy technique. In growth phase of the tree construction, this algorithm uses information gain, an entropy based measure, to select the best splitting attribute, and the attribute with the highest information gain is selected as the splitting attribute. ID3 doesn't give accurate result when there is too much noise or details in the training data set, thus an intensive pre-processing of data is carried out before building a decision tree model with ID3[6]. One of the main drawbacks of ID3 is that the measure Gain used tends to favor attributes with a large number of distinct values[8]. It only accepts categorical attributes in building a tree model. This decision tree algorithm generates variable branches per node.

2.4. C4.5

C4.5 algorithm is an improved version of ID3, this algorithm uses Gain Ratio as a splitting criteria, instead of taking gain in ID3 algorithm for splitting criteria[9] in tree growth phase. Hence C4.5 is an evolution of ID3[10]. This algorithm handles both continuous and discrete attributes- In order to handle continuous attributes, C4.5 creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it[31]. Like ID3 the data is sorted at every node of the tree in order to determine the best splitting attribute. The splitting ceases when the number of instances to be split is below a certain threshold. The main advantages of C4.5 is when building a decision tree, C4.5 can deal with datasets that have patterns with unknown attribute values. C4.5 can also deal with the case of attributes with continuous domains by discretization. This algorithm handles training data with attribute values by allowing attribute values to be marked as missing. Missing attribute values are simply not used in gain and entropy calculations. It has an enhanced method of tree pruning that reduces misclassification errors due to noise or too much detail in the training data set.

IV. COMPARISION AND RESULT ANALYSIS

WEKA Tool

We use WEKA (www.cs.waikato.ac.nz/ml/weka/), an open source data mining tool for our experiment. WEKA is developed by the University of Waikato in New Zealand that implements data mining algorithms using the JAVA language. WEKA is a state-of-the-art tool for developing machine learning (ML) techniques and their application to real-world data mining problems. It is a collection of machine learning algorithms for data mining tasks. The algorithms are applied directly to a dataset. WEKA implements algorithms for data pre-processing, feature reduction, classification, regression, clustering, and association rules. It also includes visualization tools. The new machine learning algorithms can be used with it and existing algorithms can also be extended with this tool.

Classifier Selection

We select some commonly used Decision tree classifiers for prediction classification in our work based on their qualitative performance. These classifiers are described in this section

Performance Measure

We use different metrics for comparing the classifiers' predictive performance in our experiment. These are presented below:

Confusion Matrix: The columns of the confusion matrix represent the predictions, and the rows represent the actual class. Correct predictions always lie on the diagonal of the matrix. Given below is the general structure of confusion matrix.

TP FN

FP TN

wherein, True Positives (TP) indicate the number of instances of the minority that were correctly predicted, True Negatives (TN) indicate the number of instances of the majority that were correctly predicted. False Positives (FP) indicate the number of instances of the majority that were incorrectly predicted as minority class instances and False Negatives (FN) indicate the number of the minority that were incorrectly predicted as majority class instances. Though the confusion matrix gives a better outlook on how the classifier performed than accuracy, a more detailed analysis is preferable which are provided by the further metrics.

Recall: Recall is a metric that gives a percentage of how many of the actual minority class members the classifier correctly identified. (TP + FN) represent a total of all minority members. Recall is given below

Recall = TP / (TP + FN)

Precision: It gives us the total the percentage of how many of minority class instances as determined by the model or classifier actually belong to the minority class. (TP + FP) represents the total of positive predictions by the classifier.

Precision is given by

Precision = TP / (TP + FP)

Thus in general it is said that Recall is a Completeness Measure and Precision is an Exactness Measure. The ideal classifier would give value as 1 for both Recall and Precision but if the classifier gives higher (closer to one) for one of the above metrics and lower for the other metrics in that case choosing the classifier is difficult task. In such cases some other metrics as discussed further are suggested in the literature.

F-Measure: It is a harmonic mean of Precision & Recall. We can say that it is essentially an average between the two percentages. It really simplifies the comparison between the classifiers. It is given by

F-Measure = 2 / (1/Recall + 1/Precision)

Dataset Description

We performed computer simulation on a Lung Cancer Dataset available UCI Machine Learning Repository[11]. The detailed description of dataset is shown in Fig.3.1.

International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No. (02), January 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

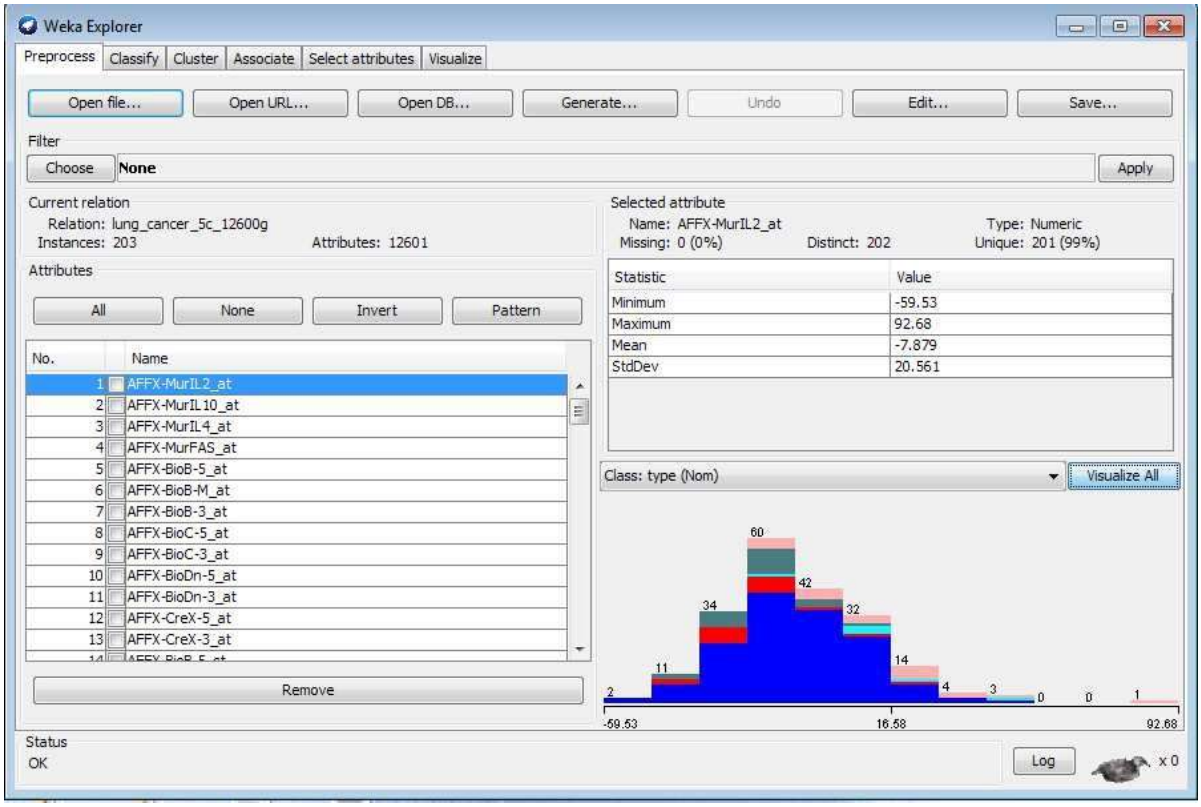

Figure-3.1: Lung Cancer Data Set

Experiment Design

In the study, we use Weka data mining tool to conduct the experiment. We compared the Decision Tree classifier's performance of the chosen models. We use 10-fold cross validation as the test mode to record classification accuracy. This approach is suitable to avoid biased results and provide robustness to the classification. Also, the parameters of a classification algorithm are chosen to their default values.

The following steps have been applied to generate experimental data in order to draw inference:

1. Find the Classifier's performance of the Decision Tree classifiers with original features in the dataset.

Results Analysis

Following the experimental procedures described in the previous section, we performed several runs in Weka tool and gathered the data for the inference. Table-3.1 summarizes the classification accuracy in percentage of all the classifiers across the dataset with original features. Decision Tree Classifiers[33,34,35,36,37,38] in WEKA is shown in Fig 3.2.

International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No. (02), January 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

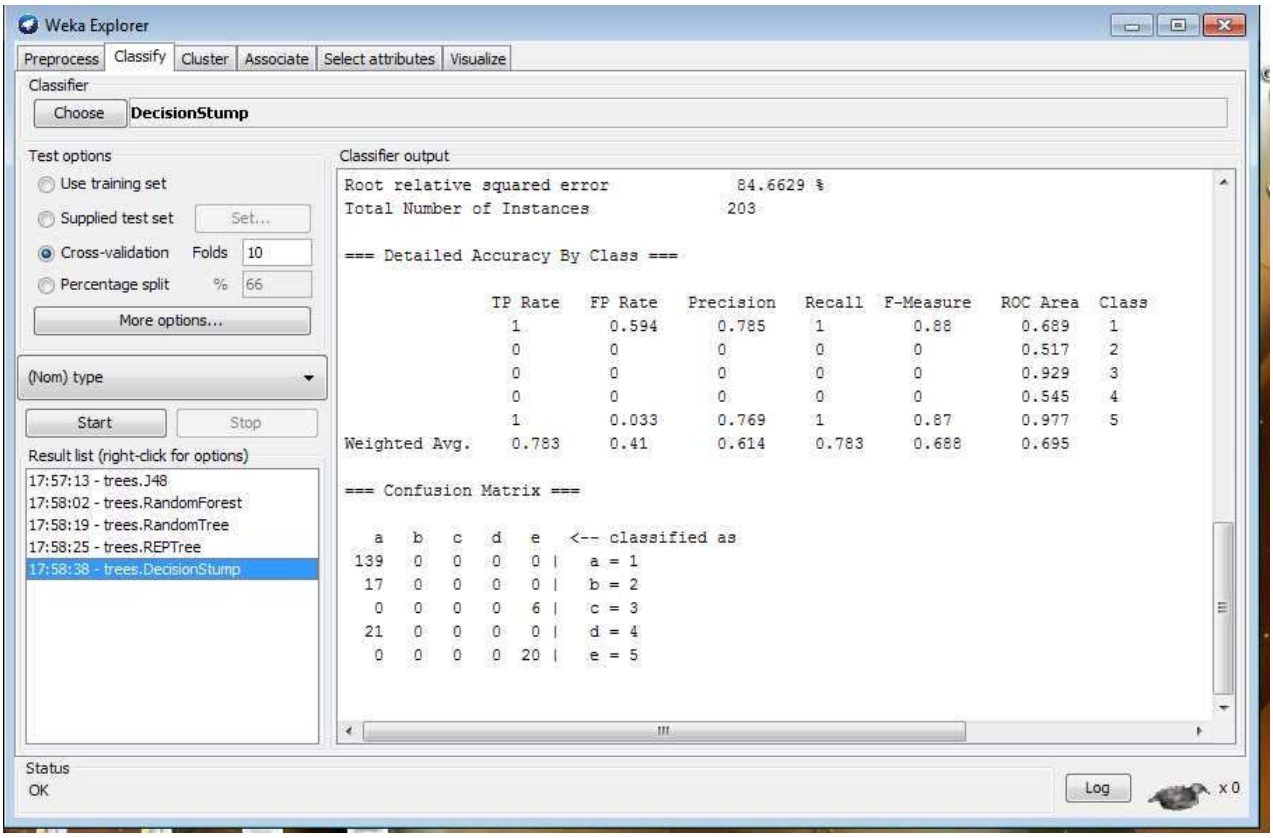

Figure-3.2: Snapshot of Decision Tree Classifiers in WEKA

Decision Tree Classifier's Name	ROOT Mean Squared Error	F-Measure	ROC Area	Accuracy	Time taken to build model
J48 pruned tree	0.1655	0.929	0.93	93.1034	8.53
Random forest	0.2075	0.834	0.952	85.7143	0.36
RandomTree	0.3232	0.74	0.74	73.8916	0.06
REPTree	0.2253	0.847	0.884	85.2217	4.62
Decision	0.2686	0.688	0.695	78.3251	2.44

Stump					

Table-3.1: Classification Accuracy in % with Original Features

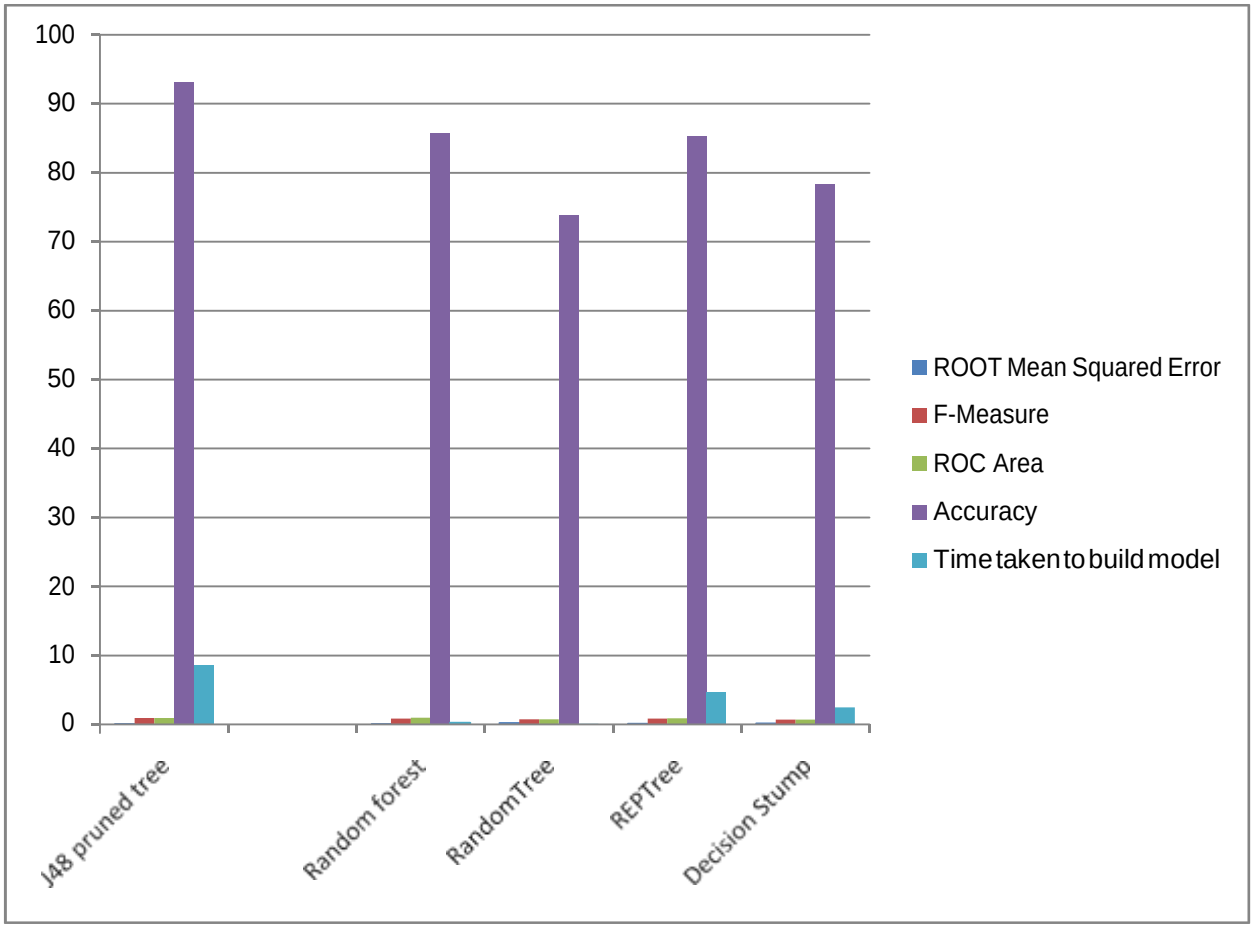

Figure-3.3: Classifiers Performance

Given the dataset, it is evident from the Figure-3.3 that the performance of the classifiers on feature reduction (or selection) is not uniform. We tested the performance of Classifiers on a Lung Cancer Dataset from UCI repository. It is observed in the tabulated data that the performance of all the classifiers is not linear across the datasets on the feature selection. The J48 Pruned tree classifiers perform better than all other remaining classifiers. This is depicted in Figure-3.3.

V. CONCLUSION AND FUTURE WORK

Decision tree is a simple representation for classifying examples. Decision tree learning is one of the most successful techniques for supervised classification learning. In this work we will discuss the various algorithms that are used for classification of data using Decision tree algorithm like ID3, Improved ID3 based on attribute importance, based on weight, based on attribute importance and weight which will be used for generating the tree. Data Classification important activity in data mining techniques. This thesis attempts

to survey this fast developing field, show some effective applications, and point out interesting trends and challenges. We conducted an experiment to compare five most commonly used classification models to classify the data taken from UCI machine learning repository. The predictive performance was recorded quantitatively using popular Decision tree techniques .The experimental data showed The J48 Pruned tree classifiers perform better than all other remaining classifiers. we analyzed the performance of most popular decision tree classifiers. The experiment considered a single dataset. We propose to extend our work by considering multiple datasets drawn from different domains, so that the results will be sound enough for generalization.

REFERENCES

[1.] Mustafa Agaoglu, "Predicting Instructor Performance Using Data Mining Techniques in Higher Education," IEEE Access ,Volume: 4 ,2016.

[2.] Tripti Mishra,Dr. Dharminder Kumar,Dr. Sangeeta Gupta,"Mining Students' Data for Performance Prediction," in fourth International Conference on Advanced Computing & Communication Technologies,2014.

[3.] Keno C. Piad, Menchita Dumlao, Melvin A. Ballera, Shaneth C. Ambat," Predicting IT Employability Using Data Mining Techniques," in third International Conference on Digital Information Processing, Data Mining, and Wireless Communications (DIPDMWC), 2016.

[4.] Bipin Bihari Jayasingh,"A Data Mining Approach to Inquiry Based Inductive Learning Practice In Engineering Education," in IEEE 6th International Conference on Advanced Computing,2016.

[5.] . S. M. Merchán,"Analysis of Data Mining Techniques for Constructing a Predictive Model for Academic,"IEEE Latin America Transactions, vol. 14, no. 6, June 2016.

[6.] Konstantina Chrysafiadi and Maria Virvou," Fuzzy Logic for adaptive instruction in an e-learning environment for computer programming," IEEE Transactions on Fuzzy Systems ,Volume: 23, Issue: 1, Feb. 2015.

[7.] M. Mayilvaganan,D. Kalpanadevi ," Comparison of Classification Techniques for predicting the performance of Students Academic Environment," in International Conference on Communication and Network Technologies (ICCNT), 2014.

[8.] Han J. and Kamber M., Data mining concept and techniques, Morgan Kaufmann Publishers, London, 2001.

[9.] Klosgen W and Zytkow J M (eds.), Handbook of data mining and knowledge discovery, OUP, Oxford, 2002.

[10.] Liu, H. Feature Extraction, Construction and Selection: A Data Mining Perspective, ISBN0-7923-8196-3, Kluwer Academic Publishers, 1998.

[11.] UCI Machine Learning Repository, Available at http://archive.ics.uci.edu/ml/machine- learning-databases/statlog/german/.

[12.] Kantardzic M., Data Mining: Concepts, Models, Methods, and Algorithms, Wiley, 2003.

[13.] Berry, M. and Linoff, G., Data mining techniques, Wiley Publishing, Inc, 2004.

[14.] Lui, H. Li, J. and Wong, L., A Comparative Study on Feature Selection and Classification Methods Using Gene Expression Profiles and Proteomic Patterns. Genome Informatics Vol. 13 p51-60, 2002.

[15.] Caruana, R. and Mizil, A. N, Data Mining in Metric Space: An Empirical Analysis of Supervised Learning Performance Criteria Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. p69 -78, 2004.

[16.] Caruana, R. and Mizil, A. N., An empirical comparison of supervised learning algorithms. Proceedings of the 23rd international conference on machine learning ICML, 2006.

[17.] S. Das, "Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection," Proc. 18th Int'l Conf. Machine Learning, pp. 74-81, 2001.

[18.] P. Langley, "Selection of Relevant Features in Machine Learning," Proc. AAAI Fall Symp. Relevance, pp. 140-144, 1994.

[19.] H. Almuallim and T.G. Dietterich, "Learning Boolean Concepts in the Presence of Many Irrelevant Features," Artificial Intelligence,vol. 69, nos. 1-2, pp. 279-305, 1994.

[20.] M. Ben-Bassat, "Pattern Recognition and Reduction of Dimensionality," Handbook of Statistics-II, P.R. Krishnaiah and L.N. Kanal, eds., pp. 773-791, North Holland, 1982.

[21.] M.A. Hall, "Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning," Proc. 17th Int'l Conf. Machine Learning, pp. 359-366, 2000.

[22.] H. Liu and H. Motoda, Feature Selection for Knowledge Discovery and Data Mining. Boston: Kluwer Academic, 1998.

[23.] U.M. Fayyad and R. Uthurusamy, "Evolving Data Mining into Solutions for Insights," Comm. ACM, vol. 45, no. 8, pp. 28-31, 2002.

[24.] L. Parsons, E. Haque, and H. Liu, "Subspace Clustering for High Dimensional Data: A Review," SIGKDD Explorations, vol. 6, no. 1, pp. 90-105, 2004.

[25.] E. Leopold and J. Kindermann, "Text Categorization with Support Vector Machines. How to Represent Texts in Input Space?" Machine Learning, vol. 46, pp. 423-444, 2002.

[26.] K. Nigam, A.K. Mccallum, S. Thrun, and T. Mitchell, "Text Classification from Labeled and Unlabeled Documents Using EM," Machine Learning, vol. 39, 103-134, 2000.

[27.] W. Lee, S.J. Stolfo, and K.W. Mok, "Adaptive Intrusion Detection: A Data Mining Approach," AI Rev., vol. 14, no. 6, pp. 533-567, 2000.

[28.] E. Xing, M. Jordan, and R. Karp, "Feature Selection for High-Dimensional Genomic Microarray Data," Proc. 15th Int'l Conf. Machine Learning, pp. 601-608, 2001.

[29.] L. Yu and H. Liu, "Redundancy Based Feature Selection for Microarray Data," Proc. 10th ACM SIGKDD Conf. Knowledge Discovery and Data Mining, 2004.

[30.] Y. Yang and J.O. Pederson, "A Comparative Study on Feature Selection in Text Categorization," Proc. 14th Int'l Conf. Machine Learning, pp. 412-420, 1997.

[31.] D.L. Swets and J.J. Weng, "Efficient Content-Based Image Retrieval Using Automatic Feature Selection," IEEE Int'l Symp.Computer Vision, pp. 85-90, 1995.

[32.] Y. Rui, T.S. Huang, and S. Chang, "Image Retrieval: Current Techniques, Promising Directions and Open Issues," Visual Comm. and Image Representation, vol. 10, no. 4, pp. 39-62, 1999.

[33.] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. Springer, 2001.

[34.] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation- Based Filter Solution," Proc. 20th Int'l Conf. Machine Learning, pp. 856-863, 2003.

[35.] W.G. Cochran, Sampling Techniques. John Wiley & Sons, 1977.

[36.] H. Liu, H. Motoda, and L. Yu, "Feature Selection with Selective Sampling," Proc. 19th Int'l Conf. Machine Learning, pp. 395-402, 2002.

[37.] H. Liu, L. Yu, M. Dash, and H. Motoda, "Active Feature Selection Using Classes," Proc. Seventh Pacific-Asia Conf. Knowledge Discovery and Data Mining, pp. 474-485, 2003.

[38.] P. Smyth, D. Pregibon, and C. Faloutsos, "Data-Driven Evolution of Data Mining Algorithms," Comm. ACM, vol. 45, no. 8, pp. 33-37, 2002.