IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-5, Issue-6 E-ISSN: 2347-2693

Analysis of Classification Approach for Data of Social Network

Anupama Tyagi[1*],Sanjiv Sharma[2]

[1*]Department of CSE &IT,Madhav institute of technology and Science, Gwalior, India
[2]Department of CSE & IT,Madhav institute of technology and Science, Gwalior ,India

[*]*Corresponding Author: anu.tyagi306@gmail.com*

Available online at: www.ijcseonline.org

Received: 17/May/2017, Revised: 28/May/2017, Accepted:14/Jun/2017, Published: 30/Jun/2017

***Abstract*—** Recently, Social networks are linked to the one another which are made of actors called as node and a variety of social familiarities of relationships are offered by edges. Classification is essential approach with the broad applications to categorize the different types of data needed in nearly all type of field of the life. Classification used for classifying the item according to features of item w. r. t the classes which are predefined. In the classification tree modeling data is being classified to do the predictions about the data which is new. This type of paper describes use of the classification trees and the shows three methods of pruning them. An experiment has set up using the various types of the algo which is classification tree algorithms having various methods of pruning to test performance of algorithm and the method of pruning . Here paper is about analyzing about the properties of data set to search relations among them. MEKA used inside the implementation of proposed work which provide the result and show that our work is much better than existing work.

***Keywords*—**Social network,classifier, MEKA, Random forest tree, J48, REP tree

I. INTRODUCTION

A social network where a group formed by individuals, and is connected by some characteristics based on some associations, friendly relationship, family connections, similar tastes in arts, literature, likes and the dislikes of beliefs and knowledge. Such a network can be made of web pages, citations, collaborations, neurons, proteins and dynamic network analysis (DNA) sequences. Participants in many Sites in Social Networking not necessarily want to look out for connecting to new people. But they need to communicate with individuals who are part of their enlarged network earlier. So by giving importance to this articulating feature these sites are named as Social Network Sites [1].Ramsey theorem says party with at least 6individuals, there are 3 individuals who are either type of mutual acquaintances or mutual strangers.

Different entities and their relationships that are among them form a network and that could be drawn as a graph. There are millions of the people socializing over internet finding other people with same type of interests in the hobbies, religion, or politics. They socialize on sites by reading profile pages of the other members and contacting them. The Social networks could be drawn as graphs. Further networks could be formed from protein

Folding, combination of genes, citations, there are the neural networks when nodes represent neurons, worldwide web networks when nodes represent uniform resource locators, collaboration networks when node represents individual actor. Whatever may be the network it can be analysed with the help of graphs? A graph is a collection of nodes (vertices) and edges (links or ties).

Fig.1 Social Network

Each node may represent individual, state, organization, workgroup and household. Network can be one mode or two mode networks. Social Network Analysis (SNA) helps in perceiving and investigating these relationships through visual and mathematical procedures. Organizations utilize these relationships and make Use of them to find out insights and take better decisions. They search the number of individuals in a network, their bonding with other individuals

in the network. Some of the prominent social networks are given below:

A. Facebook

Facebook is used commonly as a social networking website. In this network registration is free and allows peoples to generate their profiles, share photos and video, send messages. It helps in finding and keeping in touch with, acquaintances, relatives and famous personalities.

B. Twitter

Twitter is a small blogging site that permits its members to share opinion on certain information. These small posts are called tweets. Registration is free for this network.

C. Wikipedia:

In Wikipedia, knowledge about any product or information about articles, people can be found. It is created by a group of people called Wikipedia's. It may be called as an online encyclopaedia. Users who are registered on this page, which can create article for publication.

D. LinkedIn

LinkedIn is type of social networking site to link people who are working and are in business. Mostly people connect here to share data about their companies, to advertise their type of products, discuss latest technologies. Further it is useful for users who are looking for jobs and employers for their prospective employees.

E. Reddit

Reddit is one of the social news website and frown where stories, music, movies and technologies can be shared. Based on these topics, sub-communities are formed. They are formally called as "sub-reddits. Reddit site of members, submit their content to this website and this content is voted by other reedit members [2].

II. DATA CLASSIFIER

There are three classifiers; Random forest tree algorithm, the decision tree algorithm J48 and the REP tree Algorithm are used for comparison.

A. Random forest tree

The Random Forests algorithm was evolved by Leo Breiman and Adele Cutler. Random decision forests[3][4] are ensemble type of learning the method for the classification, the regression and the tasks, which will operate by the building a multitude of the decision trees at the time of training and the class outputting that is mode of classes (classification) or the mean prediction (regression) of the user trees.[5] Random decision forests which correct for the decision tree's habit of the over fitting to the training set. Random Forests grows many of the classification trees. Each of the tree is follows:

- If the no: of cases inside set of training is N the sample N is cases at casual but with the replacement, from real data. This will be training set for the growing tree.
- If their exist M types of input variables, number is as specified like at the each node, M types of variables are being selected at the random out of the M and the topmost split on M used for node splitting . The result of the m is to be held to constant during growing of forest.
- Each type of tree grown to largest extent possible. There is not a pruning.

1) Advantages

- Accuracy
- Efficiently running on the large type of databases.
- thousands of the input variables are Handled without deletion of variable
- Gives the guess of what are the variables essential in classification
- Generates internal estimate which are unbiased of error generalization as forest building type of progresses
- Provides methods those are effective for missing data estimation
- Accuracy is Maintained when large proportion of data are misplaced
- Methods are Provided for the balancing of error inside class population data sets which are unbalanced
- Generated forests could be covered for the use on future the other type of data
- Prototypes are being calculated that give data about relation between variables and classification.
- Computes the proximities between the two cases which is used inside clustering, the locating the outliers, or the (by scaling) the give views those are interesting of data
- Capabilities of Above can extended to the data that are unlabelled , leading to the clustering which are unsupervised, views of data and the outlier detection
- Offers a method which is experimental for detecting the variable for interactions

2) Disadvantages

- Observation of the Random forests to over fit for some type of datasets with the noisy classification task or regression tasks.
- For the including of data categorical of the variables with the various number of the levels, the random forests biased in the favor of attributes with the more of data levels. Therefore, the variable importance scores from random forest are not reliable for this type of data.

B. J48

J48 classifier is simple C4.5 type of the decision tree for classification. It creates a binary tree. The approach of decision tree most useful in classification problem. With this approach, a tree is built to the model process of classification. Once tree built, it is needed to each of the tuple in database and the results inside classification for those tuple [6].

C. REP Tree

Reduces Error Pruning (REP) Tree where Classifier is a fast type of decision tree learner which is build for the decision tree or for the regression tree by using the information gain or information variance and trim it using the reduced error trimming (with back fitting) [3]. It is based on principle of computing for information gain with the entropy and minimizing the error arising from variance [4]. This algorithm is first recommended in [5] only sort's data values for the attributes those are numeric once. REP Tree applies the logic named as regression tree and generates multiple trees in altered iterations. Afterwards it picks best one from all spawned trees. As in C4.5 Algorithm, which deals with the missing values by breaking the corresponding instances into pieces [6]?

At each node inside tree it's possible to establish the no: of instances that misclassified on training set propagating by the errors upwards from leaf nodes. This could be compared to the rate of error if the node was replaced by the most of the common class resulting from that node. If difference is reduction in error, then the sub tree below the node can be considered for pruning. The calculation is for all nodes inside the tree and which has highest rate of reduced error is pruned. The procedure then iterated over pruned tree which are freshly prepared till there is not a possible reduction inside the error rate at any node. The error calculated by using pruning set, part of test set. This is the disadvantage of needing larger amounts of data, but advantage of resulting in the accurate classification tree [14].

III. LITERATURE SURVEY

Yunhu Jin et al. [7] proposed his work based on idea of multiple classifiers combination, constituted by the decision tree, each tree relies on independent samples, and all the trees in a forest with same values of random vector distribution. When classifying, each tree to vote and return the class with the most votes, which makes network security situation assessment is more accurate. Experiments show model can be quicker and more accurate to assess your current network security situation compared with Bayesian network.

Veena N. Jokhakar et al. [8] presents machine learning technique and the methodology for cooling temperature deviation, defect, diagnosis that consists of four phases namely data structuring, Association identification, Statistical derivation and classification. We also provide comparative results obtained with various type of data mining algorithms for ex. Neural networks, decision trees ,SVM, ensemble techniques (boosting and random forest) in the terms of performance parameters and prove that random forest outperforms rest of the techniques by achieving an accuracy of 95%.

Qing Li, et al.[9] proposed the deportment of the pattern-recognition system based on a classifier which is random forest (RF) is optimized by revising the no: of the decision trees and the number of the variables inside decision trees of the RF. Raw data characteristics of the Chinese liquors is being collected from QCM based e-nose that were in use by RF classifier without feature processes extraction and data pre-treatment, which can reserve detailed information as much as possible. The prediction accuracies and computation times indicate a superior type of performance by RF classifier over three other classifiers [linear discriminated analysis (LDA), back propagation artificial neural network (BP-ANN), and the support vector machine (SVM)]. Taking both application of thee nose and validation of RF classifier in account, an available method is obtained to identify flavors of Chinese liquors.

P. Kalaiselvi et al.[10] uses the Data Mining approach for weather predictions and studies the benefit of using it. Decision tree J48, EM(Expectation Maximization) and clustering algorithm of k-means has used in this research work to identify the variation in the weather conditions in the terms of Temperature, Sunny, Rainfall, Overcast and Wind Fall. Data mining is process of computer assisted of through digging and the analysing of immense sets of data and then mining the relevant data. Tools of Data mining predicts the behaviours and the future mode, allowing the to make businesses proper and good decisions. It can answer the questions that will traditionally were very strong in time to resolve.

Therefore can be applied to predict the meteorological data. That is called as weather prediction. Weather forecasting is

important type of application in the meteorology and being one of most challenging and also challenging problems technologically which is all around world. Predicting weather is useful to help preparing for best climates well the worst climate. We need on alert to the adverse type of the weather conditions by adapting the precautions and using the mechanisms of prediction for early warning of the hazardous weather phenomena. Many predictions on weather for ex: rainfall prediction, the augury of thunderstorm , predicting conditions of cloud are the challenges for the atmospheric type of research.

Manish Kumar et al.[11] studied, the experiments were conducted for the prediction task of the Chronic Kidney Disease which get from the UCI Machine Learning repository by using aloof 6 machine learning algorithms, these are namely: Random Forest (RFC) classifiers, the Sequential Minimal Optimization (SMO), Naïve Bayes, Radial Basis Function (RBF) and the Multilayer Perceptron Classifier (MLPC) and the Simple Logistic (SLG).The feature which is selected is required for training and the tee testing of each of the classifier individually with ten-fold cross validation. The results obtained show RF classifier will outperforms the other type of classifiers in the terms of the Area which is under the ROC curve (AUC), the accuracy and the MCC with the respective values which are1.0, 1.0 and 1.0.

Kittipol Wisaeng et al.[12] goal to explore the performance of data classification for set of large data amount of data. The tested algorithms are the functional algorithm, the logistic model trees algorithm, REP tree algo and best 1st algo of decision tree. The repository of the UCI will used for the test and to justify performance of algo named as decision tree algorithms. Subsequently, the classification algorithm that has the optimal potential will be suggested for use in large scale data.

Tina R. Patil et al.[13] put light on a performance based on evaluation of the correct instances and instances which are incorrect of the classification of data using Naïve Bayes and the J48 classification algorithm. Naive Bayes algo which is based on probability and the J48 algorithm that is based on algo which is called decision tree. The paper is sets out that to make the comparative type of evaluation of the classifiers Naive Bayes and the J48 in context of dataset of bank to maximize the rate in a true positive way and minimize the rate in a false positive way of the defaulters rather than achieving the higher classification of accuracy using MEKA tool. The results for the experiments shown here this paper is about the accuracy in classification, sensitivity and the specificity. Here The results here on dataset show that efficiency and the accuracy of approach which is j48 is much better than Naïve bayes.

R. Nithya et al. [14] the taxonomy approach for breast masses utilizing the DT approach. The comparison outcome of 12 DT algorithms containing BF Tree, Decision Stump, FT, C4.5, LAD Tree, AD Tree, LMT, NB Tree, Random Forest, Random Tree, REP Tree and CART. In link, four performance metrics were utilized. The object of the analysis is to determine the best DT classifier for mass taxonomy from BI-RADS features. In the experimental studies, all these DT algo are applied on the UCI data set. Experimental outcomes demonstration that LAD Tree and LMT has a better performance than Az tree, BFTree, Decision Stump, FT, C4.5, NBTree, Random Forest, Random Tree, REPTree and CART.

IV. Proposed Work

In the existing work, J48 tree and the Random forest tree has been performed, but for generating the better results we can also use Reduces Error Pruning (REP).

Proposed Algorithm:

Input: Training Data

Output: REP Tree

1. Start
2. Create the root node
3. Apply community detection for the feature extraction
4. Create an example with computed community features and the event label
5. REP Tree Classifier applied
6. Post Pruning performed and split tree
7. If (accuracy of pruned tree <accuracy of pruning set)

   Then

   Prune the tree

   Else

   Continue
8. Generate tree until we get optimal result
9. Stop

Flowchart For REPTREE Algorithm

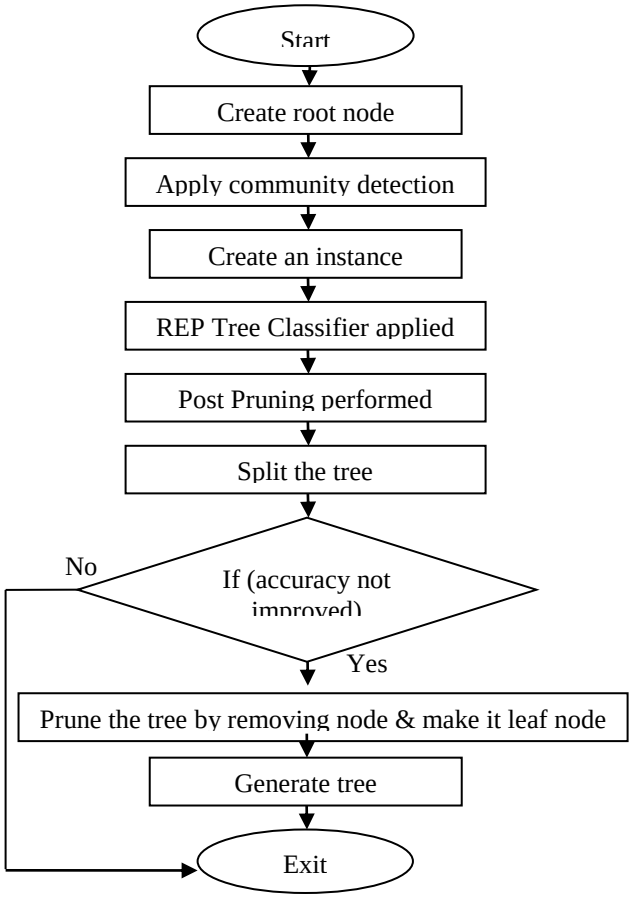

Fig. 2 Proposed flow diagrams

V. RESULTS ANALYSIS

We performed analysis on Enron and Slashdot Database in the MEKA tool. Three methods are taken such as J48, Random Forest Tree and REP Tree and then perform comparison between them.

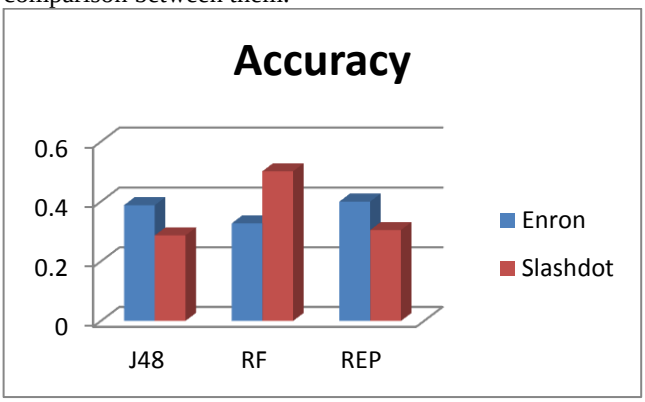

Fig. 3 accuracy

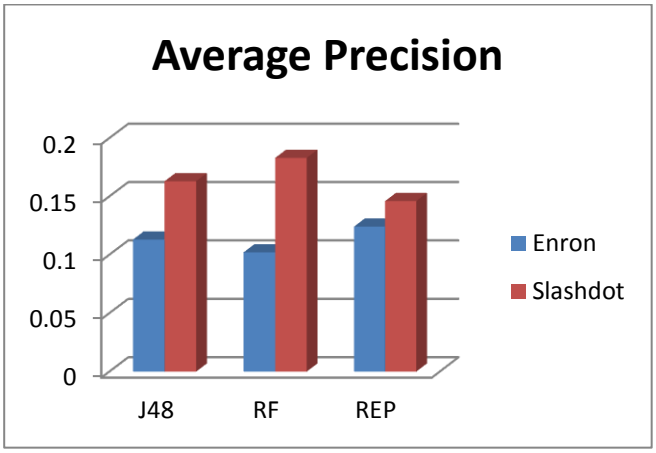

Fig. 4 Accuracy

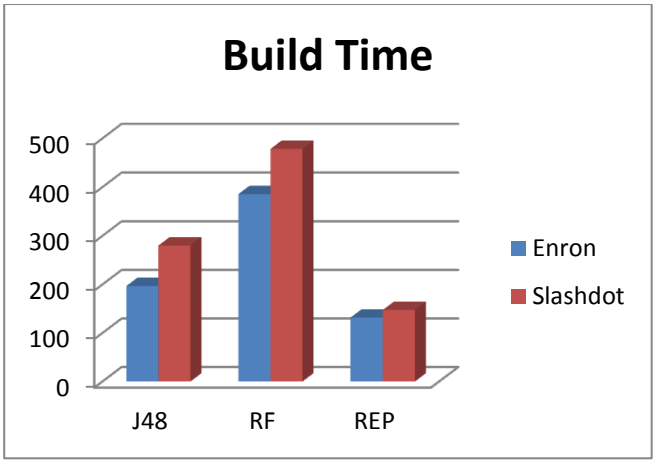

Fig. 5 Build Time

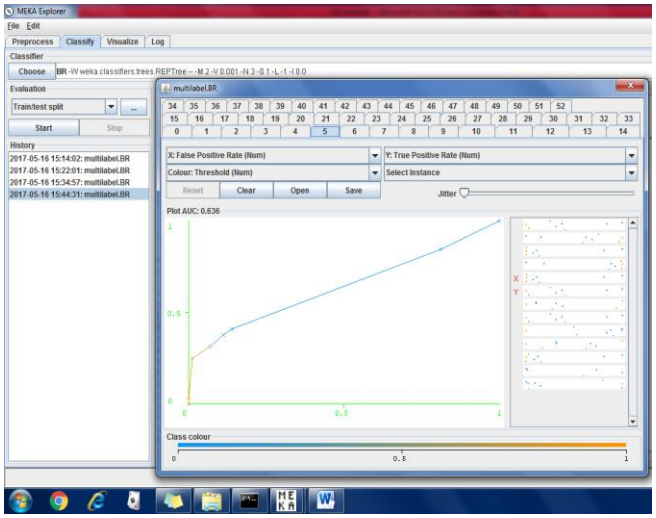

Fig. 6 ROC of Enron Database

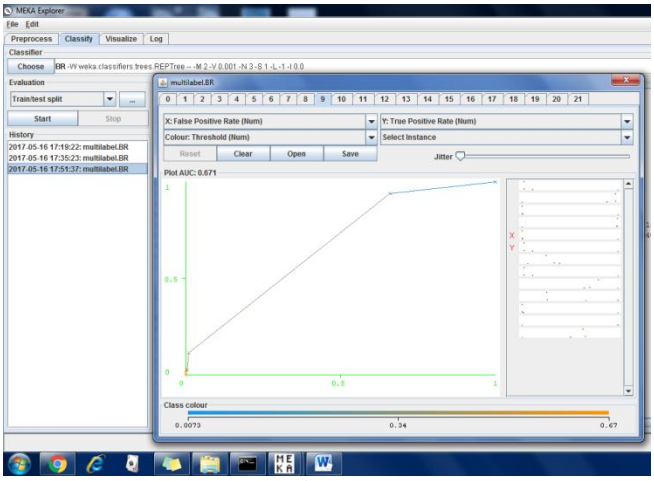
Fig. 7 ROC of Slashdot Database

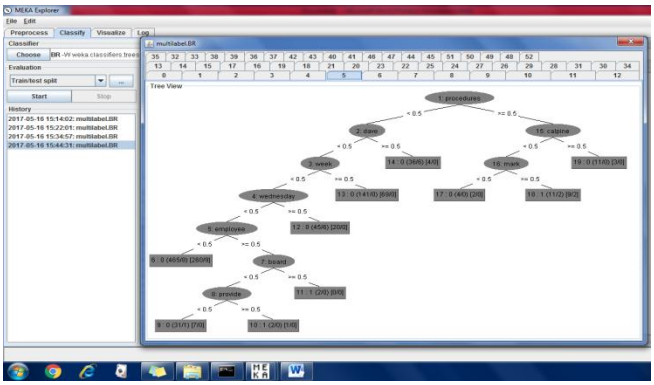
Fig. 8 REPTree of Enron Database

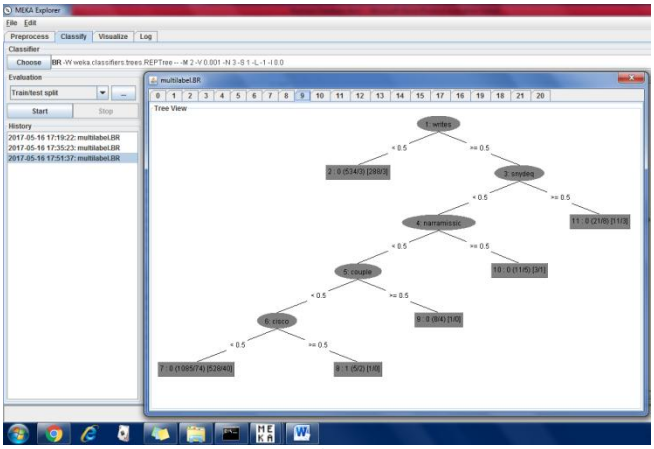
Fig. 9 REPTree of Slashdot Database

VI. CONCLUSION

Social network is a process that consists of applying the data analysis and the discovery algorithms, under acceptable computational, efficiency, limitations. Produce a particular guide of the patterns (or models) over data. This work investigated the efficiency of 3various types of the classifiers namely, the Random Forest, the REP Tree and the J48 Classifiers. From the results, we concluded that (REP) which is called reduce the error pruning is much better the other techniques for pruning.

REFERENCES

[1] Ho, Tin Kam, "*Random Decision Forests* (PDF)", Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16. pp. 278–282, August 1995.

[2] Tin Kam, "*The Random Subspace Method for Constructing Decision Forests*", (PDF).IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume: 20, Issue: 8, Pages: 832 – 844, 1998.

[3] Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome, " *The Elements of Statistical Learning*" (2nd ed.). Springer. USA, pp,1-764, 2008

[4] Danah m. boyd, Nicole B Ellison "*Social Network Sites: Definition, History, and Scholarship*" Journal of Computer-Mediated Communication, vol l3, issue no. 1, pp: 1-18 (2007).

[5] A.Malathi, D.Radha, "Analysis and Visualization of Social Media Networks", IEEE, ISBN**:** 978-1-5090-1022-6, Pages: 58 - 63, 2016.

[6] N. Landwehr, M. Hall, and E. Frank, "*Logistic Model Trees*," Machine Learning,, spinger, Volume 59, Issue 1, pp 161–205, 2005.

[7] Yunhu Jin, Yongjun Shen, Guidong Zhang, Hua Zhi "*The Model of Network Security Situation Assessment Based on Random Forest*", lEEE, ISSN**:** 2327-0594, Pages: 977 – 980, 978-1-4673, 2016.

[8] Veena N. Jokhakar, S. V. Patel "*A Random Forest Based Machine Learning Approach For Mild Steel Defect Diagnosis*", IEEE, ISSN**:** 2473-943X, Pages: 1 - 8 978-1-5090-061, 2016.

[9] Qiang Li, Student Yu Gu,, and Nan-Fei Wang, "*Application of Random Forest Classifier by Means of a QCM-based E-nose in the Identification of Chinese Liquor Flavors*", IEEE, Volume: 17, Issue: 6, Pages: 1788 – 1794, 2016.

[10] P. Kalaiselvi, D. Geetha "*Weather Prediction Using J48, EM And K-Means Clustering Algorithms*", International Journal of Innovative Research in Computer and Communication Engineering,, ISSN(Online): 2320-9801, Vol. 4, Issue 12, PP: 20889- 20895, December 2016.

[11] Manish Kumar, "*Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm*", International Journal of Computer Science and Mobile Computing A Monthly Journal of Computer Science and Information Technology, IJCSMC, ISSN 2320–088X, Vol. 5, Issue, pg.24 – 33, 2, February 2016.

[12] Kittipol Wisaeng "*A Comparison of Decision Tree Algorithms For UCI Repository Classification*" International Journal of Engineering Trends and Technology (IJETT), ISSN: 2231-5381,Volume 4 Issue 8, Page 3393- 3397, August 2013.

[13] Tina R. Patil, Mrs. S. S. Sherekar "*Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification*"

International Journal Of Computer Science And Applications, ISSN: 0974-1011, Vol. 6, Issue No.2,PP; 256- 261, Apr 2013.

[14] R. Nithya, B. Santhi, "*Decision tree classifiers for mass classification*",IJCSE, Decision tree classifiers for mass classification, Vol. 8, Nos. 1/2, Page: 39- 45, 2015

Authors Profile

Anupama Tyagi, pursued Bachelors' of engineering from Madhav Institute of Technology and Science, Gwalior (MP), India in 2014.She is currently pursuing her Master of Engineering (IT) from Madhav Institute of Technology and Science, Gwalior (MP),India

Dr. Sanjiv Sharma PhD, M.Tech (IT), B.E.(IT)is an Assistant Professor in the Department of Computer Science Engineering and Information Technology at Madhav Institute of Technology & Science Gwalior (MP), India. He received his PhD degree (Computer Science & Engineering) from Banasthali University, Jaipur (Raj.), India in 2014 and M.Tech (Information Technology) with honors from School Of Information Technology, Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal(MP), India in 2007. His current research interests include Social Network Analysis, Data Mining, Network Security and Ad hoc Network and Mobile Computing and their interdependency. The results of his work have led to 40+ articles published in various prestigious international journals, book chapter and conferences.

