International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 2
ISSN: 2454-4248
303 – 309

Hybrid Classifier using Evolutionary and Non-Evolutionary Algorithm for Performance Enhancement in Data Mining

Dr. Mohd Ashraf
Department of Computer Science & Engineering
Maulana Azad National Urdu University
Hyderabad, India
ashraf.saifee@gmail.com

***Abstract*—** Data Mining has been found to be the most active fields of research for the concluding couple of decades and is still alive for many challenging problems in academic and industries. Classification is a significant data mining technique which is utilized to find out in which group each data instance is linked up within a given information set. It is also applied for classifying information into different classes according to some constrains for different predefined categories. A set of effective approaches have already been notified and verified to solve the problem of the classification. In this research paper, the author suggested a new hybrid classifier model by uniting evolutionary and non-evolutionary algorithms; specifically, by integration of Genetic Programming and Decision Tree to improve the accuracy, comprehensiveness and time taken in classification

***Keywords*—** *Data Mining, Decision Tree (DT), Genetic Programming (GP), Evolutionary, Clustering.*

*****

I. INTRODUCTION

In today's universe, people are stunned with data increasing exponentially. The omnipresent computer makes it excessively easy to deliver things that previously people would have thrashed [7]. As the amount of information is increasing, its understanding is equally decreasing. Nowadays there is so much information that the useful information is hidden in between the layers of useless, irrelevant and redundant data and this leads to submitting the determination based on intuition and thinking instead of proper informed logistics. Wikipedia defines Data Mining as, "Data mining, an interdisciplinary subfield of computer science, is the computational process of finding patterns in heavy data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. [8]". Classification in Data Mining is a technique to predict the end product of unknown data using the information obtained from known information. Several algorithms are set for classification and there is a continuous battle to improve the algorithms. This improvement can be obtained by crossing two or more algorithms. In this paper, we will discuss the hybridization of Decision Tree and Genetic Algorithm and show how it affects the classification.

Through the diverse methodologies that was studied most of them talked about feature selection in order to improve the accuracy, timeliness and comprehensibility. Hence, feature selection is a process by which one can automatically search for best subset among the dataset. 'Best' means the subset that will achieve the highest accuracy. As well, as a number of properties are now less, it will also improve the comprehensibility and time requirement. Feature selection works on the precept that the search space consists of all possible combinations of attributes form data set and it is suitable to locate or look for the best combination of attributes that will ameliorate the public presentation of sorting. It is effective because it mainly reduces the data overfitting and thus improves accuracy.

Clustering basically allows users to make groups of data to determine the patterns from it. The advantage of clusters over classification is that every single attribute will be used to analyze the data. So, by first selecting the features which are most important in the given dataset and then using the clustering technique for further classification, one can extract the benefits of both supervised and unsupervised technique and thus increasing the relevancy of data. However, one disadvantage of clustering is the fact that the number of clusters to be formed need to be known beforehand. This problem is taken care by making as many clusters, as the number of classes in the data.

So, the proposed research in this paper suggests an algorithm that will explore the qualities of both Clustering and Feature Selection. Decision tree usually has some internal feature selection using the gain ratio. However, using the feature selection explicitly will increase the accuracy of the data and irrelevant attributes will be removed. After that, using the clustering technique by ignoring the class attributes increased the accuracy even more as new dataset has only the relevant and related attributes in the cluster.

II. LITERATURE REVIEW

The work carried out thus far by other researchers that are related to the proposal of a Genetic Algorithm/ Decision Tree (GA/DT) hybrid using feature selection or clustering is concisely presented. Radhika Kotecha et al. [1] The described importance of Data Mining and the need to work on Multiclass Classification problem. For this, the author

 ISSN: 2454-4248
303 – 309

took various algorithms (DT, k Nearest Neighbor (KNN), Genetic Programming (GP), Naïve Bayes (NB)) and some meticulous datasets and tabulated that how the DT performs best with the accuracy and training time whereas GP performs best in Comprehensibility. So, building a hybrid of GP and DT will give the best performance.

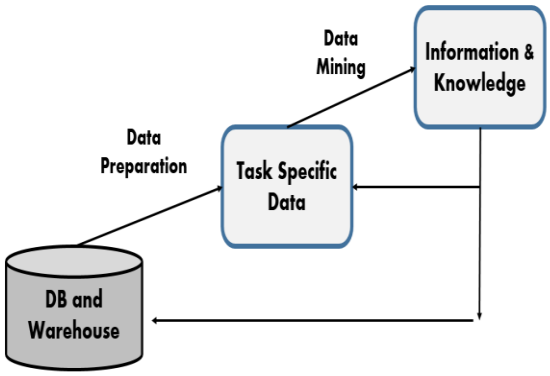
Ron Kohavi et al. [18] has given a wrapper approach for feature subset selection and used it for recognizing the relevant features. Author considered the wrapper as a 'black box' and fed subsets to it as an input and further noticed that the feature selection not only improves the accuracy but also improves training time and comprehensibility. Features are selected that is based on strong and weak relevance to the given dataset. Then it was presented that how selecting features using GA and then classification using DT gives an improved performance. Similarly, J. Bala et al. [19] described that a GA can be used for searching the space for all the possible subsets in a dataset and then DT can be invoked on the subset. DT's classification performance on unseen data can be used to identify the fitness of the feature subsets and that subset can in turn be used to evolve a better feature subset using GA. This process can be iterated till required performance level is achieved. However, the time requirement of this process was quite large and the process was over-complicated.

On the other hand, Deborah R. Carvalho et al. [6] Proposed a hybrid DT/GA hybrid in order to discover classification rules. His idea was to consider 'small disjuncts' for an increased performance. Small disjunct is basically a rule that covers a small number of examples. As they tend to overfit, they are error prone and mostly ignored while performing classification. However, although small disjuncts have less examples, a collection of disjuncts will cover a large set of examples and thus may help to increase the accuracy. Author created two types of hybrid, GA-Small and GA-Large-SN. Whereas GA Small tried to identify individual rules in each small disjunct, GA-Large-SN combined all the disjuncts and then find several rules in one go. GA-Small gave better accuracy, but took more time, whereas GA-Large-SN took considerably less time but accuracy was lower.

Thus, by looking at all the above techniques it can be predicted that using feature selection helps to increase the accuracy. The proposed GA/DT hybrid data mining algorithm for classification via feature selection and clustering is explained in the following section.

III. PROPOSED METHOD

Data mining is a process to find important and hidden relations among the vast amount of data and present them as information and/or knowledge. This process is a multi-step method that includes data collection, data management, data cleaning, data integration, data transformation, data mining and knowledge presentation [19]. This process can be depicted as in the Figure below.



Generally, data cleaning, integration and transformation are held in one step which is called the Data Preparation. After the data is ready the main concerns remaining is the data-mining itself.

After data collection and preparation of data by cleaning is the third step in the data mining process, where a suitable data mining method for the data and the problem at the hand is applied. Data mining methods can be generalized under two main categories. These are Supervised Learning and Non Supervised Learning methods [20].

Another kind of machine learning is reinforcement learning. The training information provided to the learning system by the environment (external trainer) is in the form of a scalar reinforcement signal that constitutes a measure of how well the system operates. The learner is not told which actions to take, but rather must discover which actions yield the best reward, by trying each action in turn.

Author uses a hybrid learning strategy, in this approach, different strategy have been combined on the basis of with their merits and demerits. But any classification algorithm can be evaluated by using the following parameters

1. **Accuracy:** It is defined as the percentage of correct predictions made by a classification algorithm. Equation (1) states how accuracy can be calculated:

   Accuracy= (No. of CP)/ (No. of CP *No. of ICP)* 100

   Where CP= Correct Prediction
   ICP= Incorrect Prediction

2. **Comprehensibility:** It shows a degree of simplicity in rule sets obtained after classification. Higher degree of comprehensibility is required. The greater number of nodes, lesser will be the comprehensibility.

3. **Training Time:** It is defined as the time that an algorithm takes to build a model of data sets. Minimum training time is desirable.

 ISSN: 2454-4248
303 – 309

Hybridization of algorithms for an increased performance has been done in various ways by various researchers over a long period of time. Hybrid is prototyped in order to take advantage of the algorithms involved and to remove their disadvantages simultaneously. There are many methods and methodologies which are defined for the hybridization of genetic algorithm and decision tree.

1. **Non-Evolutionary Classification Algorithm:** Different kind of non-evolutionary classification algorithms have been suggested and tested to predict the class level of data. Some common non-evolutionary classification algorithms are:
   i. Decision Trees,
   ii. Naïve Bayes Classifiers
   iii. K-Nearest Neighbor Classifiers.
2. **Evolutionary classification algorithms:** Evolutionary algorithms (EAs) are the search methods that take their inspiration from natural selection and survival of the fittest in the biological world. Several different techniques are grouped under the generic denomination of EA, which are
   i. Genetic Algorithm (GA),
   ii. Genetic Programming (GP),
   iii. Evolutionary Programming (EP).

The proposed approach uses the Non- evolutionary classification algorithm -Decision Tree (DT and Genetic Programming (GP) as an Evolutionary classification algorithm

1. Decision Tree (Non-Evolutionary): A Decision Tree is a flow-chart-like tree structure. Each internal node denotes a test on an attribute. Each branch represents an outcome of the test. Leaf nodes represent class distribution. The decision tree structure provides an explicit set of ―if-then rules rather than abstract mathematical equations, making the results easy to interpret.

   A decision tree is a predictive model most often used for classification. Decision trees partition the input space into cells where each cell belongs to one class. The partitioning is represented as a sequence of tests. Each interior node in the decision tree tests the value of some input variable, and the branches from the node are labelled with the possible results of the test. The leaf nodes represent the cells and specify the class to return if that leaf node is reached. The classification of a specific input instance is thus performed by starting at the root node and, depending on the results of the tests, following the appropriate branches until a leaf node is reached.

Application:

1. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal.
2. Descriptive means for calculating conditional probabilities. The decision tree is a predictive model that can be viewed as a tree where each branch of the tree is a classification question and leaves represent the partition of schematic tree-shaped diagram used to determine a course of action.

Advantages:

1. Comprehensibility i.e. people can easily understand why a decision tree classifies an instance as belonging to a specific class, high accuracy, low training time, etc.
2. Decision trees are inexpensive to construct

Disadvantage:

1. When data are given incrementally, Decision Trees cannot be used directly.
2. Also, when the number of classes is more and data is large, the size of the decision tree becomes very large.

Genetic Programming (Evolutionary Classification Algorithm)

GP is an evolutionary algorithm-based methodology inspired by biological evolution to find computer programs that perform a user-defined task. It is a specialization of GA where each individual is a computer program and uses a fitness measure to optimize this population.

GP steps to solve a problem:

a) Generate an initial population (computer programs) of random compositions of the functions and terminals of the problem.

b) Execute each program in the population and assign it a fitness value according to how well it solves the problem.

c) Create a new population of computer programs.
   (1) Copy the best existing programs
   (2) Create new computer programs by crossover
   (3) Create new computer programs by mutation

The best computer program that appeared in any generation, the best-so-far solution, is designated as the result of genetic programming.

Advantages:

1. GP is a flexible evolutionary technique with some features that can be very appropriate for the evaluation of classifiers. GP can be employed to construct classifiers using different kinds of representations, e.g., decision trees, classification rules, and many more.
2. Any preference criterion for e.g., accuracy can be expressed in terms of the fitness function that guides the search process of GP.
3. GP can automatically eliminate attributes unnecessary for the classification performing the task of feature extraction.

 ISSN: 2454-4248
303 – 309

Disadvantage: GP requires large training time. But once trained, the execution time of GP classifiers is much less. Hence, it can be used when there are no constraints regarding training time, but execution time matters.

After an introduction to above techniques, some meticulous datasets will be considered and then apply the above algorithms to them using Weka which is an Open Source, Portable, GUI based workbench and then tabulate the result on the basis of accuracy, training time and Comprehensibility.

So, the proposed research suggests an algorithm that will explore the qualities of both Clustering and Feature Selection. Decision tree usually has some internal feature selection using the gain ratio. However, using the feature selection explicitly will increase the accuracy of the data and irrelevant attributes will be removed. After that, using the clustering technique by ignoring the class attributes increased the accuracy even more as new dataset have only the relevant and related attributes in the cluster.

The working method of the proposed approach has been depicted in following block diagram. Firstly

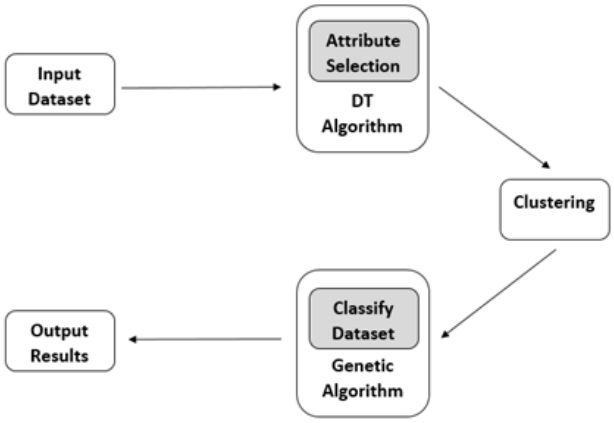


Fig. 1: *Block diagram for the proposed GP/DT/ Hybrid*

A. Algorithm:

Step-1: Take a dataset as input

Step-2: Divide dataset into Training and Testing dataset else, use n-fold cross validation

Step-3: Select the AttributeSelection which is used for feature selection.

Step-4: Choose evaluator as ClassifierSubsetEval as it let us select features using a classifier.

Step-5: After that choose J48 under the Decision Tree as the classifier

Step-6: Choose Best First searching method to search in feature selection

Step-7: Apply the feature selection to the dataset.

Step-8: Choose AddCluster to perform clustering the dataset

Step-9: Use SimpleKMeans cluster, which is the simplest type of clustering technique.

Step-10: Choose the number of clusters as the number of classes in the dataset.

Step-11: Put the index of the Class attribute under the ignoredAttributeIndices option so that unbiased clusters are created.

Step-12: Apply the Clustering technique on the filtered dataset

Step-13: Now classify the dataset using Genetic Programming

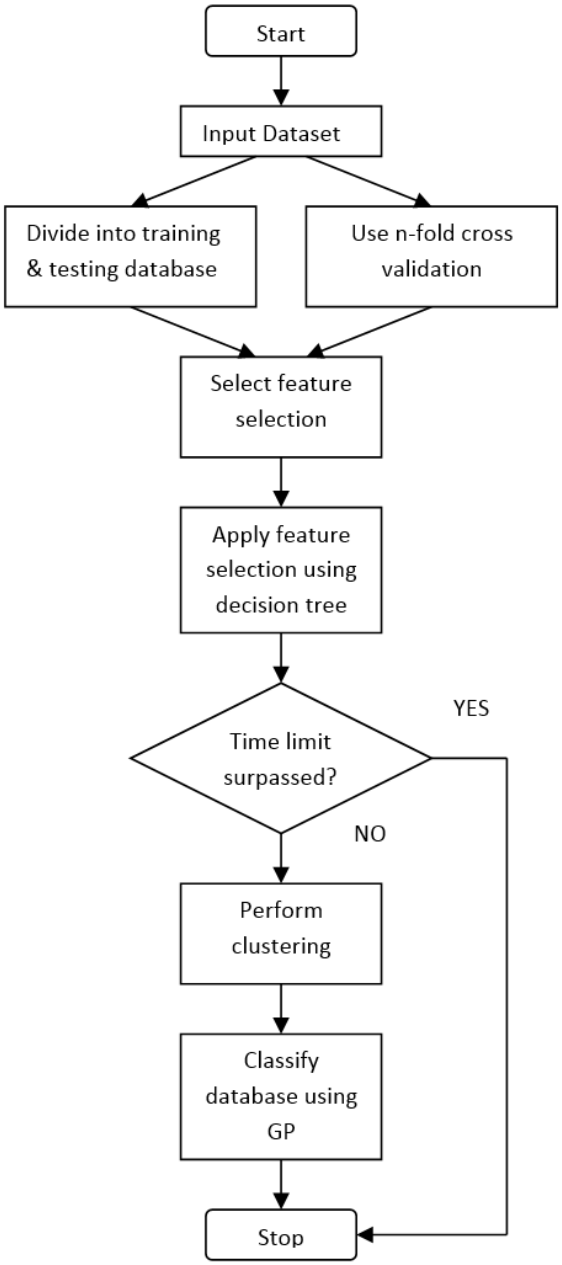
Step-14: End.

B. Algorithm explanation

To implement the proposed algorithm, Weka Software has been used to validate the algorithm.

1. First step is to input the dataset in the Weka Software.
2. Then use Attribute Selection, in order to obtain reduced/ranked data that not only helps to decrease the training time and comprehensibility, but also increase accuracy as the irrelevant or less relevant attributes are removed and they will now not interfere with the classification process.
3. Now there is a need to evaluate the attribute subset that was obtained from above filtering. So, use ClassifierSubsetEval which let the tool, select any classifier for estimating the accuracy of the subset obtained. Use J48 for attribute selection. DT was used for feature selection because it doesn't suffer from over-sensitivity to irrelevant or/and redundant attributes. In other algorithms, if two attributes are highly co-related then they get too much weight in its classification and presence irrelevant or redundant attributes lead to decline in accuracy. However, in DT, if two attributes are co-related, it's impossible to use both of the attributes for splitting into testing and training set, since that will result in same exact split and that will make no difference to the already existing tree.
4. For exploring the benefits of Clustering in the classification, select AddCluster. It's a filter which adds a new nominal attribute to the dataset and it represents the cluster that is assigned to each instance of the given clustering algorithm. Here SimpleKMeans cluster is used which is the simplest type of clustering algorithm. It's important to choose the number of clusters equal to the number of the classes in the datasets so that efficient clusters can be created. Also the class attributes of the dataset should be ignored during clustering to avoid the bias.
5. After applying all the above filters, one can classify the algorithm using Genetic Programming and obtain a result which is better than the GP and DT individually. This algorithm is then applied on five

data sets that was downloaded from UCI Repository and the results obtained are explained in the next section.



IV. RESULT

The above algorithm is applied on giving five datasets. Executing the hybrid on them verified the result showing the improvement in the accuracy and timing and concluded how the hybrid approach is working better than that of the GP and DT individually.

TABLE I. TABLE DEPICTING THE ACCURATE RESULTS BY USING GP, DT AND HYBRID CREATED.

Sr. No.	*Datasets*	*DT*	*GP*	*Hybrid*
1	Lymph	77.02	80.4	82.43
2	Heart-statlog	76.67	82.96	100
3	Molecular-Biology	21.69	24.52	33.96
4	Sonar	71.15	74.52	94.71
5	Zoo	92.07	93.06	96.03

As it can be seen that the accuracy has significantly increased and it became 100% in one of the case. Thus, this hybrid algorithm works really well for the accuracy. The increase in accuracy can further be depicted using the following bar graph:

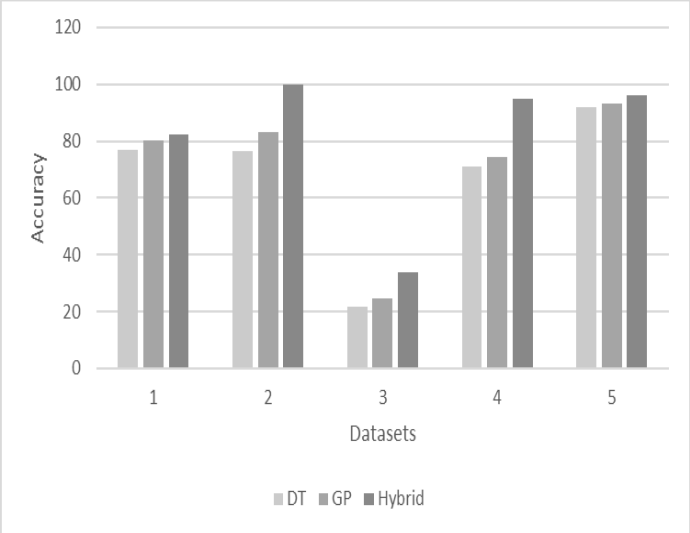


Fig. 2, Graph depicting the accuracy percentages using DT, GP and Hybrid approach

The results can be further verified by taking into account the detailed result analysis of one of the dataset, say, Sonar. arff. As one can see, all the statics are improved and performance of hybrid is higher as compared to DT and GP individually.

TABLE II. TABLE DEPICTING THE SUMMARY OF STRATIFIED CROSS-VALIDATION FOR GP, DT AND HYBRID

Sonar	***GP***	***DT***	***Hybrid***
Correctly classified Instances (%)	74.519	71.153	94.731
Incorrectly Classified Instances (%)	25.48	28.846	5.288
Kappa Statistics	0.485	0.422	0.894
Mean Absolute Error	0.254	0.286	0.052
Root Mean Squared Error	0.504	0.52	0.23
Relative Absolute Error (%)	51.185	57.504	10.592
Root Relative Squared Error (%)	101.174	104.37	49.025

TABLE III. TABLE DEPICTING THE DETAILED ACCURACY BY CLASS FOR GP, DT AND HYBRID.

SONAR Classes	**ROCK**			**MINE**		
Algorithms	***GP***	***DT***	***Hybrid***	***GP***	***DT***	***Hybrid***
TP Rate	0.68	0.711	0.944	0.802	0.712	0.95
FP Rate	0.2	0.288	0.05	0.32	0.289	0.056

 ISSN: 2454-4248
303 – 309

Precision	0.75	0.683	0.953	0.742	0.738	0.941
Recall	0.68	0.711	0.944	0.802	0.712	0.95
F-Measure	0.71	0.697	0.949	0.771	0.725	0.945

Above tables conclude that the hybrid proposed is working effectively in all dimensions of the classification and is more effective than DT or GP individually.

V. Conclusion and Future Work

In conclusion, the Hybrid approach suggested is working well and giving a better result than the decision tree and genetic programming individually. The future work will involve comparing it with other hybrids available and then tweaking it so that it will better performance than the hybrids created before.

The future work will involve comparing it with other hybrids available such as knee and GA or NB and GA and also work out their advantages and disadvantages. The process of hybridization is totally manual right now and in future we can create a Java class that can be uploaded in the Weka software to directly apply the hybridization. Also, we need to tweak the algorithm a bit so that it will provide better performance than the hybrids created before.

References

[1] Radhika Kotecha, Vijay Ukraine and Sanjay Garg, “An Empirical Analysis of Multiclass Classification Techniques in Data Mining”, INTERNATIONAL CONFERENCE ON CURRENT TRENDS IN TECHNOLOGY, Vol.2, NUiCONE, DECEMBER, 2011

[2] J Sangari. R. S, Dr. M. Balamurugan, “A SURVEY ON RAINFALL PREDICTION USING DATAMINING”, International Journal of Computer Science and Mobile Applications, Vol.2 Issue. 2, February- 2014.

[3] G. Kalyani, A. Jaya Lakshmi, “Performance Assessment of Different Classification Techniques for Intrusion Detection”, Dept of CSE, DVR & Dr HS MIC College of Technology, Kanchikacherla, Krishna.

[4] Pedro G. Espejo, Sebasti´an Ventura, and Francisco Herrera, “A Survey on the Application of Genetic Programming to Classification”, IEEE transactions on systems, man, and cybernetics—part c: applications and reviews, VOL. 40, NO. 2, MARCH 2010

[5] Rikard König, Ulf Johansson, Tuve Löfström and Lars Niklasson, “Improving GP Classification Performance by Injection of Decision Trees”, Information Fusion Research Program at the University of Skövde, 2010

[6] Deborah R. Carvalho, Alex A. Freitas, “A Hybrid Decision Tree/Genetic Algorithm Method for Data Mining”, Universidade Tuiti do Parana (UTP) Computer Science Dept., Brazil.

[7] Ian H. Witten, Eibe Frank, Mark A. Hall, “What’s It All About?,”] Data Mining Practical Machine Learning Tools and Techniques, Third Edition. USA, 2011.

[8] Wikipedia. (2014, November, 11), Data Mining[Online].Available: http://en.wikipedia.org/wiki/Data_mining

[9] Matthieu Cord, and Sarah Jane Delany, “Supervised Learning,” P´adraig Cunningham

[10] Harvinder Chauhan, Anu Chauhan, “Evaluating Performance of Decision Tree Algorithms,” International Journal of Scientific and Research Publications, Volume 4, Issue 4, April 2014

[11] UCI repository. (2008, July, 15). Index of /Datasets/UCI/arff [Online]. Available: http://repository.seasr.org/Datasets/UCI/arff/

[12] leyan. (2013, April, 05). Genetic Programming Classifier for Weka [Online]. Available: http://sourceforge.net/projects/wekagp/

[13] Machine Learning Group at the University of Waikato. (2014). Weka 3: Data Mining Software in Java [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/

[14] Jiawei Han and Micheline Kamber, “Introduction, Data Mining: Concepts and Techniques”, Second Edition. University of Illinois at Urbana-Champaign, USA, 2006.

[15] Medeswara Rao, Kondamudi, Sudhir Tirumalasetty, “Improved Clustering And Naïve Bayesian Based Binary Decision Tree With Bagging Approach,” International Journal of Computer Trends and Technology (IJCTT) – volume 5 number 2 –Nov 2013

[16] MIT Press. (2013). The GP Tutorial [Online]. Available: http://www.geneticprogramming.com/Tutorial/

[17] R.S. Michalski and R.L. Chilausky "Learning by Being Told and Learning from Examples: An Experimental Comparison of the Two Methods of Knowledge Acquisition in the Context of Developing an Expert System for Soybean Disease Diagnosis", International Journal of Policy Analysis and Information Systems, Vol. 4, No. 2

[18] Ron Kohavi, George H. John, “Wrappers for feature subset selection”, Data Mining and Visualization, Silicon Graphics, Landings Drive, Mountain View, CA 94043, USA Received September 1995; revised May 1996

[19] J. Bala, J. Huang and H. Vafaie, “Hybrid Learning Using Genetic Algorithms and Decision Trees for Pattern Classification”, School of Information Technology and Engineering, George Mason University, IJCAI conference, Montreal, August 19-25, 1995

[20] Gary Stein, Bing Chen, Annie S. Wu, Kien A. Hua, "Decision Tree Classifier For Network Intrusion Detection With GA-based Feature Selection", University of Central Florida

[21] Huan Liu, Lei Yu. Toward Integrating Feature Selection Algorithms for Classification and Clustering, IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 4, pp. 491-502, April-2005