International Journal of Modern Trends in Engineering and Research
www.ijmter.com
e-ISSN No.:2349-9745, Date: 2-4 July, 2015

A classification scheme for plotting Microarray Gene Expression Datasets to Decision Tree Algorithm using Distributed Systems

Neha V.Bhatambarekar[1], Prof.Archana Vaidya[2]

[1]*Computer Engineering, G.E.S.R.H.S. COE, b.neha29@gmail.com*
[2]*Department name, G.E.S.R.H.S. COE, archana.s.vaidya@gmail.com*

Abstract- scrutinizing gene expression data is a challenging endeavor since the substantial number of features against the shortage of available samples can incline to over fitting. In order to circumvent this pitfall and achieve high performance, some schemes build complex classifiers, using state-of-the-art or well-established approaches. In medical decision making (classification, diagnosing) there are numerous situations where decision must be made competently and reliably. One of the recurrently employed techniques to extract knowledge from data is decision tree induction, since the representation of knowledge is very intuitive and easily interpreted by humans. Decision trees are authentic and effective decision making methodology that provides elevated classification accuracy with a simple representation of gathered knowledge which has been used in distinct areas of medical decision making. Instead of manually improving the design components of a decision tree algorithm as it has been done for the past 40 years, a novel approach of a hyper-heuristic evolutionary algorithm using distributed systems for optimally combining design components from decision-tree algorithms called HEAD-DT is proposed. Thus, the hyper-heuristic automatically designs novel decision-tree algorithms, tailored to a particular type of data sets, associated with a given application domain. Hyper-heuristic methodology is capable of providing a faster, less-strenuous and at least equally effective strategy for improving decision-tree algorithms for particular application domains. By the end of the evolution, the proposed system HEAD-DT is expected to generate a new and possibly better decision-tree algorithm for a given application domain. The performance of HEAD-DT is assessed in real-world microarray gene expression data sets and it is compared against very well-known decision-tree algorithms such as REPTree ,CART,C4.5. HEAD-DT is expected to significantly outperform the baseline manually-designed decision-tree algorithms regarding predictive accuracy and F-Measure.

Keywords- automatic algorithm design; decision trees; evolutionary algorithm; hyper-heuristics; distributed systems; machine learning.

I .INTRODUCTION

The DNA microarray technology enables monitoring the expression of thousands of genes concurrently which leads to better grasping of many biological processes, reinforced diagnosis, and treatment of various diseases [1]. However data gathered by DNA microarrays are not suitable for direct human examination, since a single experiment contains thousands of measured expression values. Various perspectives have been stated towards exploiting data mining from microarray data which includes supervised and unsupervised machine learning algorithms. The machine learning task of conjecturing a function from labeled training data is termed as supervised learning. In this approach, each sample is a pair consisting of an input object and an output value called the supervisory signal. Supervised learning algorithm examines the training data and produces an inferred function, which can be used for plotting new examples. Decision tree is a classifier

represented in a flowchart like tree structure which has been mainly used to represent classification models. Decision tree induction algorithms provides numerous advantages over other learning algorithms, such as resilience to noise, economical computational cost for generating the model, and ability to deal with undesirable attributes. Besides, the induced model also presents good generalization ability any of the decision tree induction algorithms are based on a greedy top-down recursive approach for tree growth. One major obstacle of greedy search is that it usually tends to generate sub-optimal solutions. The alternative to overcome this problem can be the initiation of an ensemble of trees. Ensembles are created by inducing multiple trees from training samples and the conclusive mapping is customarily given through a voting scheme. However, the comprehensibility of analyzing a single decision tree is a disadvantage of ensembles. Therefore an approach of Evolutionary Algorithms (EAs) is been widely used in the induction of decision trees. Rather than the local search, EAs perform a robust global search in the space of candidate solutions which tends, to cope better with attribute interactions than greedy methods. These algorithms are motivated by the principle of natural selection and genetics. Operations inspired by genetics, such as crossover and mutation, are performed on the selected individuals producing new offspring which replaces the 2 parents, creating a new generation of individuals. This process is frequently repeated until a stopping criterion is satisfied.

II. LITERATURE SURVEY

R. C. Barros, R. Cerri, P. A. Jaskowiak, and A. C. P. L. F. de Carvalho [4] proposed hill-climbing bottom-up induction an iterative searching methodology. Hill climbing is simply a loop like strategy that continually moves in the direction of increasing value. The algorithm suffers from three drawbacks as Local Maxima, Plateau and Ridges. Deborah R. Carvalho and Alex A. Freitas [5] proposed a hybrid decision tree algorithm method. A hybrid approach which uses rule induction and clustering techniques elevates the accuracy resulting in fast processing time. Hence, small disjuncts are error prone due to their nature. It covers only a few small disjunct of examples. R. C. Barros, D. D. Ruiz, and M. P. Basgalupp [6] suggested use of Model Trees. Model trees are alike the regression trees, which are hierarchical structures for predicting continuous dependent variables. The model trees induction is sequential in nature as the the greedy algorithms and locally optimal at each node split, hence the convergence for a global optimal solution is hardly feasible. In addition, minor modifications in the training set often lead to large changes in the final model due to the intrinsic instability of these algorithms .L. Breiman, "Random forests," Hong Hu, Jiuyong Li, Hua Wang ,Grant Daggard, Mingren Shi Ensemble methods [7][8] were proposed to take advantage of these unstable algorithms by growing a forest of trees from the data and averaging their predictions. Techniques that aggregate the prediction of multiple models in order to improve predictive performance are known as ensemble methods. They tend to perform better than any single model alone when in conformity with two necessary conditions: (i) the base models should be independent of each (ii) the base models should do better than a model that performs random guessing. Well-known examples of ensembles are: Bagging [Leo Breiman] proposed that the training set is sampled according to a uniform probability distribution, and for each sample a base model is trained. Bagging's effectiveness depends on the (in)stability of the base models. Boosting [Freund and Schapire] stated another technique in which the distribution of the training set is changed iteratively so that the base models will focus on examples that are hard to predict. Unlike bagging, boosting assigns a weight to each training instance and dynamically changes the weights at the end of each boosting iteration. Ross Quinlan in 1993 proposed an algorithm named C4.5 [2] the successor of ID3,for classification of decision trees. It improves the ID3 algorithm by dealing with both continuous and discrete attributes, missing values and pruning trees after construction. However

C4.5 suffers from a few drawbacks as empty branches (nodes with value zero), insignificant branches(branches which reduce the usability of the tree) and over fitting. Breiman and Freidman in 1984 suggested a method of classification called Classification and Regression Trees(CART) [3] which is a non-parametric decision tree learning technique that produces either classification or regression trees, depending on whether the dependent variable is categorical or numeric, respectively. CART uses historical data to construct the decision trees.

III. PROPOSED SYSTEM

3.1 Problem Definition

"To develop a system that deals with correctly classifying the huge uncertainties of gene expressions data inducing a decision tree tailored to a specific type application domain consisting various data sets using hyper heuristic algorithm where parameters like F-measure aid in to improve the classification accuracy using distributed systems. "

3.2 System Architecture

Hyper-heuristics comprise a set of approaches with the common goal of automating the design and tuning of heuristic methods to solve hard computational search problems [1]. A hyper-solicits to automate, often by the embodying the machine learning techniques, which includes the process of selecting, combining, generating or adapting several simpler heuristics to efficiently solve computational search problems.

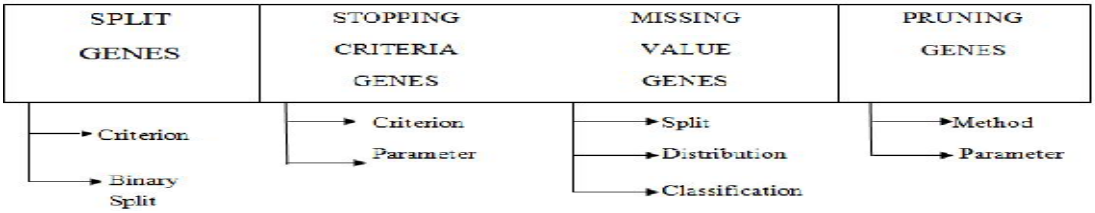


Fig 1: Major building blocks of Decision Tree

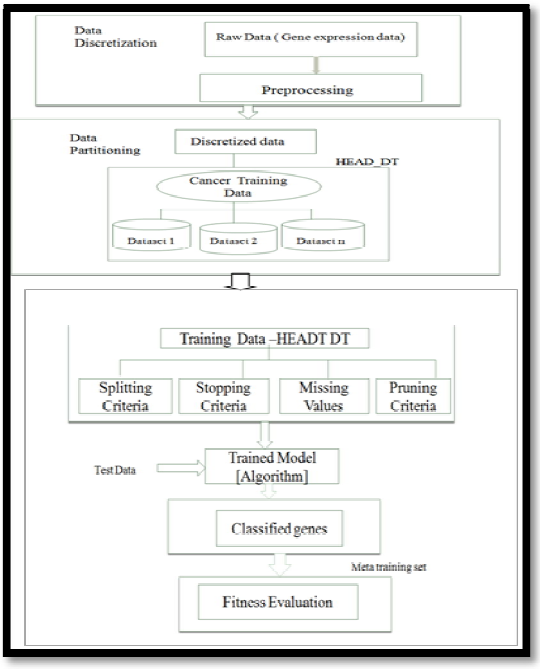


Fig 2: System Architecture

Every individual component is encoded as an integer string as shown in figure below , and each gene has a different range of supported values. The genes are divided into four categories that represent the major building blocks of a DT algorithm: (i) split genes (ii) stopping criteria genes (iii) missing values genes and (iv) pruning genes. combing through an evolutionary algorithm, components or building-blocks of human designed heuristics. HEAD-DT individuals are collections of building blocks of DT algorithms[9].

Splitting Gene Criteria: These genes are concerned with the task of selecting the attribute to split the data in the current node of the decision tree.

Stopping Criteria Genes: The top-down induction of a DT is recursive and it keeps on going until a stopping criterion (per-pruning) is satisfied. Different stopping criterion used are as :

a) Class homogeneity: when all instances that reach a given node belong to the same class, then the nodes are not split further.

b) Maximum tree depth: a parameter tree depth can be
Specified to avoid deep trees.

c) Minimum number of instances for a non-terminal node: a parameter minimum number of instances for a non-terminal node can be specified to alleviate the data fragmentation.

d) Accuracy threshold within a node: when a parameter reaches the accuracy which is specified for halting the growth of the tree within a node i.e. majority of instances have reached a defined threshold. Missing Values Gene: Missing values can be an issue during tree induction and also during classification therefore handling missing values is an important task in decision tree induction. During tree induction there are two instances where missing value are need to be dealt with are as:

a) Split Gene

b) Distributed Gene

Pruning Genes: Pruning is usually performed in decision trees for enhancing tree comprehensibility by reducing its size while maintaining accuracy. It was originally conceived as a strategy for tolerating noisy data, however it was found to improve decision tree accuracy in many noisy data sets [3][4].Pruning helps to avoid over-fitting to the training set and to reduce the size of decision tree which makes simpler interpretation for users. The well-known approaches for pruning a decision tree are : Error-based pruning (EBP), Cost-complexity pruning (CCP) ,Reduced error pruning (REP) ,Pessimistic error pruning (PEP),Minimum error pruning (MEP).

3.3 System Flow

The raw gene expressions are fetched as a text file which is then preprocessed that is converted to .csv format for discretizing the data. Once the data is arranged the system is then trained with HEAD –DT's building blocks steps. Thus the supervised learning model(training data) is employed. Further the test data is uploaded and again HEAD-DT steps are run to classify the unseen instances through supervised learning. The gene expressions are then assigned the class labels in the decision tree. Finally the Fitness evaluation is done with respect to the parameters as F-Measure and Fitness function (Meta-training set) to check the accuracy of correctly classified samples.

IV. IMPLEMENTATION DETAILS

4.1 Environment

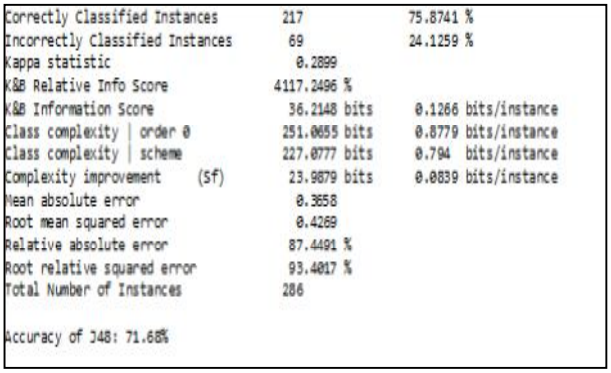
The proposed system is implemented in java jdk-7 environment. Weka 3.6 tool is integrated with java system for data analysis. For system development Eclipse (JUNO)IDE is used. The user GUI is created using java swing control.

4.2 Datasets

To classify Gene Expressions, there is a need of cancer dataset which has been downloaded.
a) It contains 35 real world Cancer gene expression datasets
b) 14 datasets are used for parameter optimization and remaining 21 datasets are used for performing experiments. Initially for testing purpose, Breast Cancer datatset was used which has been downloaded. a) Extracted fields are Age, Tumor-Size, Breast-quad etc.
As the proposed system deals with multiple datasets the computation required is considerably high which can be handled by dividing the system in multi node processing ,for balancing the load and which would also increase the efficiency. This distribution of workload can be done using Java RMI feature.

V.RESULTS

For carrying out the test on breast cancer dataset using baseline algorithms we had installed Weka 3.6 tool and jre-7 on windows platform. The datasets were downloaded in text format. The text file contained each data record in a row separated with space. This text file was converted into .csv format to perform the test. The results for baseline approaches like C4.5 and CART were calculated initially and following results were obtained on the breast cancer dataset. The testing is done on Affymetrix datasets. The obtained results are shown below in Figure 3. Figure 4 shows results obtained after running DCSM splitting measure .



Correctly Classified Instances	217	75.8741 %
Incorrectly Classified Instances	69	24.1259 %
Kappa statistic	0.2899	
K&B Relative Info Score	4117.2496 %	
K&B Information Score	36.2148 bits	0.1266 bits/instance
Class complexity | order 0	251.0655 bits	0.8779 bits/instance
Class complexity | scheme	227.0777 bits	0.794 bits/instance
Complexity improvement (Sf)	23.9879 bits	0.0839 bits/instance
Mean absolute error	0.3658	
Root mean squared error	0.4269	
Relative absolute error	87.4491 %	
Root relative squared error	93.4017 %	
Total Number of Instances	286	

Accuracy of J48: 71.68%

Fig 3 : C4.5 applied on Breast Cancer dataset

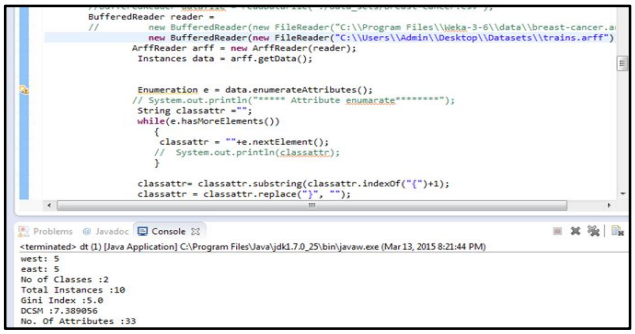


Fig 4 : DCSM applied on Breast Cancer dataset

The comparative results of C4.5 and DCSM is shown in Figure 5 where DCSM exceeds in identifying more classes and generating therefore unbiased decision tree. For the analysis of results the proposed system is expected to have the following factors: 1. Good Generalization ability: As hyper heuristic tend to classify across many datasets of a given application domain. Also it reports the components of decision tree that were most frequently selected by HEAD-DT in order to create decision-tree algorithms customized to microarray. 2. Increase in Accuracy: The proposed system uses parameters like Fitness function & f-measure which measures the precision and recall values to measure the correctly classified samples. 3. Reduced Time Complexity : The automated approach can be considered a much more cost-effective approach to the design of decision-tree algorithms tailored to a given type of data set (or application domain) than the manual algorithm design approach currently used in machine learning.

Sr.No	Datasets	No. of Instances	No. of Attributes	No. of Classes	Gini Index	DSCM Criteria
1	Diabetes	768	9	2	4.543	9.621
2	Trains	10	33	2	5.0	7.389
3	Zoo	101	18	7	7.593	8.02
4	Waveforms	5000	41	3	6.666	20.390

Fig 5: Comparative analysis of C4.5(Gini Index) and DCSM splitting criteria

CONCLUSION

Decision trees are one of the most recurrently used representations for classifiers .In our system we propose HEAD-DT algorithm to classify and analyze microarray gene expression datasets. HEAD-DT, a hyper-heuristic algorithm improves the design of top-down decision-tree induction algorithm by automating the manual classification process. For testing all possible modifications in the design components of decision-tree algorithms it would be infeasible with the human manual approach. We believe the evolutionary search strategy of HEAD-DT would constitute a robust and cost effective solution to that problem. The proposed system also gives good generalization ability as it provides the solution across multiple datasets for a given domain. We focus on increasing the efficiency and reducing the processing time of the system by using distributed system approach.

REFERENCES

[1] R.C.Barros, M. P. Basgalupp , A. A. Freitas and A. C. P. L. F. de Carvalho ,Evolutionary Design of Decision Tree Algorithms Tailored to Microarray Gene Expression Datasets, IEEE Transactions On Evolutionary Computation, Vol.18, No. 6, December 2014.

[2] C4.5: programs for machine learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993.

[3] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees. Wadsworth, 1984

[4] R. C. Barros, R. Cerri, P. A. Jaskowiak, and A. C. P. L. F. de Carvalho, "A Bottom-Up Oblique Decision Tree Induction Algorithm," in 11th International Conference on Intelligent Systems Design and Applications,2011, pp. 450.

[5] Deborah R. Carvalho and Alex A. Freitas , "A hybrid decision tree/genetic algorithm for coping with the problem of small disjuncts in Data Mining". Proc 2000 Genetic and Evolutionary Computation Conf. (Gecco-2000), 2000, 1061-1068. Las Vegas, NV, USA. July.

[6] R. C. Barros, D. D. Ruiz, and M. P. Basgalupp, "Evolutionary model trees for handling continuous classes in machine learning," Information Sciences, vol. 181, pp. 954–971, 2011.

[7] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp.5–32, 2001.

[8] Hong Hu, Jiuyong Li, Hua Wang ,Grant Daggard ,Mingren Shi "A Maximally Diversified Multiple Decision Tree Algorithm for Microarray Data Classification."

[9] Ms. Neha V.Bhatambarekar, Prof. Payal S. Kulkarni "A Survey on: Stratified mapping of Microarray Gene Expression datasets to decision tree algorithm aided through Evolutionary Design" IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 6, Ver. V (Nov – Dec. 2014), PP 01-06

