International Journal of Computational Bioinformatics and In Silico Modeling
Vol. 2, No. 6 (2013): 257-261
Research Article
Open Access

ISSN: 2320-0634

Comparison of Rule Based Classifiers by Pre-Learning for Clustering of Gene Expression Data

Sudhakar Tripathi* and R.B.Mishra

Department of Computer Engineering, IIT(BHU), Varanasi (U.P.), India

*Corresponding author: Sudhakar Tripathi; e-mail: stripathi.rs.cse@iitbhu.ac.in

Received: 11 September 2013 Accepted: 23 September 2013 Online: 01 November 2013

ABSTRACT

Recent advancement in microarray technology has helped to generate huge amount of gene expression data sets very rapidly. Major challenge is to analyze and explore these data sets to find the genes having similar profiles and hence predict their functions and pathways. To achieve this majorly used technique is clustering. Clustering is to find appropriate number of clusters as well as subsets belonging to those clusters. Many clustering techniques have been used to cluster time series as well as sample gene expression data sets but no single one is reported to be best in general conditions. In this research we have performed clustering of gene expression data sets using rule based classifiers (CART, C5, CHAID, QUEST) by training them using train data sets prepared by using some efficient heuristic clustering (we have used k-means).We have shown comparison of these models for testing and validation data sets and then these models can be generalized for clustering gene expression data sets by selecting appropriate model corresponding to preferences. Here we have assumed that all the data is being generated from same or similar source. Main benefit of using these models is simplicity and efficiency in terms of speed and storage. Hence we have used supervised and unsupervised techniques to generate and compare the models for efficient and accurate clustering of gene expression data sets.

Keywords: Rule Based Classifier; Semisupervised Clustering; Gene Expression Data Set

INTRODUCTION

Large amount of gene expression data are available due to the advancement of micro-array chip technology now a days. But the main challenge is to analyze these data for meaningful predictions and conclusions. One of the most important techniques is clustering microarray data. The main challenges of microarray data analysis include gene selection, clustering, and classification. Microarray datasets in contrast with other application domains, contain a small number of records (less than a hundred), while the number of fields (genes), is typically in thousands.

An important issue in data analysis is feature selection. In gene expression analysis the features are the genes. Gene selection is a process of finding the genes most strongly related to a particular class. One benefit provided by this process is the reduction of the foresaid dimensionality of dataset. Moreover, a large number of genes are irrelevant when classification is applied. The danger of overshadowing the contribution of relevant genes is reduced when gene selection is applied. Clustering is the far most used method in gene expression analysis. Hierarchical clustering is currently the most frequently applied method in gene expression analysis. In microarray analysis classification is applied to discriminate diseases or to predict outcomes based on gene expression patterns and perhaps even identify the best treatment for given genetic signature.

In this paper we have used rule based models having the rule sets generated by CRT, C5, CHAID and QUEST as rule bases. Here we present a comparative view of these models in contrast with feature selection, importance and cluster predicting accuracy by pre learning. Once we find the appropriate model for our case we can generalize that model for clustering. The assumption that we have made is the data is being generated from similar or same source.

MATERIALS AND METHODS

A. MODELS

Following models are used in this paper:-

1) C&RT(CRT)

C&RT stands for Classification and Regression Trees, originally described in the book by the same name [1]. C&RT partitions the data into two subsets so that the records within each subset are more homogeneous than in the previous subset. It is a recursive process—each of those two subsets is then split again, and the process repeats until the homogeneity criterion is reached or until some other stopping criterion is satisfied (as do all of the tree-growing methods). The same predictor field may be used several times at different levels in the tree. It uses surrogate splitting to make the best use of data with missing values. C&RT is quite flexible. It allows unequal misclassification costs to be considered in the tree growing process. It also allows you to specify the prior probability distribution in a classification problem. You can apply automatic cost-complexity pruning to a C&RT tree to obtain a more generalizable tree.

C&RT works by choosing a split at each node such that each child node created by the split is more pure than its parent node. Here purity refers to similarity of values of the target field. In a completely pure node, all of the records have the same value for the target field. C&RT measures the impurity of a split at a node by defining an impurity measure.

2) C5

C5 (improvement of C4.5) is an algorithm used to generate a decision tree developed by Ross Quinlan [3-4].The changes in various versions of C5 are available at [4]. The decision trees generated by C5 can be used for classification, and it is used as a statistical classifier. C5 builds decision trees or corresponding rule sets from training data set, using the concept of information entropy. The training data set is a set $S_T = s_1,s_2,...$ of already classified samples, known as supervised training. Each sample $s_i = x_1,x_2,...$ is a linear vector where $x_1,x_2,...$ represent input features of the sample. The training data is augmented with a vector $C = c_1,c_2,... c_n$ where $c_1,c_2,...$ represent the class to which each sample belongs and n is total number of classes[2]. At each node of the tree, C5 chooses one feature of the data set that most effectively splits its set of samples into subsets enriched in one class or the other. The criterion for splitting is the normalized information gain (difference in entropy) that results from choosing a feature for splitting the data into subsets [2]. The feature with the highest normalized information gain is chosen to make the decision. The C5 algorithm then follows the same steps on the smaller sub lists [2].

To maximize interpretability, C5.0 classifiers are expressed as decision trees or sets of if-then rules, forms that are generally easier to understand than neural networks. C5.0 is easy to use and does not presume any special knowledge of Statistics or Machine Learning.

3) CHAID (CHi-squared Automatic Interaction Detection)

CHAID stands for Chi-squared Automatic Interaction Detector. It is a highly efficient statistical technique for segmentation, or tree growing, developed by Gordon V. Kass[5]. Using the significance of a statistical test as a criterion, CHAID evaluates all of the values of a potential predictor field. It merges values that are judged to be statistically homogeneous (similar) with respect to the target variable and maintains all other values that heterogeneous (dissimilar). It then selects the best predictor to form the first branch in the decision tree, such that each child node is made of a group of homogeneous values of the selected field. This process continues recursively until the tree is fully grown. The statistical test used depends upon the measurement level of the target field. If the target field is continuous, an F test is used. If the target field is categorical, a chi-squared test is used. CHAID is not a binary tree method; that is, it can produce more than two categories at any particular level in the tree. Therefore, it tends to create a wider tree than do the binary growing methods. It works for all types of variables, and it accepts both case weights and frequency variables. It handles missing values by treating them all as a single valid category.

4) QUEST

QUEST stands for Quick, Unbiased, Efficient Statistical Tree. It is a relatively new binary tree-growing algorithm [6]. It deals with split field selection and split-point selection separately. The univariate split in QUEST performs approximately unbiased field selection. That is, if all predictor fields are equally informative with respect to the target field, QUEST selects any of the predictor fields with equal probability. QUEST affords many of the advantages of C&RT, but, like C&RT, trees can become unwieldy. You can apply automatic cost-complexity pruning to a QUEST tree to cut down its size. QUEST uses surrogate splitting to handle missing values.

B. DATA SET

We have used gene expression data set of Rat genome of hippocampal region named as 'ca3-Gene expression profiling in differential cognitive outcomes in aging---CA3' available in GEO of NCBI [7]. A raw data set of total 4544 genes and their expressions over 23 samples (features) have been extracted from above mentioned data set. For pre learning of the models used in this paper the data set have been partitioned in three subsets of Training, Testing and validation data sets having 70%, 20%,10% of the total data set respectively.

C. IMPLEMENTATION

Microsoft Excel has been used for data preparation & manipulation and SPSS Clementine 11.1 has been used for model building and performance evaluation. First

we have used K-means clustering model for K=6 to prepare the pre learning data set of total 4544 gene over 23 features(samples).Once we predicted the six cluster data set, we used this cluster identification as class for training, testing and validating the models by partitioning the pre clustered data set.

We implemented the CRT(C&RT), C5, CHAID, QUEST Models using SPSS Clementine 11.1 computing environment. Firstly all the models were trained with training data set and then tested and validated by testing and validation datasets. The comparative view of various models is presented in next section regarding feature selection (important sample experiments), importance of features and predictive accuracy.

RESULTS AND DISCUSSION

Model analysis of the models used in this paper shows input features selected and their variable importance factor (measure of impact of the feature on classification ranging [0, 1]) that was considered for the pre learning by model implementation algorithms. It is clear from fig.1 to fig.4 that among all the models implemented C5 took more balanced and important features having impact on model building with total of 7 features, CRT having 5 features less than C5 but better compared to CHAID and QUEST. CHAID is having the lowest balanced feature importance but QUEST is marginally better to it but much lower to that of C5 & CRT.

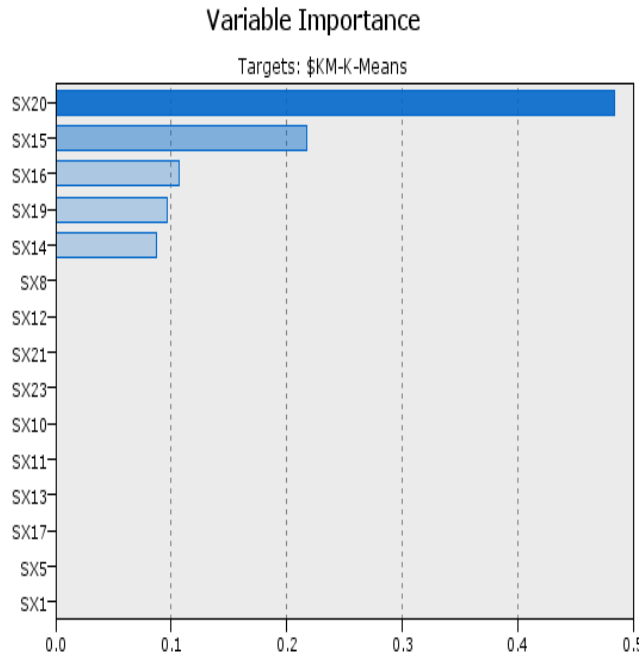

Figure 1. Input Feature and Variable Importance used by CRT

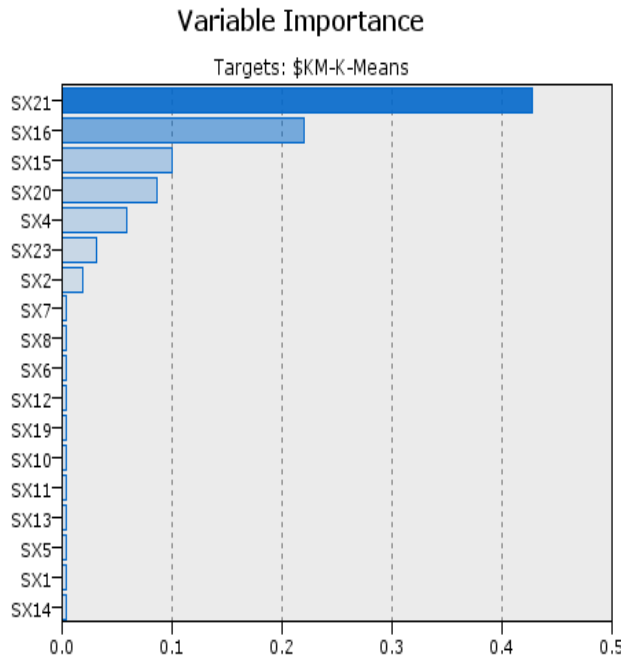

Figure 2. Input Feature and Variable Importance used by C5

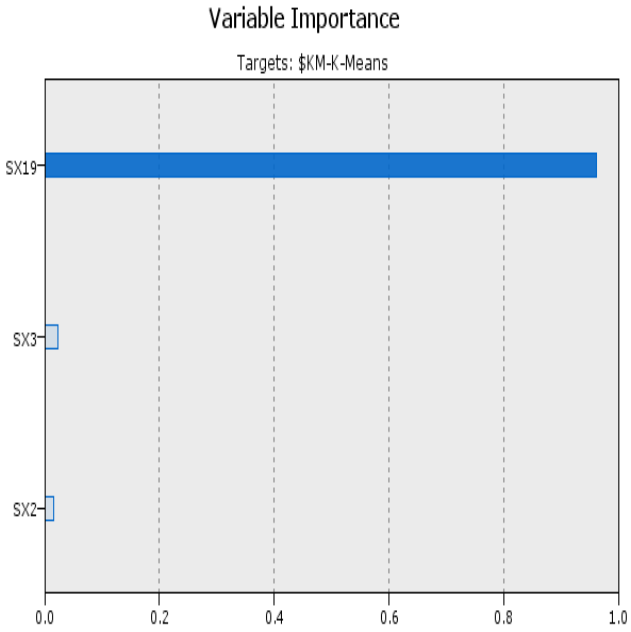

Figure 3. Input Feature and Variable Importance used by CHAID

Fig.5 to fig.6 shows the rule sets (corresponding rule bases of the models) with number of rule corresponding to each cluster class along with number of instances satisfying those rule with the rule confidence factor between 0 to 1. It is clear from the rule sets that rule set generated by C5 contains more number of rules for each cluster class having highest rule confidence factor. CRT is having least number of rules but the confidence of rules are very high. While CHAID is having least confidences of the rules generated. QUEST having more rules than CRT having good confidences of rules.

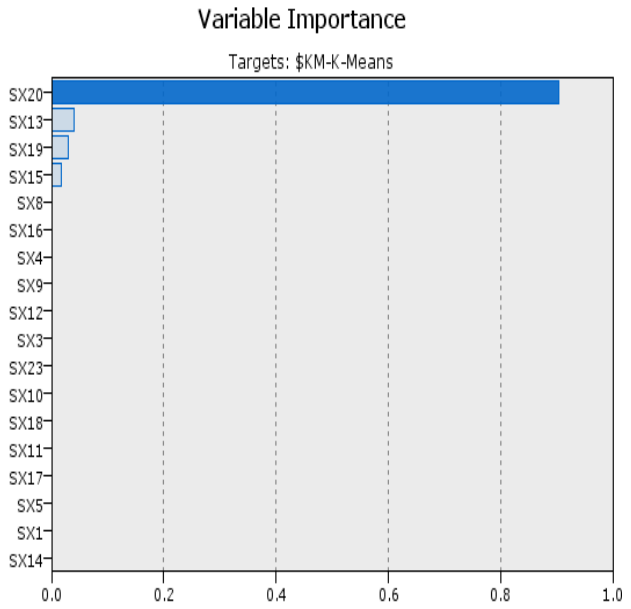

Figure 4. Input Feature and Variable Importance used by QUEST

Rules for cluster-1 - contains 1 rule(s)
Rule 1 for cluster-1 (179; 0.994)
Rules for cluster-2 - contains 1 rule(s)
Rule 1 for cluster-2 (1,505; 0.996)
Rules for cluster-3 - contains 1 rule(s)
Rule 1 for cluster-3 (439; 0.959)
Rules for cluster-4 - contains 1 rule(s)
Rule 1 for cluster-4 (382; 0.966)
Rules for cluster-5 - contains 1 rule(s)
Rule 1 for cluster-5 (369; 0.957)
Rules for cluster-6 - contains 1 rule(s)
Rule 1 for cluster-6 (290; 0.941)
Default: cluster-2

Figure 5. Rule Set (Instances, Confidence) build by CRT.

Table 1. Results obtained by models for pre learning.

S.No	Model	Data Set	% Accuracy of Pre-learning		
			% Accuracy For Training Data set	% Accuracy For Test Data set	% Accuracy For Validation Data set
1	**CRT**	Total Genes in Training data set= **3,164** Total Genes in Test data set= **932** Total Genes in Validation data set = **448**	97.76	97.21	95.54
2	**C5**		99.81	98.18	96.88
3	**CHAID**		86.57	84.23	85.49
4	**QUEST**		96.49	96.78	96.43

Increasing % Correctness

Rules for cluster-1 - contains 2 rule(s)
Rule 1 for cluster-1 (3; 1.0)
Rule 2 for cluster-1 (179; 0.994)
Rules for cluster-2 - contains 3 rule(s)
Rule 1 for cluster-2 (1,480; 1.0)
Rule 2 for cluster-2 (19; 1.0)
Rule 3 for cluster-2 (3; 1.0)
Rules for cluster-3 - contains 8 rule(s)
Rule 1 for cluster-3 (2; 1.0)
Rule 2 for cluster-3 (16; 0.938)
Rule 3 for cluster-3 (7; 0.857)
Rule 4 for cluster-3 (401; 1.0)
Rule 5 for cluster-3 (5; 1.0)
Rule 6 for cluster-3 (2; 1.0)
Rule 7 for cluster-3 (11; 0.818)
Rule 8 for cluster-3 (3; 1.0)
Rules for cluster-4 - contains 6 rule(s)
Rule 1 for cluster-4 (4; 1.0)
Rule 2 for cluster-4 (10; 1.0)
Rule 3 for cluster-4 (4; 1.0)
Rule 4 for cluster-4 (355; 0.997)
Rule 5 for cluster-4 (16; 1.0)
Rule 6 for cluster-4 (5; 1.0)
Rules for cluster-5 - contains 6 rule(s)
Rule 1 for cluster-5 (2; 1.0)
Rule 2 for cluster-5 (5; 1.0)
Rule 3 for cluster-5 (6; 1.0)
Rule 4 for cluster-5 (4; 1.0)
Rule 5 for cluster-5 (337; 1.0)
Rule 6 for cluster-5 (3; 1.0)
Rules for cluster-6 - contains 7 rule(s)
Rule 1 for cluster-6 (2; 1.0)
Rule 2 for cluster-6 (4; 1.0)
Rule 3 for cluster-6 (2; 1.0)
Rule 4 for cluster-6 (266; 1.0)
Rule 5 for cluster-6 (3; 1.0)
Rule 6 for cluster-6 (2; 1.0)
Rule 7 for cluster-6 (3; 1.0)
Default: cluster-2

Figure 6. Rule Set (Instances, Confidence) build by C5

Rules for cluster-1 - contains 1 rule(s)
Rule 1 for cluster-1 (316; 0.573)
Rules for cluster-2 - contains 2 rule(s)
Rule 1 for cluster-2 (1,265; 1.0)
Rule 2 for cluster-2 (317; 0.744)
Rules for cluster-3 - contains 5 rule(s)
Rule 1 for cluster-3 (32; 0.656)
Rule 2 for cluster-3 (34; 0.912)
Rule 3 for cluster-3 (250; 1.0)
Rule 4 for cluster-3 (32; 0.969)
Rule 5 for cluster-3 (34; 0.971)
Rules for cluster-4 - contains 1 rule(s)
Rule 1 for cluster-4 (285; 0.937)
Rules for cluster-5 - contains 1 rule(s)
Rule 1 for cluster-5 (283; 0.788)
Rules for cluster-6 - contains 1 rule(s)
Rule 1 for cluster-6 (316; 0.636)
Default: cluster-2

Figure 7. Rule Set (Instances, Confidence) build by CHAID

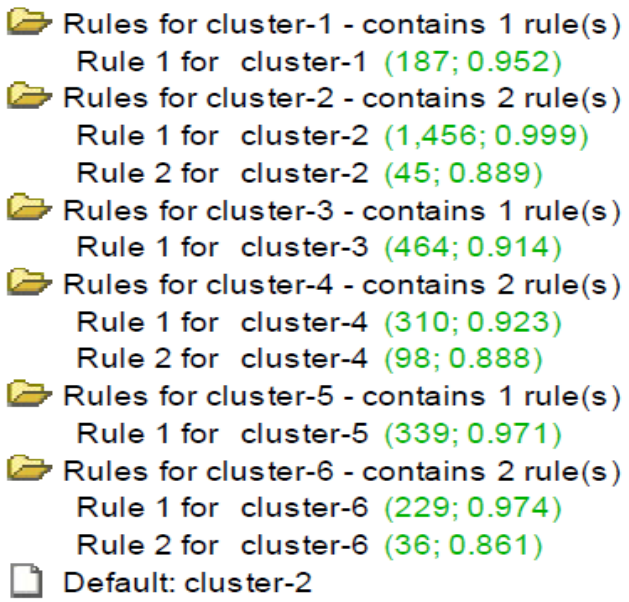

Figure 8. Rule Set (Instances, Confidence) build by QUEST

As we have discussed above that C5 model have more balanced feature selection as well as rule set with highest confidence factor, the training, testing and validation accuracy of the model is also highest (Shown in Table.1). CRT and QUEST models are also having good performance in terms of accuracy CRT nearly less than 1 % to C5 and QUEST less than 1 % to CRT. The performance of CHAID is Lowest and is nearly 12 to 14 % less compared to other models.

CONCLUSION

In this paper for the data set used and for same sources C5 has shown better performance in terms of unbiased and balanced reduced feature selection compared to other models. CRT is marginally less than C5 in terms of feature selection and accuracy. QUEST has not shown balanced feature selection but performance accuracy is good. CHAID has shown comparatively poor results in both feature selection and performance accuracy. In future work the models can be evaluated by pruning and thresholding the importance factors of the features as well as parameter changes for various models used in this paper.

REFERENCES

1. Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. Classification and Regression Trees. New York: Chapman & Hall/CRC.
2. Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.
3. Quinlan, J. (1996). Bagging, Boosting, and C4.5, Proceedings of the
4. Thirteenth National Conferenceon Artificial Intelligence, Portland ,
5. Oregon (American Association for Artificial Intelligence Press,
6. Menlo Park, California), pp. 725 – 730.
7. Rulequest Research. (2013) See5/c5.0.[Online]. Available:
8. http://www.rulequest.com/see5-info.html.
9. Kass, Gordon V.; An Exploratory Technique for Investigating Large Quantities of Categorical Data, Applied Statistics, Vol. 29, No. 2 (1980), pp. 119–127.
10. Loh, W. Y., and Y. S. Shih. 1997. Split selection methods for classification trees. Statistica
11. www.ncbi.nlm.nih.gov
12. Clementine® 11.1 Algorithms Guide , Copyright © 2007 by Integral Solutions Limited

© 2013; AIZEON Publishers; All Rights Reserved

This is an Open Access article distributed under the terms of the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

*****