An Investigation into the impact of the Feature Subset Selection Methods for Classification of Gene Expression Profiles of Microarray Dataset

[1]J.Jeyachidra, [2]M.Punithavalli,

[1] Research Scholar, Department of Computer Science and Applications, Periyar Maniammai University, Vallam, Thanjavur, Tamilnadu, India
[2] Professor, Dept. of MCA, Sri Ramakrishna Engineering College, Coimbatore, Tamilnadu, India.
chithu_raj@yahoo.co.in, mpunitha_srcw@yahoo.co.in

ABSTRACT

In the field of machine learning and pattern recognition, feature subset selection is an important area, where many approaches have been proposed. In this paper, the authors have chosen six feature selection algorithms and analyzed their performance using only one dataset called colon tumor dataset from the public domain. The authors selected the reduced number of features 10, 20, 30, 40 and 50 and calculated their accuracy with respect to the number of features, compared and analyzed with six feature selection algorithms. Hence, the author has advanced a contemporary gene expression data based on machine learning with the help of six feature selection algorithms. The same authors has already published their research work of selecting data with six algorithms. This paper extends their research work further on to performance metrics.

Keywords *:* ***Feature selection, Microarray data, Classification, Algorithms, Gene Expression, Algorithms***

1. INTRODUCTION

Microarray technology provides an opportunity for the researchers to analyze thousands of gene expression profiles simultaneously that are relevant to different fields including cancer related medicine. "By comparing different expression of genes patterns with standard expression profile, any irregularity may be recognized and diagnosed before it becomes serious for the patient. The actual problem is managing microarray data with its dimension. Since the dimension of microarray is large, classifying and handling the algorithm becomes too complex to study the gene expression characteristics. Due to the presence of more inappropriate attributes in the dataset, the accurateness of the classification algorithm also gets affected significantly. To handle those inappropriate attributes, many feature selection algorithms have been experimented by the research society. The aim of feature selection algorithm is to segregate the most important features from the microarray data to minimize the feature space in order to improve the accuracy of the classification [7]".

The authors already studied and analyzed three feature selection algorithms and compared their accuracy. Out of three algorithms, the Chi Square method performed better accuracy than other two algorithms which had been presented in conference [8]. In addition to that, the authors further calculated the accuracy with the most popular another statistical method called T-Test which has been published [9]. In addition to those four algorithms, the remaining two methods namely relief-f and information gain carried out in this paper. As a whole, all the six methods compared and analyzed. This paper is the extension of the previous work. The researchers explored the impact and the quality of the features selected and compared by the following six different feature selection algorithms for the classification of gene expression profiles of microarray data which had been tested with two different classification algorithms Bayes and C4.5 (For C4.5, the researchers used the Weka's implementation of C4.5 called J48). The performance has been validated using Leave –One-Out Cross validation (LOOCV) by considering accuracy as metrics. The research report showed that the classifier was able to achieve equally good results with the first 50 selected features of six feature selection algorithms.

- Gini Index
- Chi Square
- MRMR

- T-Test
- Relief-F
- Information Gain

2. OBJECTIVES AND SCOPE

Recent advances in microarray technology allowed the scientists to measure expression levels of thousands of genes simultaneously and determine whether the genes were active , hyperactive, or dormant in normal or cancerous tissues. Since the microarray device generated huge amount of raw data, many of the genes were irrelevant to distinguish. It was critical to identify a subset of informative genes from a large data that would give higher classification accuracy. Accuracy was very important in cancer classification since it helped diagnose accuracy. The objectives of the research were :

- To study about the existing algorithms and its behavior
- To eliminate the redundant, inappropriate data and to improve the quality of data analysis
- To save the computing space by eliminating unwanted data
- To develop an effective and efficient new algorithm to maximize classification accuracy and perform study

3. FEATURE SELECTION

Features selection was an useful preprocessing technique in data mining and it was used to reduce the dimensions of the data and improve the classification accuracy. Feature selection has become the main focus of research in data mining area. The feature selection quite became difficult and time consuming because of the nature of the data – like supervised and unsupervised learning ones. As a result, a high number of features could lead to lower classification accuracy. So, the main advantage of using feature subset selection was to remove redundant or irrelevant features from the dataset as it could lead to improve the classification accuracy performance. [2] had highlighted that the representation and quality of the incident data was first and foremost. The irrelevant or redundant data would make already the complex knowledge discovery all the more difficult.

4. MICROARRAY DATA

Microarray experiments provided an expression information of large number of genes at different conditions. The raw microarray data images had to be transformed into gene expression matrices. In the matrix table, the row represented by genes and the column represented by various samples such as tissues or experimental conditions. The numbers in each cell characterized, the expression level of the particular gene in the particular sample. The matrices had to be analyzed further to gain more knowledge. The gene expression matrix analysis could be studied by two ways.

(i) Comparing expression profiles of genes by studying the rows in the expression matrix.
(ii) Comparing expression profiles of samples by analyzing the columns in the matrix.

Microarray data for cancer classification consisted of large number of genes(dimensions) compared to the number of samples. The advent of microarray technology enabled the researchers to rapidly measure the levels of thousands of genes expressed in a biological tissue sample. One of the important applications of the microarray was to classify the tissue samples using their gene expression profiles of cancer. It was compared with the standard profiles. Microarray data was highly specialized, involved several variables which were complex to express and analyze. The challenge of the microarray data was to discover and extract useful and meaningful information from the datasets.

5. PREVIOUS WORKS

Several previous researchers [11, 12, 13, 14] were involved in the study of goodness of a feature subset in determining an optimal one. The basic feature selection was an optimization problem. [4] adopted a method called "G-S" algorithm for classification and prediction for the same data set. The mean and standard deviation were computed for each gene's level of expression. It determined according to how close its gene values were to the respective gene value for each class. Both groups achieved reasonable results for their methods to classify new samples. [3] described the use of a

supervised learning algorithm to identify patterns in gene expression data. Their data has 6817 genes or features, and their method gave reasonable results for classifying the samples.

Microarray data analysis was conducted by [1] for cancer classification. An automated system was developed for consistent cancer analysis based on gene microarray expression data. The researchers used the microarray datasets which included both binary and multi-class cancer problems.

6. FEATURE SELECTION ALGORITHMS

As stated earlier, the six popular feature selection algorithms, which were selected for the study were being explained again (as a ready reckoner) even though the same were explained in the previous paper [7].

6.1. Gini Index

The Gini coefficient or Index is a measure of inequality developed by the Italian statistician Corrado Gini and published in his 1912 paper "Variabilite e mutabilita". It is usually used to measure income inequality. The Gini coefficient is often calculated with the more practical Brown Formula shown below:

$$G = \left| 1 - \sum_{k=1}^{n} (X_k - X_{k-1})(Y_k + Y_{k-1}) \right|$$

6.2. Chi Square

Chi-Squared is the common statistical test.The formula for chi-square is

$$\chi^2(f) = \sum_{v \in V} \sum_{i=1}^{m} \frac{(A_i(f=v) - E_i(f=v))^2}{E_i(f=v)}$$

The Chi-square of a feature *f* is defined as, where *Ai(f = v)* is the number of instances in class *ci* with *f = v* and *Ei(f = v)* is the expected value of *Ai(f = v)*, calculated as *P(f = v)P(ci)N*.

6.3. MRMR

Maximum Relevance-Minimum Redundancy(MRMR) is the scheme in feature selection to select the features that correlate the strongest with a classification variable. Maximal Relevance is to search feature set S satisfying [6].

$$Max\ D(S,c), D = \frac{1}{S} \sum_{x_i \in S} I(x_i : c);$$

$I(X_i;c)$ means the mutual information between feature X_i and class c . MRMR also uses the mutual information between feature as redundancy of each feature. The following condition finds the Minimal Redundancy feature set R:

$$Min\ R(S),\ R = \frac{1}{|S|}\ 2 \sum_{x_i, x_j \in s} I(X_i, X_j)$$

Where **$I(x_i,x_j)$ indicates the mutual information between feature x_i and x_j.**

The criterion combining above two conditions is called " Minimal-Redundancy and Maximal-Relevance(MRMR). The MRMR measures has the following form to optimize D and R simultaneously:

$$Max\ \varphi(D,R), \varphi = D - R$$

Where D and R means relevance and redundancy of each feature.

6.4. T-test

The t-test is another common statistical test. The formula for the t-test is provided below:

$$t = \frac{\mu_E - \mu_C}{\sqrt{\frac{\mathrm{var}_E}{N_E} + \frac{\mathrm{var}_C}{N_C}}}$$

and also the t-statistic formula is given below.

$$\Delta = \frac{\left(\frac{\mathrm{var}_E}{N_E} + \frac{\mathrm{var}_C}{N_C}\right)^2}{\frac{\left(\frac{\mathrm{var}_E}{N_E}\right)^2}{N_E - 1} + \frac{\left(\frac{\mathrm{var}_C}{N_C}\right)^2}{N_C - 1}}$$

Where Δ gives the degrees of freedom.

6.5. RelieF

Relief-F is a feature selection strategy that chooses instances randomly, and change the weights of the feature relevance based on the nearest neighbor. By its merits, Relief-F is one of the most successful strategies in feature selection. The main idea of

Relief is to compute a score for each feature measuring how well this feature separates neighboring examples in the original space.

6.6 Information Gain

Information Gain(IG) is another method for attribute selection. The information gain of a feature *f* is defined as

$$G(f) = -\sum_{i=1}^{m} P(c_i)\log P(c_i) + \sum_{v \in V}\sum_{i=1}^{m} P(f = v)P(c_i \mid f = v)\ \log P(c_i \mid f = v)$$

Where $\{C_i\}_{i=1}^{m}$ denotes the set of classes, $v \in$ V is the set of possible values for feature *f 10].*

[5] stated that it can generalize to any number of classes.

7. CLASSIFIERS USED, PERFORMANCE EVALUATION AND VALIDATION METHOD

The authors evaluated selected feature subsets using two learning algorithms – Bayes Classifier and C4.5 classifier. Classifier performance depended on the characteristics of the

data. The performance of the selected algorithms was measured for Accuracy.

The Accuracy could be defined as follows:

Accuracy = (TP+TN) / (TP + FP + TN + FN)

Where TP was the number of True Positives
TN was the number of True Negatives
FP was the number of False Positives
FN was the number of False Negatives

In this work, the authors had used leave-one-out cross validation (LOOCV) for evaluating the performance. LOOCV was a special case of k-fold cross validation.

8. RESULTS AND DISCUSSION

About The Implementation

The researchers used the feature selection tool box called 'fspackage' of Arizona State University for doing their experiments. This toolbox would have feature selection algorithms implemented in compiled (mex format) as well as un-compiled MATLAB code (.m format). Further, it would use some of the features of Weka datamining tool in the form of a library. The authors developed a MATLAB application based on the tool box for their evaluation. WEKA is a well known machine learning tool based on JAVA.

The Colon Tumor Microarray Data Set:

In this study, only one dataset was decided to use because, some of the previous works had used and highlighted the complexity of the data set. The dataset contains 62 samples collected from Colon Tumor patients. Among them, 40 tumor biopsies were from tumors (labeled as "negative") and 22 normal (labeled as "positive") biopsies were from healthy parts of the colons of the same patients. Each sample was described by 2000 genes. Ultimately, the data set contained 62 x 2000 continuous variables and 2000 class ids. Negative was designated as 1 and positive as 2 for the ease of handling inside MATLAB code.

The Table 1 shows the accuracy of classification by Bayes while using the first 10, 20, 30, 40 and 50 features selected by six feature selection algorithms. The metrics were calculated by doing leave-one-out cross validation. For that input parameter, the Gini and Chi Square methods had provided better accuracy compared to the remaining algorithms [8].

Table 1 : LOO Cross Validation using 10, 20, 30, 40 and 50 Features using Bayes Classifier

S.No	Feature Selection Method	Bayes - Accuracy (%)				
		10 Features	20 Features	30 Features	40 Features	50 Features
1.	Gini Index	87.10	88.71	85.48	85.48	85.48
2.	Chi Square	87.10	88.71	85.48	85.48	85.48
3.	MRMR	85.48	83.87	83.87	83.87	85.48
4.	T-Test	69.35	72.58	72.58	72.58	72.58
5.	Relief-F	85.48	85.48	85.48	85.48	85.48
6.	Information Gain	85.48	83.87	83.87	83.87	85.48

Table 2 : LOO Cross Validation using 10, 20, 30, 40 and 50 Features using J48 Classifier

S.No	Feature Selection Method	J48 - Accuracy (%)				
		10 Features	20 Features	30 Features	40 Features	50 Features
1.	Gini Index	85.48	83.87	83.87	83.87	83.87
2.	Chi Square	85.48	83.87	83.87	83.87	83.87
3.	MRMR	85.48	85.48	82.26	82.26	82.26
4.	T-Test	70.97	74.19	79.03	79.03	75.81
5.	Relief-F	79.03	83.87	85.48	83.87	83.87
6.	Information Gain	85.48	85.48	85.48	83.87	83.87

The Table 3 shows the comparison between Bayes Classifier accuracy and J48 classifier accuracy with respect to the number of features [8, 9].

S.No	Feature Selection Algorithm	Accuracy (%)									
		Bayes 10 Features	J48 10 Features	Bayes 20 Features	J48 20 Features	Bayes 30 Features	J48 30 Features	Bayes 40 Features	J48 40 Features	Bayes 50 Features	J48 50 Features
1.	Gini Index	87.10	85.48	88.71	83.87	85.48	83.87	85.48	83.87	**85.48**	**83.87**
2.	Chi Square	87.10	85.48	88.71	83.87	85.48	83.87	85.48	83.87	**85.48**	**83.87**
3.	MRMR	85.48	85.48	83.87	85.48	83.87	82.26	83.87	82.26	**85.48**	**82.26**
4.	T-Test	69.35	70.97	72.58	74.19	72.58	79.03	72.58	79.03	**72.58**	**75.81**
5.	Relief-F	85.48	79.03	85.48	83.87	85.48	85.48	85.48	83.87	**85.48**	**83.87**
6.	Information Gain	85.48	85.48	83.87	85.48	83.87	85.48	83.87	83.87	**85.48**	**83.87**

Table 3: Comparison between Bayes Classifier accuracy and J48 classifier accuracy

In the Table 3, shows that the Gini Index and Chi Square methods had provided better accuracy in Bayes classifier than compared to J48 Classifier. Fig.1 shows the Comparison of Maximum Accuracy Between Between Bayes and J48 for 50 number of features.

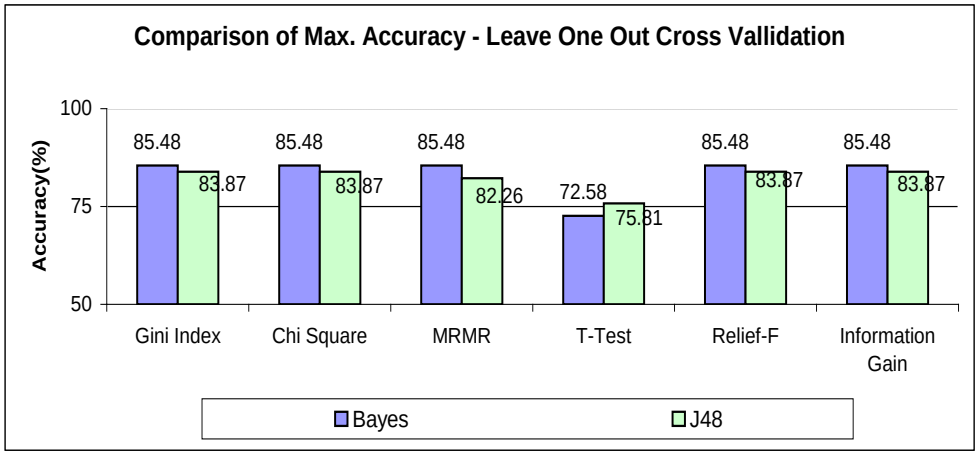

Fig. 1 : The Accuracy Found Through LOOCV for 50 Number of Features

The Table 4 shows the top 10 primary features selected by different feature selection algorithms and the time required for all the six feature selection algorithms to select the features [7].

Table 4 : The Top 10 Primary Features According to Different Algorithms

S.No	Feature Selection Method	Time Taken (sec)	Index of the First 10 Selected Features
1.	Gini Index	4.83	1671, 249, 493, 765, 1423, 513, 1771, 245, 267, 1772
2.	Chi Square	1.02	1671, 249, 493, 765, 1423, 513, 1771, 245, 267, 1772
3.	MRMR	5.48	1671, 249, 493, 765, 1772, 625, 1042, 1423, 513, 1771
4.	T-Test	0.02	1772, 1582, 513, 1771, 780, 138, 515, 625, 1325, 43
5.	Relief-F	1.45	267, 245, 249, 1423, 822, 765, 1892, 66, 493, 897
6.	Information Gain	0.61	1671, 249, 493, 765, 1772, 625, 1042, 1423, 513, 1771

Fig.2 shows the time taken by six different algorithms. If only 100 primary features were selected, then MRMR consumed a lot of time because of increase in number of required features [7].

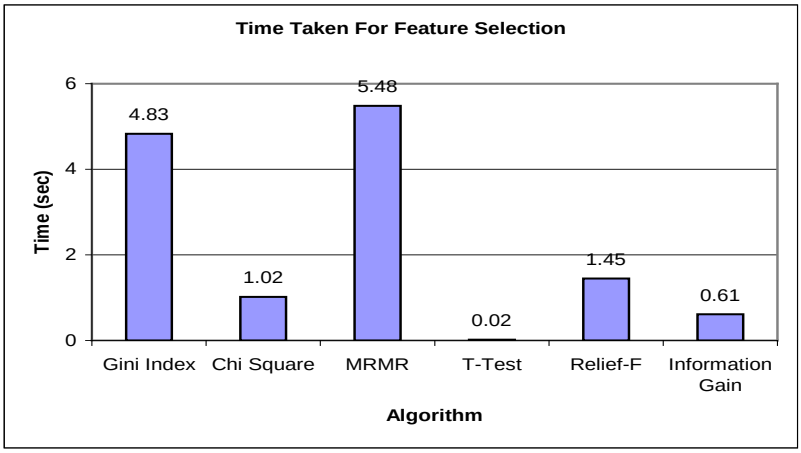

Fig. 2 : The Time Taken By The Six Feature Selection Algorithms

The Fig.2 shows performance of the feature selection algorithms in terms of run time. As shown graph, the performance of the MRMR was poorer than that of other algorithms. Even though the time consumed by T-Test was very low, it provided a poor performance in terms of accuracy.

9. CONCLUSION

In this paper, the researchers had examined the results of six different feature selection algorithms on a sample microarray dataset and the report had shown that the comparative analysis of the six different feature selection algorithms. The Table – 4 shows the order of the selected features were different from one another. Also, the different locations in the dataset showed much difference in the time taken by all the six different feature selection algorithms.

Further, the authors studied the impact of number of selected features such as 10, 20, 30, 40 and 50 features and the accuracy of the classification. In that evaluation, it was observed that the classifier was able to achieve equally good results in all the cases for the first 50 selected features. It was also experimented that even with low number of features selected by one particular algorithm, the classifier was able to produce high accuracy.

But in that evaluation, while considering 10 and above features, MRMR produced almost same results like other compared algorithms with respect to LOO cross validation. Further MRMR also taken much time compared to other algorithms. Even the time consumed for MRMR was only for finding first 100 features and all other algorithms were able to sort all 2000 features in order. For finding more than 100 features, MRMR had taken more time. According to the analysis made by the researchers, T-Test was the only algorithm performed poorly compared to the remaining algorithms.

From the result analysis one could conclude that MRMR method would take more time with better accuracy compared to other algorithms. Alternatively, the T-Test method was faster but the accuracy was poor compared other algorithms.

REFERENCES

[1] A. Osareh and B. Shadgar, "Microarray Data Analysis for Cancer Classification", *5th International Symposium on Health Informatics and Bioinformatics (HIBIT)*, 2010.

[2] A.L. Blum and P. Langley(1997), "Selection of Relevant Features and Examples in Machine Learning", *Artificial Intelligence*, 97: 245-271.

[3] A. Califano, G. Stolovitzky and Y.Tu (2000), "Analysis of Gene Expression Microarrays for Phenotype Classification", *Proceedings of the Annual Intelligent Systems in Molecular Biology*, 8:75-85.

[4] D.K. Slonim, P. Tamayo, J. Mesirov, T. Golub and E. Lander (2000), "Class Prediction and Discovery using Gene Expression Data", *Proceedings of the 4th Annual International Conference on Computational Molecular Biology*, pp. 263-272. Tokyo, Japan: Universal Academy Press.

[5] George Forman. "An Extensive Empirical Study of Feature Selection Metrics for Text Classification", *Journal of Machine Learning*. Res., 3:1289-1305, 2003.

[6] H. Peng, F.Long and C.Dong " Feature Selection Based on Mutual Information : Criteria of Max-Dependency, Max Relevance, and Min-Redundancy", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2005, Vol. 27, No. 8, pp. 1226-1238.

[7] J. Jeyachidra and M. Punithavalli, " An Evaluation of the Performance and Characteristics of Feature Selection Algorithms using Gene Microarray Dataset", *European Journal of Scientific Research*, Vol.No: , pp: 214 – 225, December 2012.

[8] J. Jeyachidra and M. Punithavalli, "A Comparative Analysis of Feature Selection Algorithms on Classification of Gene Microarray Dataset", in IEEE sponsored International Conference on Information, Communication and Embedded Systems – ICICES 2013", IEEE Conference Proceedings, Pp: 1086- 1093, 2013.

[9] J. Jeyachidra and Dr.M.Punithavalli, "A Study On Statistical Based Feature Selection Methods for Classification Of Gene Microarray Dataset", Journal of Applied and Theoretical Information Technology, 10th July 2013, Vol . 53, No.1 , pp: 107-114.

[10] J. Yang, and V. Honavar, (1997), "Feature Subset Selection using a Genetic Algorithm", *Proceedings of the Genetic Programming Conference*", pages 380-385. Stanford, CA.

[11] L. Yu, H. Liu, "Feature Selection for High Dimensional Data", A Fast Correlation-Based Filter Solution, Proceedings- International Conference ICML 2003, pp. 856-863, 2003.

[12] M. Dash, H. Liu and H. Motoda, "Consistency based Feature Selection", *Knowledge Discovery and Data Mining Proceedings*, 1805, pp. 98 -109, 2000.

[13] R. Kohavi and G.H. John, "Wrappers for Feature Subset Selection", *Applied Intelligence* 97, pp. 273 – 324, 1997.

[14] T. Djatna, Y.Morimoto, " A Novel Feature Selection in the Classification Algorithms for Strongly Correlated Attributes using Two Dimensional Discriminant Rules, *6th IETCE Data Engg. Workshop*, 2008.