A STUDY ON STATISTICAL BASED FEATURE SELECTION METHODS FOR CLASSIFICATION OF GENE 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, Department of Computer Applications, Sri Ramakrishna Engineering College, Coimbatore, Tamilnadu, India.

E-mail: chithu_raj@yahoo.co.in , mpunitha_srcw@yahoo.co.in

ABSTRACT

With the rapid development of computer and information technology, an enormous amount of data in science and engineering had been generated in massive scale. Also, the diversity of data, the data mining tasks and approaches pose many challenges to research in data mining. The data mining field has widespread applications including advance diagnosis, market analysis, business management and decision support. But in medical field, illness in common and cancer in particular have become more and more complex and complicated one. To solve the problem in data mining, knowledge discovery tools had been used mainly in research environment. The data mining algorithms are important tool and the most extensively used approach to classify gene expression data and play an important role for classification. Classification is a data mining task and is an effective method to classify the data in the process of knowledge discovery .One of the major challenges faced by many scientists today is the analysis of the explosion of microarray gene expression data. This research is based on machine learning particularly microarray gene expression data analysis . In this paper, the authors have analyzed and compared the two statistical based feature selection algorithms namely Chi Square and T – Test Methods.

Keywords: *Feature Selection, Gene, Microarray, Data Mining, Machine Learning*

1. INTRODUCTION

Microarray technology provided an opportunity for the researchers to analyze thousands of gene expression profiles simultaneously that are relevant to different fields including medicine especially cancer. The categorization of patient gene expression profile has become a common study in biomedical research. The real problem is managing microarray data with its dimension. Since the dimension of microarray is large, classifying and handling the algorithms becomes too complex to study the gene expression characteristics. Due to the presence of more improper attributes in the dataset, the accuracy of the classification algorithm also gets affected significantly. To handle that, several feature selection algorithms have been experimented by the previous researchers. The aim of feature selection algorithm is to isolate the most important features from the microarray data to minimize the feature space in order to improve the accuracy of the classification.

The authors already studied and analyzed three feature selection algorithms and compared their accuracy. Out of three, the Chi Square method performed better accuracy than other two algorithms [6]. Now, the authors explored a study on Chi Square method with the another most popular statistical method T-Test. This paper is the extension of the previous work.

In this work, the researchers explore the impact and the quality of the features selected by the following two 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 shows that the classifier was able to produce equally good results with the first 50 selected features of two feature selection algorithms.

- Chi Square
- T-Test

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

2. OBJECTIVES AND SCOPE

Recent advances in microarray technology allows the scientists to measure expression levels of thousands of genes simultaneously and determine whether the genes are active , hyperactive, or inactive in normal or cancerous tissues. The objectives of the research are :

- To eliminate the redundant or inappropriate data
- To improve the quality of data analysis
- To improve the classification accuracy

3. FEATURE SELECTION

The feature selection or variable selection or attribute selection are same and it was a heuristic for selecting the splitting criterion that " best" separated a given data partition. Feature selection measures were also known as splitting rules because they determined how the tuples at a given time were to be splitted. The feature selection provided a ranking for each attribute describing the given training tuples. The feature had the best score for the measure was chosen as the splitting attributes for the given tuples. It 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 aim of feature selection was to remove the redundant data and improve the classification accuracy.

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. The table, where the row represented by genes and the column represented by various samples such as tissues or experimental conditions and numbers in each cell characterize, the expression level of the particular gene in the particular sample. This matrices had to be analyzed further to gain the knowledge. The gene expression matrix analysis could be studied by two ways.

- Comparing expression profiles of genes by comparing rows in the expression matrix.
- Comparing expression profiles of samples by comparing columns in the matrix.

One of the important applications of microarray data was to classify the tissue samples using their gene expression profiles of cancer and it was compared with the standard profiles.

5. PREVIOUS WORKS

Many successful feature selection algorithms had been devised and the survey of feature selection algorithms might be found in [10]. Several previous researchers [2, 4, 14,15] were involved in the study of goodness of a feature subset in determining an optimal one. The basic feature selection was an optimization problem.

In the paper [3] suggested the well organized choice of discriminative genes from microarray gene expression data for cancer diagnosis. In his study [7] demonstrated about Dimension Reduction for Classification with Gene Expression Microarray Data.

The study report [9] proposed the approach for cancer classification using an expression of very few genes. There are two types involved in this method. The first type is important gene selection which was done by the use of the gene ranking scheme. The second type is the classification accuracy of gene combination carried out by using a fine classifier. .A new approach described in the paper called "Sparse Representation" using Microarray gene expression profiles for cancer diagnosis. Nine human tumor types were used as data set in their research [12].

In this study [13] projected a tumor discovery modus operand as of mammogram. Extracting features which categorized tumors. Microarray data analysis was conducted by [11] 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. In the report [1] stated that Microarray gene expression data had a large number of dimensions. The Support Vector Machine classifier was used for cancer classification with the microarray gene expression data.

6. Feature Selection Algorithms

As stated earlier, the two popular feature selection algorithms, which were selected for this study were being explained in detail again, even though the same were explained in the previous paper [5].

6.1 Chi Square

Chi-Squared was the common statistical method based. The formula for chi-square was

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

(Equa. 1)

6.2 T-Test

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

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

(Equa. 2)

and also the t-statistic formula is given below.

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

(Equa. 3)

Where Δ was the degrees of freedom.

7. CLASSIFIERS

7.1 Bayes Classifier

Bayes classifiers are statistical measures. A simple Bayes classifier was known as the Naïve Bayes Classifier. It exhibited high accuracy and speed. It was a simple induction algorithm that assumed a conditional independence model of attributes given the label (Domingos & Pazzani 1996, Langley, Iba & Thompson 1992, Duda & Hart 1973, Good 1965). The Naive Bayes classifier applied to learning tasks where each instance *x* was described by a conjunction of attribute values and where the target function f *(x)* could take on any value from some finite set V.

7.2 C4.5 Classifier

C4.5 algorithm was proposed by Quinlan (1993). The C4.5 algorithm generated a classification-decision tree for the given data-set by recursive partitioning of data. C4.5 was the most popular and the most efficient algorithm in Decision tree-based approach.

J48 is Weka's Implementation of C4.5 algorithm.

8. THE METRICS USED FOR PERFORMANCE EVALUATION

Classifier performance depended on the characteristics of the data to be classified. Performance of the selected algorithms is measured for Accuracy. The accuracy and error rate can be defined as follows [8].

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

9. VALIDATION METHOD

In this work, the authors have used leave-one-out cross validation for evaluating the performance.

Leave-One-Out Cross-Validation

LOOCV was a special case of k-fold cross validation where k was set to the number of initial tuples. It was repeated, n times, for each of the n observations and the mean square error was computed.

10. RESULTS AND DISCUSSION ABOUT THE IMPLEMENTATION

The researchers used the feature selection tool box called 'fspackage' provided by Arizona State University

for doing this experiments. The authors developed a MATLAB function based on this tool box for the evaluation [5].

The Colon Tumor Microarray Data Set:

The researchers decided to use the colon tumor data set for this study and selected this data set because, some of the previous works were used and highlighted the complexity of this data set. This dataset consists of 62 samples collected from Colon Tumor patients. Among them, 40 tumor biopsies are from tumors (labeled as "negative") and 22 normal (labeled as "positive") biopsies are from healthy parts of the colons of the same patients. Each sample is described by 2000 genes. So, the data set contains 62 x 2000 continuous variables and 2000 class ids (we represented the negative as 1 and positives as 2 for the ease of handling inside MATLAB code).

The Table-1 shows the accuracy of classification by Bayes and J48 (C4.5) while using the first 10 features selected by different feature selection algorithms. The metrics were calculated by doing leave-one-out cross validation. In this case, the Chi Square method provided better performance than the T-Test method.

Table-1 : LOO Cross Validation Using 10 Features

S.No	Feature Selection Method	Bayes Accuracy	J48 Accuracy
1.	Chi Square	87.10	85.48
2.	T-Test	69.35	70.97

The Table-2 shows the accuracy of classification while using the first 20 features selected by different feature selection algorithms. For this input parameter also, the Chi Square method performed better performance than the T-Test.

Table-2 : LOO Cross Validation Using 20 Features

S.No	Feature Selection Method	Bayes Accuracy	J48 Accuracy
1.	Chi Square	88.71	83.87
2.	T-Test	72.58	74.19

The Table-3 shows the accuracy of classification while using the first 30 features selected by different feature selection algorithms.

Table-3 : LOO Cross Validation Using 30 Features

S.No	Feature Selection Method	Bayes Accuracy	J48 Accuracy
1.	Chi Square	85.48	83.87
2.	T-Test	72.58	79.03

The Table-4 shows the accuracy of classification while using the first 40 features selected by different feature selection algorithms.

Table- 4 : LOO Cross Validation Using 40 Features

S.No	Feature Selection Method	Bayes Accuracy	J48 Accuracy
1.	Chi Square	85.48	83.87
2.	T-Test	72.58	79.03

The Table-5 shows the accuracy of classification while using the first 50 features selected by different feature selection algorithms. While using 50 features, Chi Square method provided better performance than T-Test.

Table- 5 : LOO Cross Validation Using 50 Features

S.No	Feature Selection Method	Bayes Accuracy	J48 Accuracy
1.	Chi Square	85.48	83.87
2.	T-Test	72.58	75.81

The Table -6 shows a comparative analysis of 10, 20, 30, 40 and 50 features using Bayes classifiers

Table-6 : 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.	Chi Square	87.10	88.71	85.48	85.48	85.48
2.	T-Test	69.35	72.58	72.58	72.58	72.58

The Table -7 shows a comparative analysis of 10, 20, 30, 40 and 50 features using J48 classifiers

Table-7 : 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.	Chi Square	85.48	83.87	83.87	83.87	83.87
2.	T-Test	70.97	74.19	79.03	79.03	75.81

The Table – 8 shows the comparison between Bayes Classifier accuracy and J48 classifier accuracy.

Table 8: Comparative Analysis Of 10, 20, 30, 40 And 50 Features By Bayes And J48 Classifiers

S. No	Feature Selection Method	Bayes - Accuracy (%)					J48 - Accuracy (%)				
		10 Features	20 Features	30 Features	40 Features	50 Features	10 Features	20 Features	30 Features	40 Features	50 Features
1.	Chi Square	87.10	88.71	85.48	85.48	85.48	85.48	83.87	83.87	83.87	83.87
2.	T-Test	69.35	72.58	72.58	72.58	72.58	70.97	74.19	79.03	79.03	75.81

The Figure-1 shows the accuracy of classification by Bayes and J48 (C4.5) while using the first 50 features selected by the most popular two statistical methods Chi Square and T-Test.

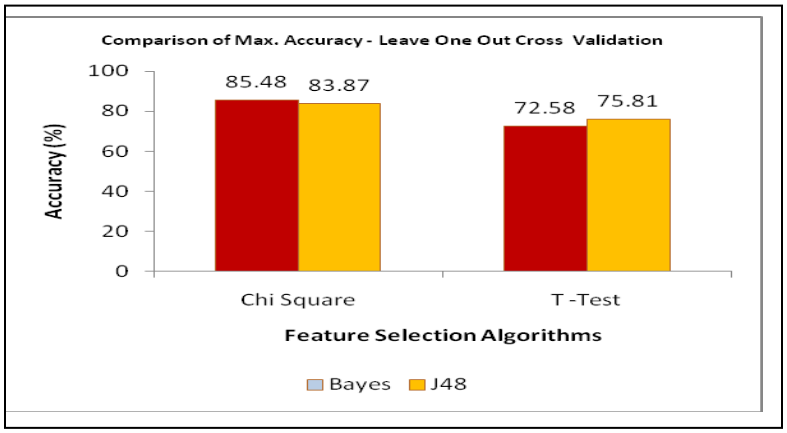

Figure 1 : The Accuracy Found Through LOOCV

The Table-9 shows the Top 10 Primary Features selected by the two feature selection algorithms and the time required for the two methods to select the features.

Table 9 : The Top 10 Primary Features According To Two Statistical Methods

S.No	Feature Selection Method	Time Taken (sec)	Index of the First 10 Selected Features
1.	Chi Square	1.02	1671, 249, 493, 765, 1423, 513, 1771, 245, 267, 1772
2.	T-Test	0.02	1772, 1582, 513, 1771, 780, 138, 515, 625, 1325, 43

The Figure-2 shows the time taken by the two different algorithms.

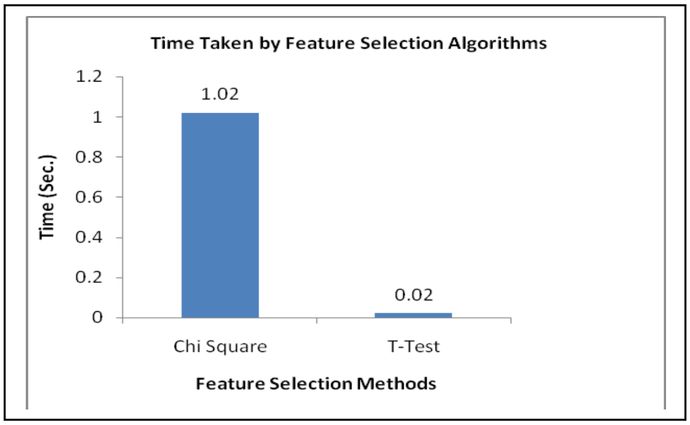

Figure 2 : The Time Taken By The Two Feature Selection Algorithms

The above chart shows performance of the feature selection algorithms in terms of run time. As shown in the graph, the performance of the T-Test was poorer than the Chi Square. Even though the time consumed by T-Test is very low, the performance in terms of accuracy is very poor.

11. CONCLUSION

In this paper, the researchers have examined the results of two different statistical based feature selection algorithms on a sample microarray dataset. The two algorithms selected the first few primary features based on different criteria are given in the Table– 9, but the order of the selected different features which were present entirely different from one another. But in this evaluation, while considering 10 and above features, according to the analysis made by the researchers, Chi Square performed better accuracy than T-Test.

If the authors observed closely the results of this study, the authors could say that, there was a opportunity in which, two different feature selection algorithms might be chosen completely different set of features as primary features and even a good classification algorithm might be capable of classifying dataset by using these two different “primary” feature sets and arrive same level of precision.

REFERENCES

[1]. Chu.F. and L.Wang, “ Applications of Support Vector Machines to Cancer Classification with Microarray Data”, International Journal of Neural Systems, Vol. 15, No.6, 475 – 484, 2005.

[2]. Djatna T. 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.

[3]. Huang.D, Chow.T.W.S, Ma.E.W.M and Jinyan Li, “ Efficient Selection of Discriminative Genes from Microarray Gene Expression Data for Cancer Diagnosis”, IEEE Transactions on Circuits and Systems, 2005.

[4]. Iffat A.,Gheyas, Leslie S.Smith, “ Features Subset Selection in Large Dimensionality Domains”, Pattern Recognition 43, pp. 5-13, 2010

[5]. Jeyachidra .J. and Punithavalli .M. ,”An Evaluation of the Performance and Characteristics of Feature Selection Algorithms using Gene Microarray Dataset”, in European Journal of Scientific Research, Vol.93, No. 2 December, 2012, pp.214 – 225.

[6]. Jeyachidra .J. and Punithavalli .M.,”A Comparative Analysis of Feature Selection Algorithms on Classification using Gene Microarray Dataset”, in the Proceedings of IEEE Sponsored International Conference on Information, Communication and Embedded Systems 2013(ICICES 2013), pp. 1106 – 1111.

[7]. Jian J. Dai, Linh Lieu and David Rocke, “ Dimension Reduction for Classification with Gene Expression Microarray Data”, Statistical Applications in Genetics and Molecular Biology, Vol. 5 No.1, Pp. 1-21, 2006.

[8]. Jiawei Han and Micheline Kamber, “Data Mining Concepts and Techniques”, Second Edition, 2006, pp: 361-363

[9]. Lipo Wang, Feng Chu, and Wei Xie, “ Accurate Cancer Classification using Expressions of Very Few Genes”, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 4, Pp: 40-52, 2007.

[10]. Molina, L.C. Belanche, L., Nebot, A. : “Attribute Selection Algorithms: A survey and Experimental Evaluation . Proceedings of 2nd IEEE KDD, pp. 306 – 313, 2002.

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

[12]. Xiyi Hang, “Cancer Classification By Sparse Representation Using Microarray Gene Expression Data”, IEEE International Conference on Bioinformatics And Biomedicine Workshops, Pp. 174-177, 2008.

[13]. Y. Ireaneus Anna Rejani, Dr.S.Thamarai Selvi, “ Early Detection of Breast Cancer using SVM Classifier Technique and Engineering” Vol 1, Issue 3, pages 127-130, 2009.

[14]. Yang J. V. Honavar, Feature Subset Selection Using a Genetic Algorithm", IEEE Intelligent Systems and their Applications 13, pp. 44-49, 1998.

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