ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

# Classification of Microarray Data Based On Feature Selection Method

C.Lavanya[1], M.Nandihini[2], R.Niranjana[3], C.Gunavathi[4]
Computer Science and Engineering, K.S.Rangasamy College of Technology, Tiruchengode, Tamilnadu, India [1, 2, 3, 4]

**Abstract- Genes are encoding regions that form necessary building block inside the cell and show the way to proteins which are achieving a variety of functions. However, some genes may get mutated. Such genes are responsible for cancer occurrence. It can be discovered by closely examining samples taken from patients to identify faulty genes. Gene expression dataset usually comes with only dozens of tissues/samples but with thousands or even tens of thousands of genes/features. In this paper, we employ feature selection techniques for analyzing cancer microarray gene expression data. Feature selection technique is used to select the most possibly cancer-related genes from huge microarray gene expression data. It aims to achieve improved classification performance. This can be achieved by the measures of T-Test, Chi-Square Test and Information gain. Cancer classification using microarray data poses another major challenge because of the huge number of genes compared to the number of tissue samples. Only a small number of genes in the microarray data which consisting of thousands of genes show strong correlation with the target phenotypes. This paper presents the Naive Bayes algorithm for the classification task. A comprehensive framework that incorporates feature selection and classification techniques is capable of successfully classifying new samples as infected or normal.**

**Index Terms-Gene Expression Data, Classification, Feature selection method, Naive Bayes algorithm**

## I. INTRODUCTION

Data mining is the computational process of analysis of large quantities of data. It uses information from past data to analyze the outcome of a particular problem or situation that may arise. Data classification is the form of data analysis that extracts models of describing important data classes. Such models called classifiers, predict categorical class labels. Such analysis can help us with better understanding of data at large. The class label of each training label is provided a state called supervised learning that is the learning of the classifier of supervised in that it is told to a class each training tuple belong. In learning, training data are analyzed by classification algorithm in which test data are used to estimate the accuracy of classification rules.

### 1.1 BASICS OF GENE EXPRESSION DATA

Gene expression is the process by which information from a gene is used in the synthesis of a functional gene product. These products are often proteins, but in non protein coding genes such as rRNA genes or tRNA genes, product is a structural or housekeeping RNA. Gene expression studies can also involve looking at profile or patterns of expression of several genes whether quantitating changes in expression levels or looking at overall patterns of expression, real time PCR is used by most scientists performing gene expression. Based on the levels of the gene expression data optimized genes are classified based on different classifiers.

### 1.2 MICROARRAY DATA CLASSSIFICATION

The micro array data are images, which have to be transformed into gene expression matrices in which rows represent genes, columns represent various samples such as tissues or experimental conditions, and numbers in each cell characterizes the expression level of particular gene in the particular sample. Microarray based disease classification system takes labeled gene expression data samples and generates a classifier model that classifies new data samples into different predefined diseases. Microarray data classification is a supervised leaning task that predicts the diagnostic category of a sample from its expression array phenotype.

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

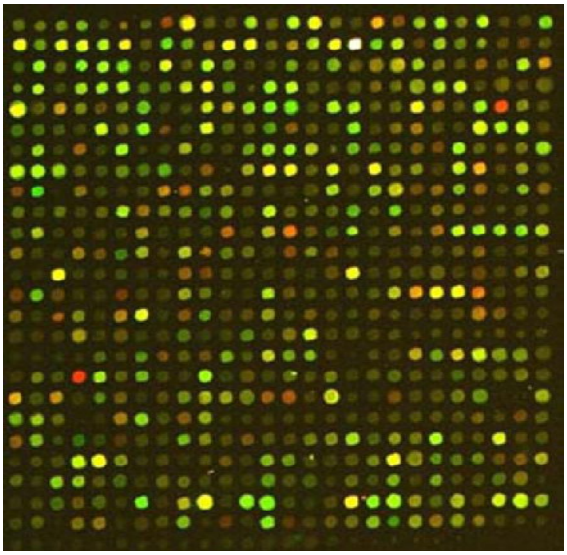

Fig.1 Microarray data

## II. GENE EXPRESSION DATA SETS

The datasets considered in the simulation are Iris, yeast, Spellman dataset, breast cancer. All these data sets are publicly available and are two class gene expression

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | t1 | t2 | t3 | t4 | t5 | t6 | t7 | t8 | t9 | t10 |
| 2 | YAL022C | 3.8518 | 1.4433 | 4.9007 | 1.5214 | 0.41538 | -1.6848 | -0.04249 | -2.3666 | -3.0103 | -0.06228 |
| 3 | YAL023C | -2.4834 | -1.8468 | -1.8762 | -4.0699 | -1.8187 | -4.0882 | -2.8394 | -2.4193 | -2.2955 | -0.83477 |
| 4 | YAL026C | -0.83282 | -0.03952 | 1.1755 | 0.061383 | -0.15474 | 0.3497 | 2.3273 | -0.31494 | 1.1737 | -0.04292 |
| 5 | YAL037W | 0.95071 | -0.34414 | 1.7837 | 1.375 | 0.14011 | 0.90882 | -0.31151 | 0.65269 | 0.025381 | 0.23071 |
| 6 | YAL041W | -2.9395 | -2.7089 | -3.2279 | -5.6413 | -2.2626 | -4.3877 | -4.3092 | -2.9473 | -2.8837 | -1.6481 |
| 7 | YAL042W | 0.86046 | 1.4121 | 0.70091 | 1.5236 | 0.536 | 1.8386 | 1.4524 | 0.93259 | 1.4878 | 0.62257 |
| 8 | YAL043C | -2.3066 | -1.9819 | -2.6073 | -4.7824 | -2.2387 | -4.2046 | -3.0809 | -2.895 | -2.3703 | -0.88976 |
| 9 | YAL043C-A | 0.59475 | 0.74273 | 1.3922 | 1.359 | 0.99807 | 1.1322 | 1.2846 | 1.2975 | 1.0213 | 0.49802 |
| 10 | YAL044C | 0.13819 | 0.51711 | 0.28241 | 0.6641 | 0.34171 | 1.5442 | 1.0322 | 1.0142 | 0.84372 | 0.34719 |
| 11 | YAL045C | -0.85836 | -2.7762 | -2.94 | -2.832 | -3.1648 | -4.5947 | -3.3343 | -4.1272 | -4.5737 | -2.8244 |
| 12 | YAL054C | -0.61552 | -0.8198 | -0.29818 | -0.84141 | -0.75644 | -1.1779 | -1.1553 | -0.6179 | -0.60902 | 0.4404 |
| 13 | YAL063C | -0.61299 | 0.055744 | -0.16914 | -0.73895 | -0.1452 | -0.39563 | 0.644 | 0.10609 | 0.21114 | -0.57642 |
| 14 | YAR007C | -1.1401 | -0.68046 | -0.17562 | -0.93679 | -0.26384 | 0.10037 | -0.69386 | -0.20379 | -0.8507 | -0.4815 |
| 15 | YAR008W | -0.89949 | -0.32658 | -0.45516 | 0.28005 | -0.68723 | -0.03708 | -0.17731 | 0.031561 | -0.41564 | -0.55937 |
| 16 | YAR009C | 0.37513 | 0.57632 | -0.4956 | 0.27061 | -0.28603 | 0.40515 | -0.53192 | -0.65724 | 0.45586 | 0.034053 |
| 17 | YAR050W | -0.03397 | 0.62255 | -2.586 | 0.40751 | -0.69945 | 2.1786 | -0.29562 | -2.1935 | 3.1602 | 0.14045 |
| 18 | YBL007C | 0.40774 | 0.40606 | 0.15697 | 0.63259 | 1.1127 | 0.8843 | 1.0171 | 0.85515 | 0.99982 | 0.37357 |
| 19 | YBL008W | 0.060519 | -0.33747 | -1.0013 | -0.95188 | -0.81554 | -0.54217 | 0.25262 | 0.39317 | 0.16779 | -0.35719 |
| 20 | YBL017C | -0.41402 | 0.16599 | -0.08462 | -0.08169 | 0.045784 | 0.82145 | 0.54198 | -0.24443 | 1.0108 | -0.24005 |
| 21 | YBL029W | 0.75188 | -1.2895 | 1.2904 | 2.3651 | 0.89355 | 0.63978 | -0.29606 | 0.97384 | -0.78985 | 0.37852 |
| 22 | YBL030C | -0.10457 | -0.89976 | 0.55978 | 0.25046 | 0.37137 | 0.34062 | 0.2064 | -0.0273 | -0.73219 | 0.28942 |
| 23 | YBL038W | -0.41717 | -0.14104 | -0.01782 | -0.4331 | -0.06168 | -0.34599 | -0.09384 | -0.18862 | -0.25844 | 0.2349 |
| 24 | YBL039C | -1.5146 | -1.7394 | -1.4437 | -3.001 | -1.1918 | -2.1556 | -2.2566 | -2.0463 | -1.7803 | -0.61371 |
| 25 | YBL079W | 1.711 | 2.2494 | 2.4589 | 4.1205 | 1.5702 | 3.4163 | 3.5612 | 2.2029 | 2.3966 | 0.44653 |

Fig.2 Dataset

### 2.1 GENE SELECTION

In this study, a number of gene selection methods have been introduced to select informative genes. The different dataset genes are classified using classifier like SVM, Naive Bayesian and optimized genes obtained through feature selection methods like T-Test, information gain and mutual information.

## III. FEATURE SELECTION METHODS

The importance of feature selection methods is selecting informative genes prior to classification of microarray data for cancer prediction and diagnosis. Feature selection method removes irrelevant and redundant features to improve classification accuracy. Feature selection methods can be categorized into filter, wrapper, and embedded or hybrid. The filter approach selects features without involving any data-mining algorithm. The filter algorithms are evaluated based on four different evaluation criteria namely, distance, information, dependency and consistency. The wrapper approach selects feature subset based on the classifier and ranks feature subset using predictive accuracy or cluster goodness. It is more computationally expensive than the filter model.

### 3.1 T –TEST

To measure the relevance of a gene, the t-test is widely used, assuming that there are two classes of samples in a gene expression data set. When there are multiple classes of samples, the t-test is computed for one class versus all the other classes. It compares the actual difference between two means in relation to the variation in the data. T Test values are calculated as

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

where $\overline{x_1}$ and $\overline{x_2}$ are means, $s_1$ and $s_2$ are variance of samples, and $n_1$ and $n_2$ are number of subjects.

### 3.2 CHI-SQUARE TEST

Chi-square is a statistical test commonly used to compare observed data with data we would expect to obtain according to a specific hypothesis. This is

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

designed specifically for multiple tests having at least two discrete outcomes (such as normal and mutated gene). The chi-square test is always testing what scientists call the null hypothesis, which states that there is no significant difference between the expected and observed result. The formula for calculating chi-square ($x^2$) is:

$$x^2 = \sum_{i=1}^{\text{colomns}} \sum_{j=1}^{\text{rows}} \frac{(\text{observed}_{ij} - \text{expected}_{ij})^2}{\text{expected}_{ij}}$$

A Chi Square Test is often used to measure a goodness of fit between an observed and expected distribution of values knowing how to perform a Chi Square test can be useful for testing probable to expected outcomes, fitting points to a curve, or testing a statistical hypothesis.

### 3.3 INFORMATION GAIN

Information gain, of a term measures the number of bits of information obtained for category prediction by the presence or absence of the term in a document. Information Gain measures the decrease in entropy when the feature given is absent. This is the application of a more general technique, the measurement of informational entropy, to the problem of deciding how important a given feature is. Informational entropy, when measured using Shannon entropy, is notionally the number of bits of data it would take to encode a given piece of information. The more space a piece of information takes to encode, the more entropy it has. Intuitively, this makes sense because a random string has maximum entropy and cannot be compressed, while a highly ordered string can be written with a brief description of the string's information. In the context of classification, the distribution of instances among classes is the information in question. If the instances are randomly assigned among the classes, the number of bits necessary to encode this class distribution is high, because each instance would need to be enumerated.

On the other hand, if all the instances are in a single class, the entropy would be lower, because the bit-string would simply say "All instances save for these few are in the first class." Therefore function measuring entropy must increase when the class distribution gets more spread out and be able to be applied recursively to permit finding the entropy of subsets of the data. The following formula satisfies both of these requirements:

$IG(X) = H(D) - H(D|X)$ where
$H(D) = - \Sigma (ni/n) \log(ni/n)$ i=1,…l and

$H(D|X) = - \Sigma (|Xj|/n) H(D|X-Xj)$

$p\pm(S)$ is the probability of a training example in the set $S$ to be of the positive/ negative class. We discretized continuous features using information theoretic binning.

For each dataset we selected the subset of features with non-zero information gain. Information Gain can be used only on discrete features and hence for numeric features discretization is necessary prior to computing Information Gain. Entropy-based discretization method is generally used for gene expression data. Similar, to t-Statistic, features are selected based on the larger values of Information Gain.

## IV. CLASSIFIERS

### 4.1 NAIVE BAYES CLASSIFIER

A Naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model". In simple terms, a naive Bayes classifier assumes that the presence or absence of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. Naive Bayes is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

## V. DISCUSSION AND CONCLUSION

We showed how combining a filtering technique for feature selection with SVM leads to substantial improvement in generalization performance of the SVM models in the five classification datasets of the competition. Another lesson learned from our submission is that there is no single best feature selection technique across all five datasets. We experimented with different feature selection techniques and picked the best

ISSN (Online) : 2319 – 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume 3, Special Issue 1, February 2014

International Conference on Engineering Technology and Science-(ICETS'14)

On 10th & 11th February Organized by

Department of CIVIL, CSE, ECE, EEE, MECHNICAL Engg. and S&H of Muthayammal College of Engineering, Rasipuram, Tamilnadu, India

one for each dataset. Of course, an open question still remains: why exactly these techniques worked well together with Support Vector Machines. A theoretical foundation for the latter is an interesting topic for future work.

REFERENCES

[1] AnirbanMukhopadhyay, UjjwalMaulik and Sanghamitra Bandyopadhyay, "An Interactive Approach to Multi-Objective Clustering of Gene Expression Patterns", IEEE Transactions on Biomedical Engineering, vol. 60, no. 1, pp. 35-41, 2013

[2] Feng Yang and K.Z. Mao, "Robust Feature Selection for Microarray Data Based on Mulitiicriiterion Fusion", IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 8, no. 4, pp. 1080-1092, 2011

[3] Garcia-Nieto, E. Albaa, L. Jourdanb and E. Talbi, "Sensitivity and Specificity Based Multi-Objective Approach for Feature Selection: Application to Cancer Diagnosis", Information Processing Letters, vol.109, pp. 887-896, 2010

[4] Jihong Liu and Guoxiong Wang, "A Hybrid Feature Selection Method for Data Sets of Thousands of Variables", IEEE, pp. 288-291, 2010

[5] Jinhuia Sheng, Hong-Wen Deng, Vince D. Calhoun and Yu-Ping Wang, "Integrated Analysis of Gene Expression and Copy Number Data on Gene Shaving Using Independent Component Analysis", IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 8, no. 6, pp. 1568-1578, 2011

[6] Jiye Liang, Feng Wang, Chuangyin Dang and YuhuaQian, "A Group Incremental Approach to Feature Selection Applying Rough Set Technique", IEEE Transactions on Knowledge and Data Engineering, pp. 1-30, 2012

[7] Meng-Yun Wu, Dao-Qing Dai, Yu Shi, Hong Yan and Xiao-Fei Zhang, "Biomarker Identification and Cancer Classification Based on Microarray Data Using Laplace Naive Bayes Model With Mean Shrinkage", IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 9, no. 6, pp. 1649-1662, 2012

[8] Patrick C. H. Ma and Keith C. C. Chan, "Incremental Fuzzy Mining of Gene Expression Data for Gene Function Prediction", IEEE Transactions on Biomedical Engineering, vol. 58, no. 5, pp. 1246-1252, 2011

[9] Shang Gao, Omar AddamandAlaQabaja, "Robust Integrated Framework for Effective Feature Selection and Sample Classification and Its Application to Gene Expression Data Analysis", IEEE, pp. 112-119, 2012

[10] YannChristinat, Bernd Wachmann, and Lei Zhang," Gene Expression Data Analysis Using a Novel Approach to Biclustering Combining Discrete and Continuous Data", IEEE/ACM Transactions On Computational Biology And Bioinformatics, Vol. 5, No. 4, pp. 583-593, October-December 2008

[11] UjjwalMaulik, Anirban Mukhopadhyay and DiebasisChakraboirty, "Gene-Expression-Based Cancer Subtypes Prediction Through Feature Selection and Transductive SVM", IEEE Transactions on Biomedical Engineering, vol. 60, no. 4, pp. 1111-1117, 2013

[12] Yuchun Tang, Yan-Qing Zhang and Zhen Huang, Xiaohua Hu, "Granular SVM-RFE Gene Selection Algorithm for Reliable Prostate Cancer Classification on Microarray Expression Data", in the Proceedings of the 5th IEEE Symposium on Bioinformatics and Bioengineering, 2005

[13] Zhenyu Wang and Vasile Palade, "A Comprehensive Fuzzy-Based Framework for Cancer Microarray Data Gene Expression Analysis", IEEE, pp. 1003-1010, 2007

[14] George Lee, Carlos Rodriguez, and AnantMadabhushi," Investigating the Efficacy of Nonlinear Dimensionality Reduction Schemes in Classifying Gene and Protein Expression Studies", IEEE/ACM Transactions On Computational Biology And Bioinformatics, Vol. 5, No. 3, pp. 368-384, July-September 2008

[15] Zhenyu Wang, Vasile Palade and Yong Xu, "Neuro-Fuzzy Ensemble Approach for Microarray Cancer Gene Expression Data Analysis", in the International Symposium on Evolving Fuzzy Systems, pp. 241-246, 2006

[16] Lipo Wang, Feng Chu, and Wei Xie, " Accurate Cancer Classification Using Expressions of Very Few Genes", IEEE/ACM Transactions On Computational Biology And Bioinformatics, vol 4, no.1, pp.40-53, January-March 2007