Investigation of Micro Array Gene Expression Using Linear Vector Quantization for Cancer

E T Venkatesh
Senior Lecturer, dept. of Computer Technology
Kongu Engineering College
Erode, India

Dr.P. Thangaraj
Dean,dept. of Computer Technology & Applications
Kongu Engineering College
Erode, India

***Abstract*— Genomics refer to the comprehensive study of genes and their task. Micro array analysis or Gene expression profiling provides methods to analyze thousands of genes in a single sample. Micro array analysis is providing challenges in various fields by providing large amount of data which can be processed to obtain useful information. In this paper we study the gene samples obtained from biopsy samples collected from colon cancer patients. We introduce a learning vector quantization method that determines artefacted states and separate malignant genes from regular genes.**

Keywords- **Neural Network, Data mining, Micro array analysis, Colon cancer.**

I. INTRODUCTION

With the availability of large amounts of data it is increasingly becoming more difficult for the user to retrieve non trivial information from them. Knowledge Discovery in Databases (KDD) is the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data [1]. The KDD process is interactive and iterative, involving numerous steps. Data mining is one step of the Knowledge Discovery in Databases (KDD) process.

A DNA microarray is an ordered array of genes immobilized on a planar substrate that allows the specific binding of labeled cDNA [2]. DNA microarray technology is a landscape-changing tool for molecular biology. It allows a genome-wide monitoring approach of gene expression and provides unprecedented opportunity to explore the biological processes underlying human diseases by providing a comprehensive survey of the cell's transcriptional landscape. The raw microarray data is basically an image with different colors indicating hybridization [3] of DNAs expressed at different conditions. The image is further converted into numerical data as pixel intensity reflecting ideally the count of photons corresponding to the amount of transcripts genetically.

The challenges faced in microarray dataset include a very large number of features, typically 2000 – 3000. But, not all of these genes are needed for classification [4]. Most genes do not influence the performance of the classification task. Taking such genes into account during classification increases the dimension of the classification problem, poses computational difficulties and introduces unnecessary noise in the process. A major goal for diagnostic research is to develop diagnostic procedures based on inexpensive microarray that have enough probes to detect certain diseases. This requires the selection of some genes which are highly related to the particular classification problem, i.e., the informative genes. This process is called Gene Selection (GS) [5], which corresponds to feature selection from machine learning in general.

II. DATAMINING AND LVQ NEURAL CLASSIFIERS

Wikipedia defines Data mining as the process of extracting patterns from data. Data mining is becoming an increasingly important tool to transform this data into information. It is commonly used in a wide range of profiling practices, such as marketing, surveillance, fraud detection and scientific discovery.

Data mining can be used to uncover patterns in data but is often carried out only on samples of data. The mining process will be ineffective if the samples are not a good representation of the larger body of data. Data mining cannot discover patterns that may be present in the larger body of data if those patterns are not present in the sample being "mined". Inability to find patterns may become a cause for some disputes between customers and service providers. Therefore data mining is not foolproof but may be useful if sufficiently representative data samples are collected. The discovery of a particular pattern in a particular set of data does not necessarily mean that a pattern is found elsewhere in the larger data from which that sample was drawn. An important part of the process is the verification and validation of patterns on other samples of data.

Linear Vector Quantization (LVQ) can be understood as a special case of an artificial neural network, more precisely, it applies a winner-take-all Hebbian learning-based

approach. It is a precursor to Self-organizing maps (SOM) and related to Neural gas, and to the k-Nearest Neighbor algorithm (k-NN)

Learning Vector Quantisation is a supervised version of vector quantisation, similar to Selforganising Maps (SOM)[6]. Vector quantization is a technique where by the input space is divided into a number of distinct regions, and for each region a reconstruction vector is defined. When presented with a new input x , a vector quantizer first determines the region in which the vector lies. Then the quantizer outputs an encoded version of the reconstruction vector *wi* representing that particular region containing x . The set of all possible reconstruction vectors *wi* is usually called the codebook of the quantizer. When the Euclidean distance similarity measure is used to decide on the region to which the input x belongs, the quantizer is called a Voronoi quantizer[7].

In LVQ networks, class information is used to fine-tune the reconstruction vectors in a Voronoi quantizer so as to improve the quality of the classifier decision regions [8]. In classification problems, it is the decision surface between classes and not the inside of the class distribution that should be described most accurately. The quantizer process can be easily adapted to optimize placement of decision surface between different classes. The method starts with the calibration of a trained Voronoi quantizer using a set of labeled input samples. Each *wi* is then labeled is according to the majority of classes represented among those samples which have been assigned to *wi* . Here the distribution of the calibration samples to the various classes, as well as the relative numbers of the *wi* assigned to these classes, must comply with the priori probabilities of the classes, if the probabilities are known. The tuning of the decision surfaces is done by rewarding correct classifications and punishing incorrect ones. When training pattern *xk* from class *c j* is presented to the network, let the closest reconstruction *wi* belong to class *cl* . Then only vector *wi* is updated according to the following supervised rule

$$\Delta w_i = \begin{cases} +\eta^k (x^k - w_i) & if\ c_j = c_l \\ -\eta^k (x^k - w_i) & if\ c_j \neq c_l \end{cases}$$

III. DATASET USED IN OUR WORK

The colon cancer data is available in Kent Ridge Biomedical Data Repository. The gene expression samples were analyzed with an Affymetrix Oligonucleotide array complementary to more than 6500 human genes. Colon dataset consists of 62 samples of colon epithelial cells taken from colon-cancer patients. Each sample contains 2000 gene expression levels. Although original data consists of 6000 gene expression levels, 4000 out of 6000 were removed based on the confidence in the measured expression levels. 40 of 62 samples are colon cancer samples and the remaining are normal samples. Each sample was taken from tumors and normal healthy parts of the colons of the same patients and measured using high density oligonucleotide arrays [10]

.

IV. EXPERIMENTAL INVESTIGATION AND ANALYSIS

Table 1 lists the classification accuracy using the Neural network algorithm along with its kappa statistic, and absolute errors.

Correctly Classified Instances	87.0968 %
Incorrectly Classified Instances	12.9032 %
Kappa statistic	0.7062
Mean absolute error	0.129
Root mean squared error	0.3592
Relative absolute error	28.08%
Root relative squared error	75.01%

Table 1 : Classification accuracy using Neural Network

Table 2 lists the precision and recall using Neural network based classifier.

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0.727	0.05	0.889	0.727	0.8	Positive
0.95	0.273	0.864	0.95	0.905	Negative

Table 2. Precision and Recall

Figure 1 shows the classification accuracy over other classification methods including Classification and regression tree and sequential minimal optimization.

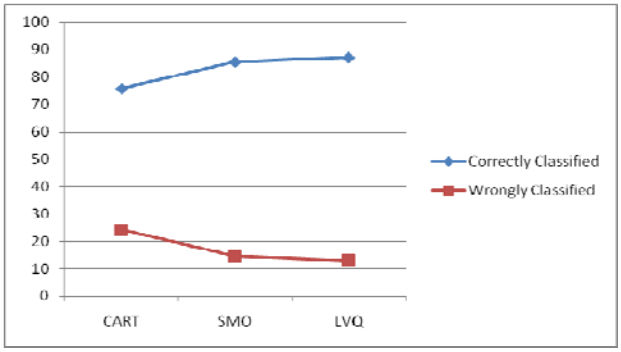

Figure I. Comparison with other classifiers

V. CONCLUSION

In this paper we provide, Micro array analysis of Gene expression profiling methods to analyze thousands of

genes in a single sample. Micro array addresses challenges in various fields by providing large amount of data which can be processed to obtain useful information. In this paper we study the gene samples obtained from biopsies collected from patients. Linear vector quantization was used to classify the provided gene expression samples for colon cancer. LVQ shows promising results compared over other methods.

REFERENCES

[1] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From datamining to knowledge discovery in databases. In AI Magazine, volume 17,pages 37–54, 1996.
[2] Schena M, *Microarray Analysis*, John Wiley & Sons, 2003.
[3] Rui Xue, Jianying Li and Denni J.Streveler. Microarray Gene Expression Profile Data Mining Model for clinical cancer research in Proceedings of the 37th Hawaii International Conference on System Sciences – 2004.
[4] Zhenyu Wang and Vasile Palade. A Comprehensive Fuzzy-Based Framework for Cancer Microarray Data Gene Expression Analysis in IEEE transaction 2007.
[5] L. Li, C. R. Weinberg, T. A. Darden, and L. G. Pedersen, "Gene Selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the ga/knn method," Bioinformatics, vol. 17, pp. 1131-1142, 2001.
[6] H. Hassoun, *Fundamentals of Artificial Neural Networks***,** MIT PRESS, 1996.
[7] Gray, R.M., "Vector quantization," *IEEE Acoustic, Speech, and Signal Processing Magazine*, *1*, pp. 4-29, 1984.
[8] T. Kohonen, *Self-Organizing Maps*, Springer, Berlin, 1995
[9] Abhishek Bansal and G. N. Pillai (2007) : High Impedance Fault Detection using LVQ Neural Networks, International Journal of Computer, Information, and Systems Science, and Engineering
[10] Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I.,Schummer, M. and Yakhini, N. (2000): Tissueclassification with gene expression profiles. *Journal of Computational Biology*, **7**:559-584.