IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 1, Issue 5, Oct-Nov, 2013
ISSN: 2320 - 8791
www.ijreat.org

Dimensionality Reduction for Gene Selection using Clustering and Information Measures

Kishore Kumar P K

Department of Mathematics, Agni College of Technology, Thalambur, Chennai-603103

Abstract

Several gene selection methods have been used to calculate the accuracy of high marker genes from micro array data. The efficient way of selecting the redundant and non-relevant gene from gene expression data is to apply the method of fuzzy equivalence partition matrix(FEPM) which effectively performs the selection from continuous gene expression data set.One of the finest way of selecting the relevant genes is adapted with the method of dimensionality reduction. The reduced gene data set has been chosen by the concept called ICA(Independent Component Analysis). After the reduction these reduced set of genes has been clustered using the method named Non-negative Matrix Factorisation (NMF) along with the extensions of SNMF and NMFSC to calculate the rate of accuracy for certain set of genes. The proposed method shows the best accuracy based on information measures involving class separability index and the comparison between the best cluster accuracy rate determined by the cluster algorithms.
Keywords *– FEPM, ICA, NMF, NMFSC, SNMF*

1.Introduction

Cancer is a genetic disease or disorder that is caused by the damages in genes, which regulates the cell growth and division[3]. The cancerous cells invade and destroy other cells either by direct growth into adjacent cells through invasion or by implantation into distant site through metastasis.There are different types of cancer namely Lung cancer, Blood cancer, Prostrate cancer, Breast cancer etc.,[4]Accurate diagnosis of different types of cancers is of great importance for doctors to choose a proper treatment. However, the similar appearances of some types of cancers are a main challenge for traditional diagnostic methods. In recent years, this problem has attracted great attention in the context of microarray gene expressions because of their ability to differentiate cancers at molecular level. After the expression profiles of cancer cells are obtained, a variety of machine learning or statistical approaches can undertake the diagnosis task. Micro array data analysis has been successfully applied in a number of studies over a broad range of biological disciplines including cancer classification by class discovery and prediction.

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, the product is a functional RNA[8]. The process of gene expression is used by all known life - eukaryotes, prokaryotes and viruses - to generate the macromolecular machinery for life. Microarray technology is a new tool that can automate the diagnostic task and improve the accuracy of the traditional diagnostic techniques[2]. With microarrays, it is possible to examine the expression of thousands of genes at once. Testing for elevated expression of certain genes can assist in predicting cancer.

The recent advancement and wide use of high throughput technology are producing an explosion in using gene expression phenotype for identification and classification in a variety of diagnostic areas. An important application of gene expression data in functionally genomics is to classify samples according to their gene expression profiles.

In most gene expression data, the number of training samples is very small compared to the large number of genes involved in the experiments. However, among the large

amount of genes, only a small fraction is effective for performing a certain task. Furthermore, a small subset of genes is desirable in developing gene-expression-based diagnostic tools for delivering precise reliable, and interpretable results[9]. With the gene selection results, the cost of biological experiment and decision can be greatly reduced by analyzing only the marker genes. Hence, identifying a reduced set of the most relevant genes is the goal of gene selection.

However for real-valued gene expression data, the estimation of different information measures is a difficult task as it requires knowledge on the underlying probability density functions of the data and the integration on these functions. Rough set theory is a new paradigm to deal with uncertainty, vagueness, and incompleteness. It has been applied to fuzzy real extraction, reasoning with uncertainty,fuzzy modelling, feature selection and so forth.

2. Rough Sets and FEPM

Rough set starts with the notion of an approximation space which is a pair <U, A> where U is an Universal Set and A is an attribute set in which a equivalence partition is formed. There are two approximations in rough set namely the lower approximation and the upper approximation respectively. The lower approximation is the union of all elementary sets of X and the upper approximation is the union of all elementary sets that have non-empty intersections with X. A set with lower and upper approximation is called a rough set. Otherwise it is also termed as not definable[2]. The indiscernibility relation IND(P) on U is given by IND(P) = $\{(x_i,x_j) \in U \times U \ / \ a \in P, f_a(x_i) = f_a(x_j)\}$ The partition of U generated by IND(P) is denoted by U / IND(P) = $\{ [x_i]_P : x_i \in U \}$ Also X may

be characterized by lower and upper approximations as $P(x) = \bigcup \{ [x_i]_P / [x_i],X\}$

$P(x) = \bigcup \{ [x_i]_P / [x_i]_P \cap X \neq \varphi\}$

A matrix formed by using the data set partitioned into fuzzy equivalence classes based on these rough sets is used to calculate the information measure. Given a finite set U, A is a fuzzy attribute set in U which generates a fuzzy equivalence partition on U. If c denotes the number of fuzzy equivalence classes[7] generated by the fuzzy equivalence relation and n is the number of objects in U, then the c partitions of U are sets of (cn) values $\{m_{ij}^A\}$ that can conveniently be arrayed as a (c x n) matrix M_A is termed as the FEPM and is denoted by

$$M_A = \begin{bmatrix} m_{11}^A & m_{12}^A & \dots & m_{n1}^A \\ \dots\dots & \dots\dots\dots & & \dots\dots \\ m_{c1}^A & m_{c2}^A & \dots & m_{cn}^A \end{bmatrix}$$

subject to $\sum_{i=1}^{n} m_{ij}^A \ \forall j$

The ith fuzzy equivalence partition matrix is given by $F_i = \{ m_{i1}^A / x_1 + m_{i2}^A / x_1 + \dots + m_{in}^A / x_1 \}$

The cardinality of the fuzzy set F_i can be calculated

by $|F_i| = \sum_{i=1}^{n} m_{ij}^A \ \forall j$

3. Information Measure

Information measure is the amount of information calculated in a particular location. Shannon's entropy is used to calculate the information measure. There are three information measures used to select the relevant and non-redundant genes. They are 1. Mutual information measure I(P,Q), 2. V- information measure and 3. Chi-square information measure

The information quantity of a fuzzy attribute set A or fuzzy equivalence partition is then defined as

$H(A) = - \sum_{i=1}^{n} \lambda Fi \log \lambda Fi$ where λ_{Fi} = ($|F_i|$ / n) called a fuzzy relative frequency, and c is the number of fuzzy equivalence partitons or classes.

3.1 Mutual information measure

Given <U,A> , P and Q are two subsets of A. The information quantity

corresponding to P and Q are given by

$$H(P) = -\sum_{i=1}^{n} \lambda Pi \log \lambda Pi ,$$

$$H(Q) = -\sum_{i=1}^{n} \lambda Qi \log \lambda Qi$$

The joint entropy of P and Q is defined as

$$H(PQ) = -\sum_{i=1}^{n} \lambda Ri \log \lambda Ri$$ Where r is the number of resultant fuzzy equivalence partitions, R_k is the corresponding kth equivalence partition, and λ_{Rk} is the joint frequency of P_i and Q_j which is given by $\lambda_{Rk} = \lambda_{PiQj} = |P_i \cap Q_j| / n$ where $k = (i-1)q + j$ The mutual information between two fuzzy attribute sets P and Q is given by $I(PQ) = H(P) + H(Q) – H(PQ)$

3.2 V- Information measures

In fuzzy approximation spaces[5], one of the simplest measures of dependence can be obtained using the function $V = |x - 1|$ resulting in V- information $V(R \| P \times Q) = \sum_{i,j,k} |\lambda_{Rk} - \lambda_{PiQj}|$ where $P = \{\lambda_{Pi} | i = 1,2,...,p\}$ and $Q = \{\lambda_{Qj} | j = 1,2,...,q\}$ and $R = \{\lambda_{Rk} | k = 1,2,...,r\}$ represent two marginal frequency distributions and their joint frequency distributions, respectively.

3.3 Chi-Square Information measure

The chi-square information measure of fuzzy approxmation spaces can be defined as follows $\chi^2 (R \| P \times Q) = \sum_{i,j,k} |\lambda_{Rk} - \lambda_{PiQj}|^2 / (\lambda_{Pi}\lambda_{Qj})$

4. Independent component Analysis

ICA is a useful extension to principal component analysis(PCA), which was originally developed for blind separation of independent sources from their linear mixtures. It has been used in various applications of auditory signal separation, medical signal processing, and so on. Unlike PCA, where the aim is to decorrelate the dataset, ICA aims to make the transformed[2] coefficients mutually independent. This implies that the higher order dependencies will be removed by the ICA expansion. Considering a pxn data matrix X, whose columns c_j (j = 1…n) represent the observational variables, the ICA model of X can be written as X = SA where S is a pxn source matrix and A an nxn mixing matrix. Vectors s_q the columns of S, are assumed to be statistically independent and are called as the ICs of S that the columns of X are linear mixtures of the ICs. The statistical independence between s_q can be measured by using mutual information $I = \sum_q H(s_q) – H(S)$, where $H(s_q)$ is the marginal entropy of the variable s_q and H(S) the joint entropy of S. Estimating the ICs can be accomplished by finding the right linear combinations of the observational variables. We can invert the mixing matrix such that $S = XA^{-1} = XW$. Then an ICA algorithm is used to find a projection matrix W such that the columns of S are as statistically independent as possible.Several algorithms have been proposed to implement ICA such as FastICA and JADE. FastICA algorithm to model the gene expression data, considering its efficiency in processing large-scale dataset. The FastICA has been widely used to process gene expression data. In FastICA, the mutual information is approximated by

$$J(S_q) = (E\{F(S_q)\} = E\{G(\zeta)\}^2$$

where G is an arbitrary nonquadratic function and ζ a Gaussian distributed variable. ICA can remove linear correlations as well as higher order dependencies in the data. It also allows some flexibility in scaling and sorting by convention. The ICs are generally scaled to unit deviation, while their signs and orders can be chosen arbitrarily.

5. Clustering with NMF

NMF can reduce the dimensionality of expression data from thousands of gene to metagenes. Coupled with a model selection mechanism, NMF can be an efficient method to identify distint molecular patterns and a powerful tool for class discovery. The ability of NMF to recover meaningful biological information from cancer-related microarray data[1]. NMF also appears to have advantages over other methods, such as HC and COMs,

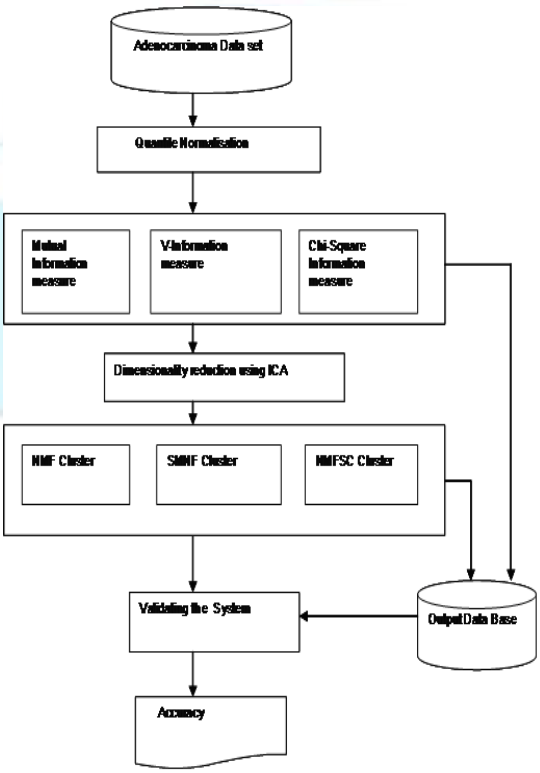
because HC impose a stringent tree structure on the data, and is highly sensitive to the the metric used to assess similarity,and SOM can be unstable, yielding different decompositions of the data with different initial conditions. However, standard NMF cannot control the sparseness of the decomposition, and thus, does not always yield a parts-based representation.

In NMf model, each entry v_{ij} in V is the coefficient gene i in metagene j and each entry h_{ij} in H represents the expression level of metagene i in sample j. In such a factorization, matrix H can be used to group the n sample into k clusters. Each cell sample is placed into a cluster corresponding to the most highly expressed metagene in the sample. That is sample j is placed in cluster i if h_{ij} us the largest entry in column j.

Although NMF has been successfully used in several applications it results in part based representations only. To solve this problem, Hoyer extended the NMF framework by including an adjustable sparseness parameter. SNMF and NMFSC are extensions to those ideas. The main improvement in them is that the sparseness can be adjusted explicitly, rather than implicitly. These algorithms were used for tumor class and clustering. These can also be used to measure the accuracy rate of microarray data set such as adenocarcinoma, leukemia and colon cancer. It is useful in viewing the nature of algorithm for voluminous data set in gene expression. NMF group the samples into clusters, there are several problems that need to be resolved. Among them one key issue is which k can decompose the samples into “meaningful” clusters. Another problem is that the NMF algorithm may or may not converge to the same solution in each run with random initial conditions. So, how to evalutate the stability of clustering associated with a given rank k? This is still an open problem. The basic idea is that if a clustering to k classes is strong, the cluster assignment of samples should not vary much from random starting points. After running with many different random initial points, a consensus matrix C is computed to evaluate the stability of clustering associated with the given k. The entries of C range from 0 to 1 and reflect the probability that each pair of samples cluster together. If a clustering is stable, the entries of C will be close to 0 or 1. The dispersion between 0 and 1 thus measures the reproducibility of the class assignments with respect to the random initial conditions. A reordered matrix of C can be used for visual inspection, which can serve as similarity measure among samples. Quantitatively, the stability for each value of k can be measured through the cophenetic correlation coefficient ρk (C),which indicates the dispersion of the consensus matrix C.

In a perfect consensus matrix, the cophenetic correlation coefficient $\rho k = 1$. When the entries are scattered between 0 and 1, the cophenetic correlation coefficient is less than 1. Roughly speaking, the more stable the cluster assignment is, the greater the coefficient ρk is. Therefore, by observing how ρk evolves as k increases, we can select the value of k when the magnitude of the coefficient starts to fall. This implies that a two-cluster split of the samples is more stable than others.

7. System Design



8. Experimental Results

The input data is cancer gene expression data. It contains only numeric values. Data set used is a Lung cancer data set called Adenocarcinoma. It consists of 675 rows and 156 columns. Using data mining techniques, data can be normalized into [0 1]. The data set collected is been preprocessed using the method called Quantile Normalisation. It is the method of normalising the data set into [0,1] where the range of the numeric value has been rounded off to get the entire data set between 0 and 1. The concept of dimensionality reduction has been introduced to overcome the problem of voluminous data set used in the analysis. The method of applying the reduction technique is to reduce the data set into a smaller one without affecting its semantics. ICA is the algorithm used to reduce the data set. This method has been used to analyse the rate of clusters grouped by the similar genes in a voluminous micro array data. This accuracy rate ensures the percentage of accuracy for the genes in the reduced set. Various algorithms have been used to cluster analysis. Some of the algorithms used in our method is NMF, NMFSC and SNMF

Different information measures have been used to calculate the accuracy of redundant and non-relevant genes. This has been achieved with the reduced set of genes. Some of the information measures used in our method are mutual information measure, v-information measure and chi – square information measure.

The subsystem has been validated using the lung cancer data set named adenocarcinoma for different samples of cancerous patients. This gives an optimum accuracy rate of cancer data set.

The output graphs of information measures obtained from the gene data set and the cluster accuracy rates has been discussed in the below[10]. This has been evaluated for the lung cancer data set and the accuracy has been tabulated. The graphs illustrates the accuracy of gene expression data set given as an input and shows the accuracy rate with class separability index obtained within scatter matrix, between scatter matrix and the vector in the data set.The graphical representation for within distance, between distance and class separability index is shown

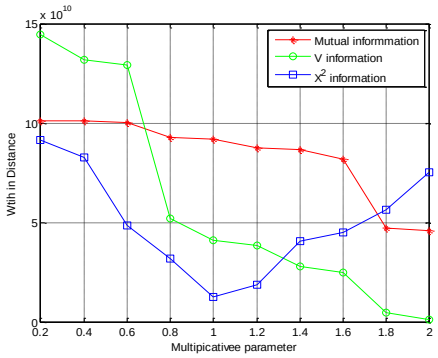


Fig 1-Within class scatter matrix

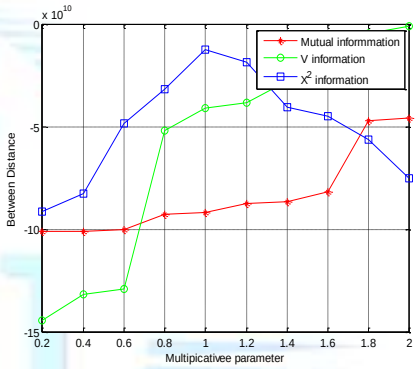


Fig 2 -Between class scatter matrix

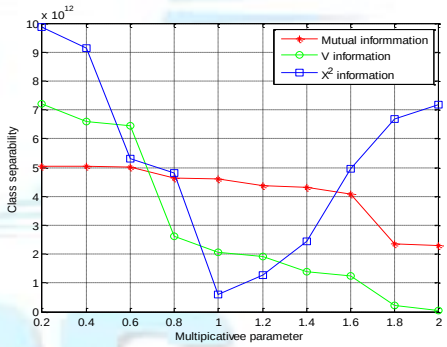


Fig 3 – Class Separability Index

Table showing the class separability distance of genes within class, between class and class separability using three information measures.

The range of multiplicative parameter varies from $1 \leq \eta \leq 1.8$ For mutual information measure, $\eta = 1$ decreases for within class and class separability and increases with between class. For V- information measure, $\eta = 1.8$ decreases for within class and class separability and increases with between class For χ^2 information measure, $\eta = 1$ decreases for within class and class separability and increases with between class

8.1 Cluster Accuracy Rates Using NMF, SNMF and NMFSC with ICA

The following figures illustrates the concepts of cluster accuracy rate using NMF, SNMF and NMFSC.

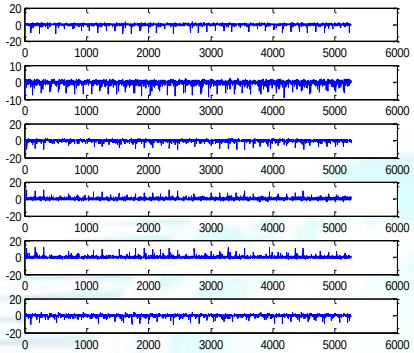


Fig 4 – Dimensionality reduction with ICA

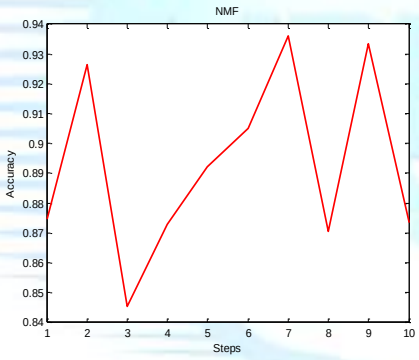


Fig 5 – NMF Cluster graph

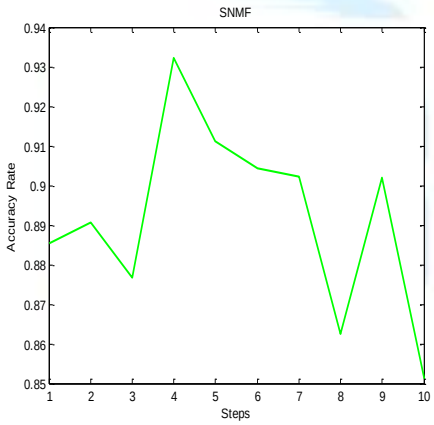


Fig 6 – SNMF Cluster graph

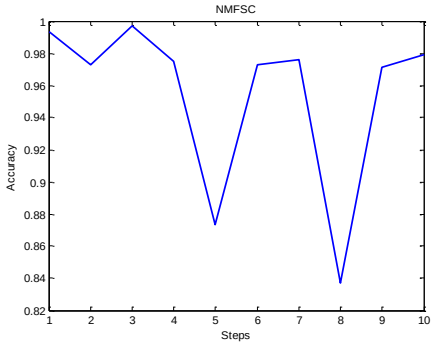


Fig 7 – NMFSC Cluster graph

The results of cluster rate analysis is shown below.

In NMF cluster around 600 genes are clustered at an accuracy rate of 93.85%

In SNMF cluster around 300 genes are clustered at an accuracy rate of 93.15%

In NMFSC cluster around 200 genes are clustered at an accuracy rate of 99.75%

9. Conclusion and Future Work

The selection of relevant and non – redundant gene by the FEPM approach is a successful method to value a continuous gene expression data set which has a true and marginal density in the existing system . The concepts of dimensionality reduction using ICA along with NMF clustering techniques are used to calculate the cluster accuracy rate for different clustering algorithms such as NMFSC and SNMF. Also the information measures shows the accuracy of marker genes using class separability index.

In future the better results may be achieved by other gene selection methods. Therefore the interaction between different gene selection methods and other clustering algorithms should be further explored.

References

[1] Chun-Hou Zheng, De-Shuang Huang, "Tumor clustering using Non-negative matrix factorization with Gene selection" , IEEE Transactions on Information technology in Biomedicine, Vol. 13, No.4, July 2009

[2] Pradipta Maji and Sankar K. Pal, "Fuzzy – Rough sets for information measures and selection of relevant genes from Microarray data", IEEE Transaction on Systems,Man and Cybernetics,Vol.40,No.3,pp741 – 752, June 2010

[3] Shenghuo Zhu, Dingding Wang, Kai Yu, Tao Li and Yihong Gong "Feature selection for gene expression using model-based entropy",IEEE/ACM Transactions on computational and bioinformatics, vol 7, No.1, March 2010

[4] Youngmi Yoon, Sangjay Bien, and Sanghyun Park, "Microarray data classifier consisting of k-top scoring rank-comparison decision rules with a variable number of genes", IEEE Transactions on Systems, man and cybernetics, vol.40, No.2, pp 216 – 226, March 2010

[5] P. Maji, "f - information measures for efficient selection of discriminative genes from microarray data," IEEE Trans. Biomed. Eng., vol. 56, no. 4,pp.1063–1069, Apr. 2009

[6] Jung-Hsien Chiang, "A Combination of Rough-based feature selection and RBF Neural network for classification using Gene expression data",IEEE Transactions on Nanobioscience,Vol.7,No.1 March 2008

[7] P. Maji and S. K. Pal, "Rough set based generalized fuzzy C-means algorithm and quantitative indices," IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 37, no. 6, pp. 1529–1540, Dec. 2007

[8] Q. Hu, D. Yu, Z. Xie, and J. Liu, "Fuzzy probabilistic approximation spaces and their information measures," IEEE Trans. Fuzzy Syst., vol. 14,No. 2, pp.191–201, Apr. 2006

[9] Feng Tan, Xuezheng Fu, Yanqing Zhang, Anu G. Bourgeois, " Improving Feature subset selection using a Genetic Algorithm for Microarray gene expression data", IEEE Transaction, pp 2529 – 2534, July 2006

[10] N. Kwak and C.-H. Choi, "Input feature selection by mutual information based on Parzen window," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 12, pp. 1667–1671, Dec. 2002