Tumor Clustering and Gene Selection Techniques - A Survey

S.Praba
Research Scholar
Bharathiar University
Coimbatore – 638 401, India

A.K.Santra, PhD.
Dean
CARE School of Computer Applications
Trichy – 620 009, India

ABSTRACT

Cancer classification has become one of the active areas of research in the field of medical sciences. Various gene selection and tumor classification techniques have been available in the literature. Gene selection comprises of an exploration for gene subsets that are capable to discriminate tumor tissue from normal tissue. Gene selection is a primary issue in gene expression based tumor classification. Recently, Tissue microarrays have become an extensively used technique to screen for protein expression patterns in a large numbers of tumors. There is increasing interest in transforming the importance of tumor classification from morphologic to molecular. Gene expression profiles provide additional data when compared with morphology and offer a substitute to morphology-based tumor classification systems. So, researchers are very much intentional to develop novel approaches for gene selection and tumor classification. This paper provides a detailed related survey of various gene selection techniques and tumor classification approaches.

General Terms

Survey on Tumor Clustering and Gene Selection Techniques.

Keywords

Gene selection, Tumor Classification, DNA, Genetic Algorithm, Particle Swarm Intelligence.

1. INTRODUCTION

The discovery of cancer classes has conventionally been based on histomorphology. In recent years, DNA microarrays have been effectively employed to automatically identify cancer classes via clustering of the expression profiles. It has been indicated that a number of tumors can be clustered into clinically appropriate groups based exclusively on gene expression (mRNA) profiles [1].

It has been found that tumors that have identical histopathological look may follow significantly different clinical courses with different responses to therapy and thus, inefficient diagnosis is often possible if the diagnosis is based primarily on morphological appearance. Microarray technology is observed to categorize tumor samples based on gene expressions and therefore has been extensively used in systems biology and iatrology [2].In recent years, there is extensive research in transforming the importance of tumor classification from morphologic to molecular [3]. Array technologies have become popular in evaluating the level of expression of thousands of genes concurrently [4, 5 and 6]. A number of research works have used arrays to examine gene expression in various tumors, and these investigations have illustrated the potential utility of expression profiling for classifying tumors [7, 8]. Gene expression profiles may provide additional information than conventional morphology systems.

Gene selection is a crucial part for gene expression based tumor classification systems. The main benefit of microarray is that it is able to competent to monitor the expression of a large number of genes and provide extremely useful biological information.

The identification of discriminant genes is of prime importance and attention. Various investigations in Biology and Medicine are supported by the thorough study of the top ranking genes which would be very helpful in recent discoveries in cancer research. Medical diagnostic examinations that assess the presence of a given protein in serum may be obtained from a little subset of discriminant genes [9].

It is a fact that additional features should give more discriminating power. But, there are several reasons for minimizing the number of features to a sufficient minimum.

Initially, large numbers of features would increase the computational complexity and cost.

Secondly, when treated individually, the two features may provide appropriate classification information, but if integrated, it provides ineffective results due to high mutual correlation. Therefore, complexity increases without much gain.

Thirdly, generalization properties of the classifier will be comprised by a huge number of features. It is to be observed that, the effective generalization attributes of the classifier is obtained through the ratio of higher number of training samples to the number of free classifier parameters [10].

Fourthly, a huge number of features will affect the assessment of the classification error. A small number of features will enhance the assessment of classification error. Thus, minimizing the dimensionality of the gene expression information is a vital concern in developing an efficient gene expression-based tumor classification system [3].

Besides, minimizing noise and enhancing the accuracy of tumor classification, selected subsets of genes with significant accuracy of classification may be involved in certain biological processes which would result in tumor development [3]. The selected subsets of genes may have essential biological understanding and may be utilized for cancer discovery and other future possible research directions.

The main benefits of gene selection over other techniques are minimizing the dimensionality (e.g., principal components), its simplicity, future cost savings, and higher possibility of being adopted in a clinical setting.

2. LITERATURE SURVEY

A number of techniques have been presented in the literature to minimize the dimensionality of gene expression data [11]. A majority of the machine learning approaches have been used in cancer classification using microarray data [12]. Eisen et al., [13] developed one of the earliest techniques called the hierarchical algorithm. Other well known algorithms applied for tumor classification are neural networks, K-Nearest Neighbor (KNN), SVM, kernel based classifiers, genetic algorithms and Self- Organizing Maps (SOM) [14].

Even though, a number of groups have widely regarded model selection in SVMs, optimal parameters are generally domain specific. Yendrapalli et al., [15] proposed a technique to estimate the impact of model selection on the performance of a number of SVM implementations to classify tumors.

The issue of multiclass classification, particularly for techniques like SVMs, does not provide an easy solution. It is usually straightforward to build classifier theory and algorithms for two equally exclusive classes than for N mutually exclusive classes. Yendrapalli et al., utilized BSVM that constructs N-class SVMs [16]. A majority of the existing techniques for model selection utilize the leave-one-out (loo) related estimators which are regarded computationally costly. The author utilized Leave-one-out model selection for SVM (looms) that employs advance numerical techniques which results in efficient calculation of loo rates of different models [17].

It is observed that classification accuracy varies with the kernel type and the parameter values; therefore, with suitably chosen parameter values, tumors can be categorized by kernel machines with higher accuracy and lower false alarms. The results illustrate the capability of using learning machines in diagnosis of a tumor.

2.1 Wrapper and Filter Method

Gene selection is vital in tumor classification with the benefits such as enhancing the classification accuracy, minimizing the cost in a clinical setting and gaining considerable imminent into the mechanism of disease [18, 19, 20, 21].

Recent gene selection techniques mainly categorized into two types [18]:

- Filter methods
- Wrapper methods

Filter model executed the feature subset selection and the classification in two separate phases, employs an evaluation metric that is simple and fast for assessment. Therefore, a filter approach is not dependent of the learning algorithm used after it. The wrapper method formulated the feature subset selection and classification in the same process, links a learning algorithm to evaluate the classification accuracy.

In microarray data, it is tough to identify the best gene subset among all combinations of genes because of its high dimensions. Even though a number of heuristic search approaches can be employed, these techniques are computationally costly [19]. Filter approaches are known as gene ranking techniques in gene expression data area. These approaches identify predictive subsets of the genes through a simple criterion evaluated from the empirical distribution, and the top-genes were chosen as a feature subset. The most popular gene selection approaches are based on statistical tests or information theory to rank the genes [20]. Each gene is assessed individually and given a score reflecting its correlation with the class based on certain criterion in these gene ranking approaches. The gene is not dependent of any learning techniques in ranking gene approaches. Thus, these approaches have better generalization attribute and computational competence. But, there is an issue that these chosen genes are regularly highly correlated [21]. As these chosen genes may fit in to the same signaling pathways or function connected to the disease. Thus, if a gene has a high rank, other genes which are highly correlated with it, may have high rank in the gene ranking approach. This redundancy is an added computational border, which would result in misclassifications.

So, Li Jiangeng et al., [22] presented a new hybrid technique for choosing marker genes from gene expression data. This novel hybrid approach integrates gene ranking, heuristic clustering analysis and wrapper technique to choose marker genes for tumor classification.

In this approach, Li Jiangeng et al., initially the feature filter algorithm selects a group of top-ranked informative genes; then, in order to lessen the redundancy of the informative genes, a group of prototype genes are taken as the representative of the informative genes by heuristic Kmeans clustering; finally, SVM-FRE approach is utilized to choose a set of marker genes.

2.2 NMF-related Models for Tumor Clustering

A brief survey is presented on NMF related models which includes K-means, Probabilistic Latent Semantic Indexing, nsNMF. A novel model called Posterior Probabilistic Clustering is also presented in this section.

2.2.1. Nonnegative Matrix Factorization, NMF ([23, 24])

In general, NMF can be written as:

$$\begin{aligned} \min \quad & J(X, FG^T) \\ \text{s.t.} \quad & F, S, G \geq 0 \end{aligned} \tag{1}$$

$J(X, FG^T)$ represents certain distance function or dissimilarity function between two matrices X and FG^T. F and G are updated alternately until convergence. If the least square error $||X - FG^T||_F^2$ is chosen as objective function J to optimize, the equivalent update rules of F and G are:

$$\begin{aligned} F_{ia} &:= F_{ia} \frac{(XG)_{ia}}{(FG^TG)_{ia}} \\ G_{ia} &:= G_{ia} \frac{(X^TF)_{ia}}{(GF^TF)_{ia}} \end{aligned} \tag{2}$$

Otherwise, if the K-L divergence

$\sum_{i,j}(X_{ij}\, log\, \frac{X_{ij}}{(FG^T)_{ij}} - X_{ij} + (FG^T)_{ij})$ is chosen to optimize, the rules of F and G are:

$$F_{ia} := \frac{F_{ia}}{\sum_j G_{ja}} \sum_j \frac{X_{ij}}{(FG^T)_{ij}} G_{ja} \tag{3}$$

$$G_{ia} := G_{ia} \frac{(X^TF)_{ia}}{(GF^TF)_{ia}} G_{ja} \tag{4}$$

2.2.2 K-means

Ding et al., [25] describes that NMF which factorizes symmetric matrix X, which is the similarity matrix of the original samples, with orthogonal constraints on the factor matrices F (or G) is equivalent to K-means [26].

2.2.3 Probabilistic Latent Semantic Indexing, PLSI [27]

PLSI is one of models that are effectively utilized for the purpose of information retrieval. Ding et al., [28] described that PLSI and NMF optimize the same objective function (K-L divergence) with different update rules.

If X is normalized to satisfy $\sum_{ij} X_{ij} = 1$ the model can be written as:

$$min \sum_{i,j} (X_{ij}\, log \frac{X_{ij}}{(FSG^T)_{ij}} - X_{ij} + (FSG^T)_{ij}) \quad (5)$$

$$\text{s.t. } F, S, G \geq 0 \quad (6)$$

$$\sum_i F_{ik} = 1, \sum_j G_{jk} = 1, \sum_k S_{kk} = 1 \quad (7)$$

S is diagonal. The results reveal that the update rules of F and G in PLSI are certainly obtained from NMF simply by normalizing F and G in equation (3) and (4) at each iteration.

2.2.4 Nonsmooth NMF, nsNMF ([29])

nsNMF optimizes $X = FSG^T$ instead of $X = FG^T$, where $S = (1-\theta)I + \frac{\theta}{k} II^T$, I represents the identity matrix and parameter θ is used to control the sparseness of both F and G.

2.2.5 Posterior Probabilistic Clustering

Unlike PLSI, which considers the factor matrices F, S and G as class-conditional probabilistic matrices, i.e., F, S and G satisfy the condition (7), PPC ([30]) considers F, S and G as posterior probabilistic matrices, i.e., $\sum_k F_{ik} = 1, \sum_k G_{jk} = 1, \sum_k S_{kk} = 1$. In order to simply the model, only constraint G is added.

Unlike [30], in Zhong-Yuan Zhang et al., [26] selected K-L divergence as the objective function. The model can be written as:

$$(PPC) = \min_{F,G\geq 0} \sum_{i,j} (X_{ij} log \frac{X_{ij}}{(FG^T)_{ij}} - X_{ij} + (FG^T)_{ij}) \quad (8)$$

$$\text{s.t. } \sum_{k=1}^{K} G_{jk} = 1, j = 1,2, \ldots n$$

Mimic the derivative process of PLSI; the update rules of F and G are obtained.

$$G_{jk} = \frac{G_{jk}}{\sum_j F_{jk} \sum_k G_{jk}} \sum_i \frac{X_{ij} F_{ik}}{(FG^T)_{ij}} \quad (9)$$

$$= \frac{G_{jk} \left(\frac{X^T F}{GF^T}\right)_{jk}}{\sum_j F_{jk} \sum_k \left[\frac{G_{jk} \left(\frac{X^T F}{GF^T}\right)_{jk}}{\sum_i F_{ik}}\right]}$$

Update rule of F is the equivalent to standard NMF:

$$F_{ik} := \frac{F_{ik}}{\sum_j G_{jk}} \left(\frac{X}{FG^T} G\right)_{ik} \quad (10)$$

The results reveal the fact that this PPC model is an efficient one among the six models.

For further research, the author generalized PPC to simultaneous feature and sample clustering. F is used as the posterior probability for feature clustering, and the posterior probability normalization is $\sum_{k=1}^{K} F_{ik} = 1$.

The simultaneous PPC (SPPC) becomes $\min_{F,S,G\geq 0} J(X, FSG^T)$, s.t. $\sum_{k=1}^{K} F_{ik} = 1, \sum_{k=1}^{K} G_{jk} = 1$, where $J(X, FSG^T)$ can be conventional least squares error or K-L divergence, the corresponding algorithms can be derived similarly to PPC.

2.3 Feature Selection based on Evolutionary Algorithms

Feature selection is often regarded as an essential preprocess step to examine these data, as this approach can minimize the dimensionality of the datasets [31]. Two models of feature selection are available based on whether the selection is integrated with a learning approach. They are filter approach and wrapper approach.Wrapper approaches clearly have more benefits than filter approach based on conceptual perspective, as the features are chosen by optimizing the discriminate power of the ultimately used induction algorithm.

In this work, Enrique Alba et al., [32] are focussed in gene selection and classification of DNA Microarray data in order to differentiate tumor cells from normal cells. For this task, the author presented two hybrid techniques that utilize metaheuristics and classification approaches. The first approach comprises of a Particle Swarm Optimization (PSO) [33] integrated with a SVM technique. PSO is a population based metaheuristic motivated by the social nature of birds. Particularly, a recent version called Geometric PSO [34] has been used in this research. The second model is based on the widely used GA using a specialized Size-Oriented Common Feature Crossover Operator (SSOCF) [35], which keeps constructive informative blocks and constructs offsprings which have the same distribution than the parents. This model will be also integrated with SVM in this approach.

2.4 Machine Learning in Gene Selection

For gene selection, there are also a number of techniques available in the literature for tumor classification. In general, gene selection is considered as a variable selection issue in statistics and a dimension reduction issue in machine learning. Efficient gene selection regularly results in a significant classifier with effective accuracy and interpretability [36].

A number of greedy algorithms have been presented in the literature [37, 38]. In these approaches, gene-ranking techniques are widely used, which choose genes based on certain predetermined ranking criteria. There are two main kinds of ranking criteria, i.e., correlation coefficients and hypothesis testing statistics [39]. Two-sample t-test techniques consist of parametric tests [40, 41] and nonparametric tests [42]. A majority of the modern approaches are based on classical statistical approaches such as Bayesian variable selection [43], logistic regression [44], and Analysis Of Variance [45]. Moreover, certain recent techniques such as ICA [46] and SVM [47] have also been used. In recent years, neighborhood rough set based approach is also presented for gene selection [48, 49].

A comparative investigation of many discrimination approaches based on filtered sets of genes can be observed in Dudoit et al., [11]. Most of these approaches choosing essential genes based on individual gene information therefore are inefficient to consider mutual information among genes.

2.5 IVGA based Gene Selection

Independent Variable Group Analysis (IVGA) [50, 51] is an approach for grouping variables that are mutually dependent collectively so that not dependent or only weakly dependent variables are placed to different sets. The main aim of independent variable group analysis is to division a group of variables into separate sets so that the statistical dependencies of the variables within each group are strong [52]. These dependencies are modeled, in which the weaker dependencies between variables in different sets are unnoticed.

Chun-Hou Zheng et al., [53] presented a new method for gene selection based on IVGA. Additional to the feature selection approach presented in the literature [52], the author initially employed t-statistics approach to choose a segment of genes from the original data. Subsequently, the author chose the independent key genes through IVGA from the chosen genes for tumor classification. Ultimately, SVM is used to categorize tumors based on the key genes selected by IVGA. In order to validate the efficiency, the presented approach is applied to categorize three different DNA microarray data sets which include colon cancer data [7], acute leukemia data [39], and prostate cancer data [54]. The results reveal the fact that this approach is efficient and feasible.

2.6 Recent Advanced Techniques in Gene Selection

In cancer diagnosis and treatment, it is very essential to exactly recognize the location of origin of a tumor. With the rapid development in DNA microarray technologies, generating gene expression profiles for various cancer types has been a potential technique for cancer classification. Further to the study on binary classification such as normal versus tumor samples, which is effective in a number of domains, the discrimination of multiple tumor types has become an essential aspect. In the mean time, the selection of genes which are related to a particular cancer type enhances the efficiency of the classifiers and also offers molecular imminent for treatment and drug development. Rui Xu et al., [55] utilized semisupervised ellipsoid ARTMAP (ssEAM) for multiclass cancer discrimination and PSO for informative gene selection. ssEAM is a neural network approach based on adaptive resonance theory and appropriate for classification purposes. ssEAM features rapid, steady, and finite learning and produces hyperellipsoidal clusters, inducing complex nonlinear decision boundaries. PSO is an efficient technique for global optimization. A discrete binary version of PSO is utilized in this approach to show whether genes are chosen or not. The performance of ssEAM/PSO for multiclass cancer diagnosis is illustrated by evaluating it on three publicly available multiple-class cancer data sets. ssEAM/PSO attains significant performance on all these data sets.

Even though adopting feature reduction in standard rough set theory to choose informative genes is an efficient approach, its classification accuracy rate is generally not higher compared with other tumor-related gene selection and tumor classification techniques; for gene expression values must be discretized before gene reduction, which results in information loss in tumor classification. Thus, the neighborhood rough set model presented by Hu Qing-Hua is presented to tumor classification, which leaves out the discretization process, so no data loss occurs before gene reduction. Experiments on two popular tumor datasets indicates that gene selection by means of neighborhood rough set model clearly outperforms classic rough set theory and experiment results also show that majority of the chosen gene subset not only has higher accuracy rate but also are related to tumor [56].

Kai-Bo Duan et al., [57] proposed a new feature selection approach that utilizes a backward elimination process similar to that employed in Support Vector Machine Recursive Feature Elimination (SVM-RFE). Different from SVM-RFE approach, at each step, this approach evaluates the feature ranking score from a statistical analysis of weight vectors of multiple linear SVMs trained on subsamples of the original training data. This approach is validated on four gene expression datasets for cancer classification. The results reveal that this feature selection approach chooses better gene subsets than the original SVM-RFE and enhances the accuracy of classification. A Gene Ontology-based similarity assessment shows that the chosen subsets are functionally diverse, further testing this gene selection approach.

A novel Partial Least Squares (PLS) based gene-selection approach which synthesizes genetic relatedness and is appropriate for multicategory classification is presented by Guoli Ji et al., [58] for the discovery of tumor specific genes on microarray. By means of the explanation difference of independent variables on dependent variable (class), the author formulated three indicators for global gene selection, which considers all the combined impacts of all the genes and the correlation among the genes. Integrated with the linear Kernel Support Vector Classifier (SVC), this approach is validated by MIT acute myeloid leukemia/acute lymphoblastic leukemia (AML/ALL) and Small Round Blue Cell Tumors (SRBCT) data sets. A subset of specific genes with small numbers and high identification are obtained. The results reveal that this PLS-based approach for tumor-specific genes selection provides significant performance. Moreover, this approach is observed to be very robust. This approach can effectively solve feature-selection problem on high-dimensional small sample. Meanwhile, it has significant performance for multicategory classification.

The main issues in tissue classification by means of DNA Microarray data are choosing genes appropriate for a given tumor and generating the optimized classifiers. Shutao Li and Mingkui Tan [59] proposed a novel gene selection and tissue classification approach based on SVM and Genetic Algorithm (GA). Initially, the Wilcoxon-test is utilized as a coarse gene selection approach to eliminate most of the irrelevant genes.

Then the fine selection on the source of its classification potential of a single gene with SVM is conducted to obtain the final gene subset. Ultimately, GA is employed to optimize the parameters of SVM to identify the best parameters with the gene subset. The results reveal that this approach is more effective when compared with the previous approaches.

The application and combination of efficient and reliable approaches of computational intelligence provide a great potential for handling the feature selection and classification. Garcia et al.,[60] proposed a Differential Evolution (DE) technique for the effective automated gene subset selection. In this model, the selected subsets are validated through their classification rate using a SVM classifier. This technique is validated on DLBCL Lymphoma and Colon Tumor gene expression datasets. The results reveal that this DE-SVM model is highly reliable and significant when compared with other approaches.

An integration of Integer Coded Genetic Algorithm (ICGA) and Particle Swarm Optimization (PSO), coupled with the neural-network based Extreme Learning Machine (ELM), is employed for gene selection and cancer classification. ICGA is employed with PSO-ELM to choose an optimal set of genes, which is then utilized to construct a classifier to develop an algorithm (ICGA_PSO_ELM) that can deal with sparse data and sample imbalance. The author evaluated the significance of ICGA-PSO-ELM and compared with conventional techniques. A study into the functions of the selected genes, by means of a systems biology approach, indicated that a number of identified genes are involved in cell signaling and proliferation. An examination of these gene sets indicates a larger representation of genes that encode secreted proteins than seen in randomly selected gene sets. Rising biological confirmation has recognized the tumor microenvironment as a vital aspect that find out tumor survival and growth. Thus, the genes discovered by this study that encode secreted proteins might offer important insights to the nature of the essential biological features in the microenvironment of each tumor type that assist these cells to thrive and proliferate [61].

3. PROBLEMS AND DIRECTIONS

From the thorough analysis from the previous section, it is very clear that attaining genome-wide expression data from cancerous tissues provides an imminent into the gene expression variation of several kinds of tumor, which would intern help in the cancer classification of individual samples.

Since the number of immune-histochemical marker measurements increases, it is a general aspect to know whether tissue microarray data (protein abundances) could also be used for tumor class discovery. Class discovery in this context comprises of two challenges: (a) Presenting approaching to cluster tumors based on tissue microarray data and (b) Finding out whether reputed classes (clusters) build by such approaches are biologically and clinically significant.

Diagnostic pathology has conventionally depended on macro and microscopic histology and tumor morphology as the basis for classifying tumors. Existing classification techniques do not significantly discriminate among tumors with similar histopathologic features that vary in clinical course [63,64].

There are various challenges which are to be given prime importance in gene selection and tumor classification. One of the vital challenges of microarray investigations is to obtain biological insights from the extraordinary quantities of data on gene expression patterns. Partitioning genes into closely associated sets has become an element of practically all analyses of microarray data [15, 62]. The other most important challenge is the irresistible number of genes compared to the smaller number of available training samples. In machine learning techniques, these data sets have high dimension and small sample size [15]. Moreover, most of these genes are irrelevant to the distinction of samples. These irrelevant genes greatly affect the performance of the classifier. The other vital issue is that, DNA array data consists of technical and biological noise. Therefore, it is essential to recognize a subset of informative genes from a large data that will give greater performance.

Several new and advanced clustering techniques have to be used in tumor clustering for obtaining the best results. Advanced clustering techniques such as Nonnegative Matrix Factorization (NMF) [65], Penalized Matrix Decomposition (PMD) [66], Normalized Expectation-Maximization (EM) algorithm [67] can be used in tumor clustering and gene selection for improving the results.

4. CONCLUSION

Cancer has become one of the dangerous diseases in the recent scenario. As the number of cancer victims has been increasing day by day, there have increasing attention in the area of gene selection and tumor classification. A number of approaches have been developed by various researches for gene selection with tumor classification. This paper has provides various existing techniques available in the literature. The characteristic features, advantages and drawbacks of the existing gene selection approaches are also examined. The future directions for the better performance of the gene selection and tumor classificaiton approaches are also clearly discussed. This paper would provide an efficient platform for the researchers doing research in the field of medical sciences.

5. REFERENCES

[1] Tao Shi, David Seligson, Arie S Belldegrun, Aarno Palotie and Steve Horvath, “Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma”, Modern Pathology Vol. 18, pages: 547–557, 2005.

[2] Zhong-Yuan Zhang† Xiang-Sun Zhang, “Two Improvements of NMF Used for Tumor Clustering”, First International Symposium on Optimization and Systems Biology (OSB’07) Beijing, China, 2007.

[3] Xiong M, Li W, Zhao J, Jin L, Boerwinkle E., “Feature (gene) selection in gene expression-based tumor classification”, Molecular Genetics and Metabolism, Volume: 73, Issue: 3, Pages: 239-247, 2001.

[4] Tlsty TD, Margolin BH, Lum K. “Differences in the rates of gene amplification in nontumorigenic and tumorigenic cell lines as measured by Luria-Delbruck fluctuation analysis”, Proc Natl Acad Sci USA 86:9441–9445, 1989.

[5] Theillet C. “Full speed ahead for tumor screening”, Nature Med 4:767–768, 1998.

[6] Strausberg RL, Austin MJF. “Functional genomics: Technological challenges and opportunities”, Physiol Genomics 1:25–32, 1999.

[7] Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal

colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 96:6745–6750, 1999.

[8] Perou CM, Jeffrey SS, Rijn MVD, Rees CA, Eisen MB, Ross RT, Pergamenschikov A, Williams CF, et al. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Proc Natl Acad Sci USA 96:9212–9217, 1999.

[9] Isabelle Guyon, Jason Weston and Stephen Barnhill, "Gene Selection for Cancer Classification using Support Vector Machines", Machine Learning, 46, 389–422, 2002.

[10] Theodoridis S. Pattern Recognition. San Diego: Academic Press, 1999.

[11] S. Dudoit, J. Fridlyand, T. Speed, "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data", J. Am. Statistical Assoc., Vol. 97, pp. 77-87, 2002.

[12] C. Peterson, M. Ringner, "Analysis Tumor Gene Expression Profiles", Artificial Intelligence in Medicine, Vol. 28, no. 1, pp. 59-74, 2003.

[13] M. Eisen, P. Spellman, P. Brown, D. Botstein, "Cluster Analysis and Display of Genome- Wide Expression Patterns", Proc. Nat'l Acad. Sci. USA, Vol. 95, pp. 14863-14868, 1998.

[14] Pablo Tamayo, Donna Slonim, Jill Mesirov, Qing Zhu, Sutisak Kitareewan, Ethan Dmitrovsky, Eric S. Lander and Todd R. Golub, "Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods and Application to Hematopoietic Differentiation", Proc. Nat'l Acad. Sci. USA, Vol. 96, pp. 2907-2912, 1999.

[15] K. Yendrapalli, R. Basnet, S. Mukkamala, A. H. Sung, "Gene Selection for Tumor Classification Using Microarray Gene Expression Data", Proceedings of the World Congress on Engineering 2007 Vol I, WCE 2007, July 2 - 4, 2007.

[16] V. Cherkassy, "Model complexity control and statistical learning theory", Journal of natural computing 1: (2002) 109–133.

[17] N. Cristianini, J. S. Taylor, "Support Vector Machines and Other Kernel-based Learning Algorithms", Cambridge, UK: Cambridge University Press, 2000.

[18] Inza, I., Larranaga, P., Blanco, R. and Cerrolaza, A.J., "Filter versus wrapper gene approaches in DNA microarray domains", Artificial Intelligence in Medicine, ELSEVIER, Amsterdam, 2004, 31(2), pp.91- 103.

[19] C.H. Ooi and P. Tan, "Genetic algorithms applied to multi-class prediction for the analysis of gene expression data," Bioinformatics, Oxford University Press, Oxford, 2003, 19(1), pp. 37-44.

[20] Li, J., Zhang, C. and Olihara, M., "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression", Bioinformatics, Oxford University Press, Oxford, 2004, 20(15), pp. 2429-2437.

[21] Jaeger, J., Sengupta, R. and Ruzzo, W. L., "Improved gene selection for classification of microarrays", Pac. Symp. Biocomput, Hawaii, USA, 2003, pp. 53-64.

[22] Li Jiangeng, Duan Yanhua and Ruan Xiaogang, "A Novel Hybrid Approach to Selecting Marker Genes for Cancer Classification Using Gene Expression Data", IEEE, 2007.

[23] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755):788–791, October 1999.

[24] Daniel D. Lee and Sebastian H. Seung. Algorithms for non-negative matrix factorization. In Annual Conference on Neural Information Processing Systems, pages 556–562, 2000.

[25] C. Ding, X. He, and H. D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In SIAM Data Mining Conf, pages 606–610, 2005.

[26] Zhong-Yuan Zhang, "NMF-based Models for Tumor Clustering: A Systematic Comparison", Third International Symposium on Optimization and Systems Biology (OSB'09), pp. 41–47, 2009.

[27] Thomas Hofmann, "Probabilistic latent semantic indexing", In SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, pages 50–57. ACM Press, 1999.

[28] C. Ding, T. Li, and W. Peng. "Nonnegative matrix factorization and probabilistic latent semantic indexing: Equivalence chi-square statistic and a hybrid method", Proceedings of the National Conference on Artificial Intelligence, 21(1):342, 2006.

[29] A. Pascual-Montano, J. M. Carazo, K. Kochi, D. Lehmann, and R. D. Pascual-Marqui. Nonsmooth nonnegative matrix factorization (nsnmf). IEEE transactions on Pattern Analysis and Machine Intelligence, 28(3):403–415, March 2006.

[30] Chris Ding, Tao Li, Dijun Luo, andWei Peng. Posterior probabilistic clustering using nmf. In SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pages 831–832, New York, NY, USA, 2008. ACM.

[31] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," Machine Learning, vol. 46, no. 1-3, pp. 389–422, 2002. [Online]. Available: citeseer.ist.psu.edu/guyon02gene.html.

[32] Enrique Alba, Jose Garcia-Nieto, Laetitia Jourdan and El-Ghazali Talbi, "Gene Selection in Cancer Classification using PSO/SVM and GA/SVM Hybrid Algorithms", IEEE Congress on Evolutionary Computation (CEC), 2007.

[33] J. Kennedy and R. Eberhart, "Particle Swarm Optimization," in Proc. of the IEEE International Conference on Neural Networks, vol. 4, 1995, pp. 1942–1948.

[34] A. Moraglio, C. D. Chio, and R. Poli, "Geometric Particle Swarm Optimization," in 10th European conference on Genetic Programming (EuroGP 2007), ser. Lecture Notes in Computer Science, vol. 4445. Springer, Abril 2007.

[35] L. Jourdan, C. Dhaenens, and E.-G. Talbi, "A genetic algorithm for feature selection in data-mining for genetics," in Proceedings of the 4th Metaheuristics International ConferencePorto (MIC'2001), Porto, Portugal, 2001, pp. 29–34.

[36] Kitter J, "Feature selection and extraction", In: Young TY, Fu K-S (eds) Handbook of pattern recognition and image processing. Academic Press, NY

[37] Bae K, Mallick BK (2004), "Gene selection using a two-level hierarchical Bayesian model", Bioinformatics 20:3423–3430.

[38] Li W, Sun F, Grosse I (2004) Extreme value distribution based gene selection criteria for discriminant microarray data analysis using logistic regression. J Comput Biol 1:215–226.

[39] Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286:531–537.

[40] Devore J, Peck R (1997) Statistics: the exploration and analysis of data, 3rd edn. Duxbury Press, Pacific Grove, CA.

[41] Thomas G et al (2001) An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. Genome Res 11:1227–1236.

[42] Troyanskaya G et al (2002) "Nonparametric methods for identifying differentially expressed genes in microarray data", Bioinformatics 18:1454–1461.

[43] Lee KE, Sha N, Dougherty ER, Vannucci M, Mallick BK (2003) Gene selection: a Bayesian variable selection approach. Bioinformatics 19:90–97.

[44] Shevade SK, Keerthi S (2003) A simple and efficient algorithm for gene selection using sparse logistic regression. Bioinformatics 19:2246–2253.

[45] Draghici S, Kulaeva O, Hoff B, Petrov A, Shams S, Tainsky MA (2003) Sorin noise sample method: an ANOVA approach allowing robust selection of differentially regulated genes measured by DNA microarrays. Bioinformatics 19:1348–1359.

[46] Calo´ DG, Galibemberti G, Pillati M, Viroli C (2005) Variable selection in cell classification problems: a strategy based on independent component analysis. In: Vichi M, Monari P, Mignani S, Montanari A (eds) New development in classification and data analysis. Studies in classification, data analysis, and knowledge organization. Springer, Berlin, pp 21–30.

[47] Zhang HH, Ahn J, Lin X, Park C (2006) Gene selection using support vector machines with non-convex penalty. Bioinformatics 22:88–95.

[48] Hu QH, Yu DR, Liu JF, Wu CX (2008) Neighborhood rough set based heterogeneous feature subset selection. Info Sci 178(18): 3577–3594.

[49] Hu QH, Yu DR, Xie ZX (2008) Neighborhood classifiers. Expert Syst Appl 34(2):866–876.

[50] Lagus K, Alhoniemi E, Valpola H (2001) Independent variable group analysis. In: Dorffner G, Bischof H, Hornik K (eds) International conference on artificial neural networks—ICANN 2001, ser. LLNCS, vol 2130. Springer, Vienna, Austria. August, pp 203–210

[51] Lagus K, Alhoniemi E, Seppa¨ J, Honkela A, Wagner P (2005) Independent variable group analysis in learning compact representations for data. In: Honkela T, Ko¨no¨nen V, Po¨lla¨ M, Simula O (eds) Proceedings of the international and interdisciplinary conference on adaptive knowledge representation and reasoning (AKRR'05). Espoo, Finland, June, pp 49–56.

[52] Esa Alhoniemi, Antti Honkela, Krista Lagus, Jeremias Seppa, Paul Wagner, and Harri Valpola, "Compact Modeling of Data Using Independent Variable Group Analysis", IEEE Transactions on Neural Networks, 2007.

[53] Chun-Hou Zheng, Yan-Wen Chong and Hong-Qiang Wang, "Gene selection using independent variable group analysis for tumor classification", Neural Comput & Applic, 2011.

[54] Singh D, Febbo PG, Ross K, Jackson DG, Manola J, Ladd C, Tamayo P, Renshaw AA, D'Amico AV, Richie JP et al (2002) Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1:203–209.

[55] Rui Xu; Anagnostopoulos, G.C.; Wunsch, D.C.I.I., "Multiclass Cancer Classification Using Semisupervised Ellipsoid ARTMAP and Particle Swarm Optimization with Gene Expression Data", IEEE/ACM Transactions on Computational Biology and Bioinformatics, Volume:4, Issue: 1, Page(s): 65- 77, 2007.

[56] Wang, Shulin; Chen, Huowang; Li, Shutao, "Gene Selection Using Neighborhood Rough Set from Gene Expression Profiles", International Conference on Computational Intelligence and Security, Page(s): 959-963, 2007.

[57] Kai-Bo Duan; Rajapakse, J.C.; Haiying Wang; Azuaje, F., "Multiple SVM-RFE for gene selection in cancer classification with expression data", IEEE Transactions on Nano Bioscience, Volume: 4, Issue: 3, Page(s): 228 - 234, 2005.

[58] Guoli Ji; Zijiang Yang; Wenjie You, "PLS-Based Gene Selection and Identification of Tumor-Specific Genes", IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, Volume: 41, Issue: 6, Page(s): 830- 841, 2011.

[59] Shutao Li; Mingkui Tan, "Gene Selection and Tissue Classification Based on Support Vector Machine and Genetic Algorithm", 1st International Conference on Bioinformatics and Biomedical Engineering (ICBBE), 2007.

[60] Garcia-Nieto, J.; Alba, E.; Apolloni, J., "Hybrid DE-SVM Approach for Feature Selection: Application to Gene Expression Datasets", 2nd International Logistics and Industrial Informatics, (LINDI), Page(s): 1 – 6, 2009.

[61] Saraswathi, S.; Sundaram, S.; Sundararajan, N.; Zimmermann, M.; Nilsen-Hamilton, M., "ICGA-PSO-ELM Approach for Accurate Multiclass Cancer Classification Resulting in Reduced Gene Sets in Which Genes Encoding Secreted Proteins Are Highly Represented", IEEE/ACM Transactions on

Computational Biology and Bioinformatics, Volume:8, Issue: 2, Page(s): 452- 463, 2011.

[62] J. Quackenbush, "Computational Analysis of Microarray Data", Nature Rev. Genteics, Vol. 2, pp. 418-427, 2001.

[63] Srinivas Mukkamala, Qingzhong Liu, Rajeev Veeraghattam, Andrew H. Sung, "Computational Intelligent Techniques for Tumor Classification (Using Microarray Gene Expression Data)", International Journal of Lateral Computing, Vol. 2 (2005), pp. 38-45.

[64] Heping Zhang, Chang-Yung Yu, Burton Singer and Momiao Xiong, "Recursive partitioning for tumor classification with gene expression microarray data", proceedings of National Academy of Sciences of the United States of America, 2001.

[65] Chun-Hou Zheng; De-Shuang Huang; Lei Zhang; Xiang-Zhen Kong, "Tumor Clustering Using Nonnegative Matrix Factorization With Gene Selection", IEEE Transactions on Information Technology in Biomedicine, Volume: 13, Issue:4 , Page(s): 599- 607, 2009.

[66] Chun-Hou Zheng; Juan Wang; To-Yee Ng; Chi Keung Shiu, "Tumor Clustering Based on Penalized Matrix Decomposition", 4th International Conference on Bioinformatics and Biomedical Engineering (iCBBE), 2010 .

[67] Nguyen Minh Phuong; Nguyen Xuan Vinh, "Normalized EM algorithm for tumor clustering using gene expression data", 8th IEEE International Conference on BioInformatics and BioEngineering, Page(s): 1- 7, 2008

6. AUTHOR'S PROFILE

S. Praba received her U.G. degree B.Sc(CS) from Bharathiar University, Coimbatore in 1999 and her P.G. degree M.C.A from Bharathidasan University, Trichy in 2003. She was worked as Lecturer under Nanjappa Institute of Technology, Karaumathampatti during the year 2003 to 2005. Currently she is working as a Head and Assistant Professor under MCA Department in Maharaja Engineering College at Avinashi. Her current research interests includes Data mining.

A. K. Santra received the P. G. degree and Doctorate degree from I.I.T., Kharagpur in the year 1975 and 1981 respectively. He has got 20 years of Teaching Experience and 19 years of Industrial (Research) Experience. His area of interest includes Artificial Intelligence, Neural Networks, Process Modeling, Optimization and Control. He has got to his credit (i) 42 Technical Research Papers which are published in National / International Journals and Seminars of repute, (ii) 20 Research Projects have been completed in varied application areas, (iii) 2 Copy Rights for Software Development have been obtained in the area of Artificial Neural Networks (ANN) and (iv) he is the contributor of the book entitled **"Mathematics and its Applications in Industry and Business"**, Narosa Publishing House, **New Delhi.** He is the recognized Supervisor for guiding Ph. D. / M. S. (By Research) Scholars of Anna University-Chennai, Anna University-Coimbatore, Bharathiyar University, Coimbatore and Mother Teresa University, Kodaikanal. Currently he is guiding 12 Ph. D. Research Scholars in the Department. He is a Life member of CSI and a Life member of ISTE.