International Journal On Engineering Technology and Sciences – IJETS™
ISSN(P): 2349-3968, ISSN (O): 2349-3976
Volume I, Issue VI, OCT - 2014

A UNIFIED APPROACH FOR CANCER CLASSIFICATION USING STEM CELLS

R.Anbu selvi[1], Dr.R.Nallusamy[2],
PGScholar[1], Professor and Principal[2],
Nandha College Of Technology, Erode.
anbuarmy5891@gmail.com[1] , principal@nandhatech.org[2]

ABSTRACT: *Microarray data play an important role in the development of efficient cancer diagnoses and classification. However, micro array expression data are usually redundant and noisy, and only a subset of them present distinct profiles for different classes of samples. Thus, selecting high discriminative genes from gene expression data has become increasingly interesting in the field of bioinformatics. In this paper, a multi-objective biogeography based optimization method is proposed to select the small subset of informative relevant to the classification. A typical microarray gene expression dataset is usually both extremely sparse and imbalanced. To select multiple highly informative gene subsets for cancer classification and diagnosis, a hybrid algorithm has been proposed of statistical learning, Principle component analysis, PSO clustering, and granular computing separately eliminates irrelevant, redundant, or noisy genes in different granules at different stages and selects highly informative genes with potentially different biological functions in balance. To show the effectiveness of the proposed approach, then compare the performance of this technique with the signal-to-noise ratio (SNR) and consistency based feature selection (CBFS) methods. Using gene microarray datasets dataset from the adult stem cell (including both binary and multi-class classification problems), demonstrate experimentally that our proposed scheme can achieve significant empirical success and is biologically relevant for cancer diagnosis and drug discovery in terms of performance factors like precision, recall and fmeasure.*

I. INTRODUCTION

Recent advances in microarrays technology allow scientists to measure the expression levels of thousands of genes simultaneously in biological organisms and have made it possible to create databases of cancerous tissues. It finally produces gene expression data that contain useful information of genomic, diagnostic, and prognostic to select informative genes that contribute to a cancerous state. However, the gene selection process poses a major challenge because of the following characteristics of gene expression data: the huge number of genes compared to the small number of samples (high-dimensional data), irrelevant genes, and noisy data. To overcome this challenge, a gene selection method is used to select a subset of informative genes that maximizes classifier's ability to classify samples more accurately. In computational intelligence domain, gene selection is called feature selection. The gene selection has several advantages.

1) It can maintain or improve classification accuracy.
2) It can reduce dimensionality of the data.
3) It can reduce computational time.
4) It can remove irrelevant and noisy genes.
5) It can reduce the cost in a clinical setting.

In classification of gene expression data, selecting a smaller subset of informative genes from thousands of genes is a critical step for accurate cancer classification. In the context of cancer classification, gene selection methods can be classified into two categories. If a gene selection method is carried out independently from a classification procedure, it belongs to the filter method. Otherwise, it is said to follow a hybrid (wrapper) method. In the early era of microarrays analysis, most previous works have used the filter method to select genes since it is computationally more efficient than the hybrid method. Many filter methods are usually mentioned as individual gene-ranking methods such as *t*-test, signal-to-noise-ratio, information gain, etc. They evaluate a gene based on its discriminative power for the target classes without considering its correlations with other genes. This mechanism may result in inclusion of irrelevant and noisy genes in a gene subset for the cancer classification. The irrelevant and nosy genes reduce the classification accuracy. Meanwhile, these genes also increase the dimensionality of the gene subset and, in turn, rise their computational time. At the moment, several hybrid methods, especially a combination between particle swarm optimization (PSO) and a classifier, have been implemented to select

informative genes. The hybrid methods usually provide greater accuracy than the filter methods since the genes are selected by considering and optimizing correlations among genes.

Recently, several gene selection methods based on PSO have been proposed to select informative genes from gene expression data. PSO is a new population-based stochastic optimization technique. This approach produced 100% classification accuracy in many datasets, but it used a high number of selected genes (large gene subset) to achieve the high accuracy. It uses the high number because of the global best particle is reset to zero position when its fitness values do not change after three consecutive iterations.

A hybrid of PSO and PCA has best classification for the same purpose. Unfortunately, the accuracy result is still not high and many genes are selected for cancer classification since there are no direct probability relations between PCA and PSO. Generally, the PSO-based methods are intractable to efficiently produce a small (near-optimal) subset of informative genes for high classification accuracy. This is mainly because the total number of genes in gene expression data is too large (high-dimensional data). The diagnostic goal is to develop a medical procedure based on the least number of possible genes that needed to detect diseases.

II. RELATED WORK

DNA microarray data id used to screen thousand of genes simultaneously and determine whether those genes are active or silent in normal and cancerous tissues. With the advancement of microarray technology, analytical methods have used to find out whether microarray data have discriminative signatures of gene expression over normal or cancerous tissues. Solution has modeled with techniques prediction scheme that combines fuzzy preference based rough set (FPRS) method for feature (gene) selection with semi supervised SVMs.

It is difficult in profiling of the tumor tissues through gene selection. Microarray dataset is extremely sparse, so semi supervised classifier will be overhead with performance through increase in noise ratio and robustness. Signal to noise ratio will be increasing due to high dimensional data.

III. PROPOSED METHOD

The hybrid methods usually provide greater accuracy than the filter methods since the genes are selected by considering and optimizing correlations among genes. Recently, several gene selection methods based on PSO have been proposed to select informative genes from gene expression data. PSO is a new population-based stochastic optimization technique.

System now permits scientists to screen thousand of cells simultaneously and determine whether those Cells are active or silent in normal and cancerous tissues. With the advancement of gene cell technology, new analytical Methods must be developed to find out whether gene cell have discriminative signatures of cell expression over normal or cancerous tissues. In this paper, we propose a unified approach that combines fuzzy preference based rough set (FPRS) Method for feature (gene) selection with unsupervised SVMs. To show the effectiveness of the proposed approach, we compare the performance of this technique with the signal-to-noise ratio (SNR) and consistency based feature selection (CBFS) methods.

Advantages of the Proposed System

- Dimensionality reduction in the microarray cell dataset.
- Computational cost is low compared with microarray.
- DNA gene analysis improves generalization of malignant cells.
- Dominance relation and preference equivalence will be obtained easily

IV. METHODS

1. Data Pre-processing of DNA Microarray Data
2. Extracting the Feature using the fuzzy preference modeling
3. Deriving a Knowledge based on SVM classification features
4. Performance Evaluation
5. Deriving a hybrid technique for cancer classification using Principle component analysis and SVM classification
6. Performance Analysis

Data Pre-processing of DNA Microarray Data

DNA microarray data is pre-processed from Dataset downloaded from the repository, since classification is a typical and fundamental issue in diagnostic and prognostic prediction of cancer, different combination of methods is compared**.**

Data Pre-processing is carried as follows for both DNA Microarray

- Data cleaning: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
- Data integration: to combine multiple databases, data cubes, or files.
- Data transformation: normalization and aggregation.
- Data reduction: reducing the volume but producing the same or similar analytical results.
- Data discretization: part of data reduction, replacing numerical attributes with nominal ones.

Extracting the Feature using the fuzzy preference modeling

Fuzzy preference based rough set (FPRS) is used for finding more relevant gene markers from microarray gene expression data. Fuzzy preference relations can reflect the degree of preference quantitatively making it more powerful in extracting information from fuzzy data than equivalence or dominance relations. Features is obtained by the equivalence relation between the two set obtained the fuzzy preference extraction. In classification analysis for features, the concepts obtained by condition attributes are used to approximately describe the decision.

Fuzzy Preference relations

In practice, fuzzy rough set models address three key issues: i) inducing a granular structure on the universe based on an attribute or a criterion; ii) aggregating the granular structures obtained from different. attributes or criteria, and iii) determining the lower and upper approximations of decisions.

Two kinds of fuzzy preference relations are used in practice for a variety of decision-making models: i) multiplicative preference relation and ii) fuzzy preference relation. The decision values of the manuscripts are *accept*, *revise* and *reject*, the task is to analyze the consistency of the decision and to compute the dependency between each criterion and decision. It is also depends up to

1. Upwards conditional significance
2. Downwards conditional significance

Deriving Knowledge based SVM classification through features

SVM is learning machine based on two key elements: a general purpose learning algorithm and a problem specific kernel that computes the inner product of input data points in a feature space. SVM was originally developed as two class pattern reorganization problem which has been extended to the multi-class problem. To alleviate the problem of small-size training set, transductive support vector machine (TSVM) was proposed. TSVMs seek largest separation in presence of both labeled and unlabeled data through regularization. At the initial iteration, the standard SVMs are used to obtain an initial discriminating hyperplane based on the labelled data alone.

The *pseudo* labels are then assigned to the unlabeled samples. These are called semilabeled samples. Subsequently, transductive samples are selected from the semilabeled samples according to a given criterion. A hybrid training set is thus obtained consisting of the original labeled set and transductive set. The resulting hybrid training set is used at the following iterations to find a more reliable separating hyperplane. Training the TSVM algorithm can be roughly outlined as the following steps:

Input: An initial training set and an unlabeled set.
Output: Transductive SVM classifier with initial training set and a transductive set.
Step 1: Specify and execute an initial inductive learning using all labeled samples, and obtain an initial SVM classifier.
Step 2: Compute the decision function values of all the unlabeled samples with the initial classifier. Obtain label vector of all the unlabeled examples. Select all the positive and negative semilabeled points within the margin band as transductive samples and add them to the initial training set to obtain a hybrid training set.
Step 3: Retrain the SVM using this hybrid training set. Compute the decision function values

of all the unlabeled samples. Obtain the label vector of the unlabeled samples. Select all the positive and negative semilabeled points within the margin band as transductive samples.
Step 4: Select the resultant transductive samples by intersecting the previous and current transductive samples.
Step 5: Remove the transductive set from the initial training set and add the resultant transductive set obtained from step 4.
Step 6 Repeat steps 3–5. The algorithm finishes after a finite number of iterations.

The algorithm is capable of reducing the misclassification rate of the transductive samples at each iteration through a process of successive filtering between the transductive sets which results in increased accuracy.

The SVMs play the role to separate positive and negative samples, while the transductive inference successively searches more reliable discriminant functions employing additional unlabeled samples. Intuitively, unlabeled patterns guide the linear boundary away from the dense regions.

Designing and Applying Hybrid technique for PCA

Cancer classification is initially classified for features using SVM and results of the SVM is passed to Principle component analysis for the classification and clustering of the data based on dimensionality reduction, the possible maximum variance is estimated by the multivariate analysis.

V. PERFORMANCE ANALYSIS

Performance of the SVM with Fuzzy preference Modeling will yield good accuracy for the microarray datasets. Microarray dataset predicts the disease with less detection speed and high accuracy with precision, recall and fmeasure. The Proposed model yields better solutions to the cancer classification using the semi supervised technique with features of the cancer cells. Performance of the system is computed against the precision, recall and f measures against the classification technique SVM with fuzzy preferences, Hybrid technique using PCA and SVM yields the better results the classification and cluster formation of multivariate data found in the dataset.

VI. CONCLUSION

A typical microarray gene expression dataset is usually both extremely sparse and imbalanced. To select multiple highly informative gene subsets for cancer classification and diagnosis, a hybrid algorithm has been proposed of statistical learning, Principle component analysis, PSO classification, and granular computing separately eliminates irrelevant, redundant, or noisy genes in different granules at different stages and selects highly informative genes with potentially different biological functions in balance. To show the effectiveness of the proposed approach, we compare the performance of this technique with the signal-to-noise ratio (SNR) and consistency based feature selection (CBFS) methods. Using gene microarray datasets dataset from the adult stem cell the proposed scheme can achieve significant empirical success and is biologically relevant for cancer diagnosis and drug discovery in terms of performance factors.

REFERENCES

[1] S. Bandyopadhyay, A. Mukhopadhyay, and U. Maulik, "An improved algorithm for clustering gene expression data," *Bioinformatics*, vol. 23, no. 21, pp. 2859–2865, 2007.
[2] S. Bandyopadhyay, R. Mitra, and U.Maulik, "Development of the human cancer microRNA network," *BMC Silence*, vol. 1, no. 6, 2010.
[3] A. J. Gentles, S. K. Plevritis, R. Majeti, and A. A. Alizadeh, "Association of a leukemic stem cell gene expression signature with clinical outcomes in acute myeloidleukemia,"*JAMA—J. Amer.Med. Assoc.*, vol. 304, no. 24, pp. 2706–2715, 2010.
[4] H. K. Kim, I. J. Choi, C. G. Kim, A.Oshima, and J. E. Green, "Gene expression signatures to predict the response of gastric cancer to cisplatin and fluorouracil," *J. Clin. Oncol.*, vol. 27, no. 15s, 2009.
[5] U. Maulik, A. Mukhopadhyay, and S.Bandyopadhyay, "Combining pareto-optimal clusters using supervised learning for identifying co-expressed genes," *BMC Bioinformat.*, vol. 10,no. 27, 2009.
[6] U.Maulik, "Analysis of gene microarray data in soft computing framework," *Appl. Soft Comput.*, vol. 11, no. 6, pp. 4152–4160, 2011.
[7] A. Mukhopadhyay, S. Bandyopadhyay, and U. Maulik, "Multi-class clustering of cancer subtypes through SVM based ensemble of paretooptimal solutions for gene marker identification," *PLoS ONE*, vol. 5, no. 11, pp. 1–14, 2010.
[8] U. Maulik, A. Mukhopadhyay, and D. Chakraborty, "Gene-expression based cancer

subtypes prediction through feature selection and transductive SVM," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 4, pp. 1111–1117, 2013.

[9] A. Dupuy and R. M. Simon, "Critical review of public microarray studies in cancer outcome and guidelines on statistical analysis and reporting," *J. Natl. Cancer 1*, vol. 99, pp. 147–157, 2007.

[10] L. Ein-Dor, O. Zuk, and E.Domany, "Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer," *Proc. Natl. Acad. Sci. USA*, vol. 103, pp. 5923–5928, 2006.

[11] F. Chiclana, J. M. Tapia Garc´ia, M. J. del Moral, and E. Herrera-Viedma,"A Statistical comparative study of different similarity measures of consensusin group decision making," *Inform. Sci.*, vol. 221, pp. 110–123,Feb. 2013.

[12] D. Dubois and H. Prade, *Fuzzy Sets and Systems: Theory and Application.*New York, NY, USA: Academic Press, 1980.

[13] Y. C. Dong, G. Q. Zhang, W. C. Hong, and Y. F. Xu, "Consensusmodels for AHP group decision making under row geometric meanprioritization method," *Decision Support Syst.*, vol. 49, pp. 281–289,Mar. 2010.

[14] Y. C. Dong, W. C. Hong, Y. F. Xu, and S. Yu, "Selecting the individualnumerical scale and prioritization method in the analytic hierarchyprocess: A 2-tuple fuzzy linguistic approach," *IEEE Trans. Fuzzy Syst.*,vol. 19, no. 1, pp. 13–25, Feb. 2011.

[15] Z. P. Fan and Y. Liu, "An approach to solve group-decision-makingproblems with ordinal interval numbers," *IEEE Trans. Syst., Man,Cybern. B, Cybern.*, vol. 40, no. 5, pp. 1413–1423, Oct. 2010.

[16] S. Genc, F. E. Boran, D. Akay, and Z. S. Xu, "Interval multiplicativetransitivity for consistency, missing values and priority weights of interval fuzzy preference relations," *Inform. Sci.*, vol. 180, pp. 4877–4891,Dec. 2010.

[17] E. Herrera-Viedma, F. Herrera, and F. Chiclana, "Some issues on consistency of fuzzy preference relations," *Eur. J. Oper. Res.*, vol. 154,no. 1, pp. 98–109, Apr. 2004.

[18] E. Herrera-Viedma, S. Alonso, F. Chiclana, and F. Herrera, " A consensus model for group decision making with incomplete fuzzy preference relations," *IEEE Trans. Fuzzy Syst.*, vol. 15, no. 5, pp. 863–877,Oct. 2007.

[19] E. Herrera-Viedma, F. Chiclana, F. Herrera, and S. Alonso, "Group decision-making model with incomplete fuzzy preference relations based on additive consistency," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 37, no. 1, pp. 176–189, Feb. 2007.

[20] Y. C. Hsueh and S. F. Su, "Learning error feedback design of directadaptive fuzzy control systems," *IEEE Trans. Fuzzy Syst.*, vol. 20, no. 3, pp. 536–545, Jun. 2012.

[21] C. L. Hwang and K. Yoon, *Multiple Attribute Decision Making: Methods and Applications, A State of the Art Survey.* Berlin, Germany: Springer- Verlag, 1981.

[22] F. Karray and C. W. de Silva, *Soft Computing and Intelligent Systems Design: Theory, Tools and Applications.* Reading, MA, USA: Addison- Wesley, 2004.

[23] S. H. Kim and B. S. Ahn, "Interactive group decision making procedure under incomplete information," *Eur. J. Oper. Res.*, vol. 116, pp. 498–507, Aug. 1999.

[24] S. H. Kim, S. H. Choi, and J. K. Kim, "An interactive procedure for multiple attribute group decision making with incomplete information: Range-based appraoch," *Eur. J. Oper. Res.*, vol. 118, no. 1, pp. 139–152,Oct. 1999.