International Journal of Information Processing, 6(3), 55-66, 2012
ISSN : 0973-8215
IK International Publishing House Pvt. Ltd., New Delhi, India

Rough Set and Statistical Method for Both Way Reduction of Microarray Cancer Dataset

Asit Kumar Das[a] and Soumen Kumar Pati[b]

[a] Department of Computer Science and Technology, Bengal Engineering and Science University, Shibpur, Howrah - 711 103, West Bengal, India, Contact: asitdas72@rediffmail.com

[b]Department of Computer Science/Information Technology, St. Thomas' College of Engineering and Technology, 4, D.H. Road, Kolkata-700023, West Bengal, India,
Contact: soumen_pati@rediffmail.com

Microarray gene dataset often contains huge number of genes and samples many of which are irrelevant and redundant with respect to classification. Therefore, the data should be pre-processed to filter out the unimportant genes and samples before passing them on to the classifier. In the paper, the concepts of Rough Set Theory (RST) and Genetic Algorithm (GA) are used for selecting only the relevant samples of the dataset. The method constructs relative discernibility matrix to compute the core attributes based on which attributes are encoded to strings used as an initial population for running the genetic algorithm. The method runs each time by adding a single attribute to the initial strings to select only a minimal attribute set known as reduct. Then statistical method uses to reduce the gene set by selecting only the informative genes. Here, genes are ranked first and select only the high ranked genes. Then Pearson Correlation Coefficients are calculated and genes are merged. Thus genes are partitioned and final gene set is obtained by selecting a gene with the highest rank from each partition. The experimental results show that, the proposed method yields better result than some well known attribute reduction algorithms. Also the goodness of the method is evaluated by computing the classification accuracy by various well known classifiers on some real world microarray cancerous datasets.

Keywords : Discernibility Matrix, Genetic Algorithm, Gene Reduction, Rough Set Theory, Sample Reduction, Statistical Method.

1. INTRODUCTION

Now-a-days, an increasing number of applications in different fields, especially in the field of natural and social sciences, produce massive volumes of very high dimensional data under a variety of experimental conditions. In scientific databases like gene microarray dataset [1], it is common to encounter large sets of observations (samples), represented by hundreds or even thousands of coordinates (genes). The performance of data analysis such as clustering and classification degrades in such high dimensional spaces. Gene microarray high dimensional data provides the opportunity to measure the expression level of thousands of genes simultaneously and this kind of high throughput data has a wide application in Bioinformatics research. In DNA microarray data analysis generally biologists measure the expression levels of genes in the tissue samples from patients, and find explanations about how the genes of patients relate to the types of cancers they had. Many genes could strongly be correlated to a particular type of cancer, however, biologists prefer to focal point on a small subset of genes that dominates the outcomes before performing in-depth analysis and expensive experiments with a high dimensional dataset [1]. Therefore, automated selection of the minimal set of samples and genes(*i.e.*, reduct), is highly advantageous. DNA microarray technology [2] has directed the focus of computational biology towards

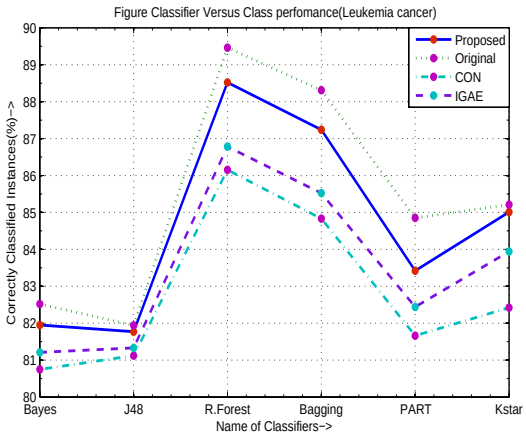

Figure 3. Classification Performance of Leukemia Dataset

Table 2
Comparison of Proposed, PCA and SVD Methods

Dataset	Method	EM(L)	MDBC(L)	K-means(S)
Prostate cancer	PRP	**-61.74**	**-73.31**	**835.98**
	PCA	-102.78	-83.82	917.70
	SVD	-89.59	-79.99	878.97
Lung cancer	PRP	**-135**	**-136.98**	**464.32**
	PCA	-152.44	-149.46	489.86
	SVD	-141.52	-146.71	481.05
Leukemia (ALL/AML)	PRP	**-208.09**	**-211.40**	**981.61**
	PCA	-231.03	-228.13	1119.47
	SVD	-225.49	-224.63	1013.51

4. CONCLUSIONS

This paper describes a new method of attribute reduction. In this method multiple reducts are generated using concepts of Rough Set Theory and Genetic Algorithm. Here, GA has been iteratively applied to search reducts in a microarray dataset and also ensure minimal length of the reducts. The GA applied was restrained by controlling the number of attributes on which it works, in each iteration. This could have been done by incorporating the minimum length criterion into the fitness function. But the disadvantage is that, one criterion may dominate the other and hence lead to inaccurate fitness value. Hence, GA was basically used to find reducts of particular length. The length was incremented by one in each iteration, starting from minimum. Future enhancements to this work may include use of boundary region of RST for designing a better fitness function. A better encoding technique may also be formulated. Also, application of optimization techniques other than GA, like PSO, Ant-colony optimization, *etc.*, is also worth a try. Moreover, the method gives multiple reducts. Some technique may be used to combine these reducts to a single reduct.
In this paper, a novel statistical method has been proposed to remove redundant genes. The genes are ranked based on discrete values and select only the high ranked genes. Then Pearson Correlation Coefficients are calculated and genes are merged based on similarity measurement and form some clusters. Then representative gene of each cluster is obtained by selecting a gene with the highest rank. This method of gene reduction is applied on microarray cancer dataset to select a subset of important informative genes. Future enhancements to this work may include integration of the gene ontology to better biological knowledge of genes.

REFERENCES

1. Lee, S hyun. and Kim Mi Na. Microarray Data, *ABC Transactions on ECE*, 10(5):120-

122, 2008.

2. D A Aerman, K Gish, S Ybarra, D Mack and A J Levine. Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays, *Proc. Natl. Acad. Sci*, 1:6745-6750, 1999.
3. J DeRisi. Use of a cDNA Microarray to Analyse Gene Expression Patterns in Human Cancer, *Nat. Genet.*, 14(4):457-60, 1996.
4. K Muralidhar and R Sarathy. Security of Random Data Perturbation Methods, *ACM Transactions on Database Syst.*, 24(4):487-493, 1999.
5. M Garey and D Johnson. Computers and Intractability- A Guide to the Theory of NP-Completeness, *Freeman*, New York, 1979.
6. Z Pawlak. Rough Sets, *International Journal of Information and Computer Sciences*, 11:341-356,1982.
7. Z Pawlak. Rough Set Theory and its Applications to Data Analysis, *Cybernetics and systems,* pages 661-688, 1998.
8. S K Pal and A Skowron (Ed.). Rough Fuzzy Hybridization: A New Trend in Decision Making, *Berlin: Springer-Verlag*, pages 93-98, 1999.
9. N Zhong, J Dong and S Ohsuga. Using Rough Sets with Heuristics for Feature Selection, *J. Intelligence Information System*, pages 199-214, 2001.
10. D E Goldberg. Genetic Algorithms in Search, Optimization and Machine Learning, *Addison-Wesley*, pages 432, 1989.
11. D Beasley, D R Bull, R R Martin. An Overview of Genetic Algorithms : Part 2 Research Topics, *University Computing*, 15:170-181, 1993.
12. L Devroye, L Gyorfi and G Lugosi. A Probabilistic Theory of Pattern Recognition, *Springer-Verlag*, 1996.
13. S K Pal and S Mitra. Neuro-Fuzzy Pattern Recognition: Methods in Soft Computing, *New York: Willey*, 1999.
14. S C Gupta and V K Kapoor. Fundamental of Mathematical Statistics, *Published by: Sultan Chand and Sons*, A.S. Printing Press, India, 1994.
15. A Skowron and C Rauszer. The Discernibility Matrices and Functions in Information Systems, Dordrecht, Kluwer, 1992.
16. Baoquing Jiang, Meng Liang and Ling Mei. Attribute Reduction Algorithm Based on Discernibility Matrix of Skowron and Itemset Lattice, *Int. Conference AICI*, 2010.
17. Y Y Yao and Y Zhao. Discernibility Matrix Simplification for Constructing Attribute Reducts, *Information Sciences*, 179(5):867-882, 2009.
18. Gisele L Pappa, Alex A Freitas and Celso A Kaestner. Attribute Selection with a Multiobjective Genetic Algorithm, Lecture Notes in Artificial Intelligence 2507, *Springer-Verlag*, pages 280-290, 2002.
19. Hong Shi and Jin-Zong Fu. A Heuristic Genetic Algorithm for Attribute Reduction, *Fifth International Conference on Machine Learning and Cybernetics*, 2006.
20. Bing Xiang Liu and Feng Liu and Xiang Cheng. An Adaptive Genetic Algorithm Based on Rough Set Attribute Reduction, *3rd International Conference on Biomedical Engineering and Informatics*, 2010.
21. X Wang and O Gotoh. Microarray-Based Cancer Prediction Using Soft Computing Approach, *Cancer Informatics*, 7:123-139, 2009.
22. Prostate Cancer Training Dataset: *http:// www-genome.wi.mit.edu/ mpr/prostate.*
23. Lung Cancer Training Dataset: *www. chestsurg.org/microarray.htm.*
24. Leukemia Cancer Training Dataset: *http:// www-genome.wi.mit.edu/cgibin/cancer/ datasets.cgi.*
25. WEKA: Machine Learning Software, *http:// www.cs.waikato.ac.nz/ ml.*
26. M A Hall. Correlation-Based Feature Selection for Machine Learning, *Ph.D Thesis, Dept. of Computer Science, Univ. of Waikato*, Hamilton, New Zealand, 1998.
27. Liu and R Setiono. A Probabilistic Approach to Feature Selection: A Filter Solution, *Proceedings of 13'th International Conference in Machine Learning*, pages 319-327, 1996.
28. M C Mozer, M I Jordan and T Petsche. A principled Alternative to the Self-Organising Map in Advances in Neural Information Processing Systems, *MIT Press*, Cambridge, MA, 9, 1997.
29. M Petrou and P Bosdogianni. Image Processing: The Fundamentals-an example of SVD, *John Wiley*, pages 37-44, 2000.

Dr. Asit Kumar Das is an Assistant Professor of Computer Science and Technology at Bengal Engineering and Science University, Shibpur, Howrah. He has received B.Sc. Honours in Mathematics, B. Tech. and M.Tech degree in Computer Science and Engg. from Calcutta University. He obtained Ph.D (Engg.) degree from Bengal Engineering and Science University, Shibpur, Howrah, India. His research interests include Data Mining and Pattern Recognition, Text Categorization, Rough Set Theory, Bio-informatics *etc.*.

Mr. Soumen Kumar Pati is an Assistant Professor of Computer Science/Information Technology at St. Thomas' College of Engineering and Technology, Kolkata, India. He has received M.Tech degree in Computer Technology from Jadavpur Computer Technology from Jadavpur University. He is registered for PhD (Engg) degree at Bengal Engineering and Science University, Shibpur, Howrah, India. His research interests include Bio-informatics, Data Mining and Pattern Recognition, Rough Set Theory, *etc.*.