A Survey on Stable Feature Selection for Micro Array Data

G.Baskar
PhD Research Scholar
Department of Computer Science
Government Arts College (Autonomous)
Coimbatore, Tamil Nadu, INDIA

P.Ponmuthuramalingam, Ph.D.
Associate Professor& Head
Department of Computer Science
Government Arts College (Autonomous)
Coimbatore, Tamil Nadu, INDIA

ABSTRACT

Feature selection has recently attracted strong interest in knowledge discovery from high-dimensional data. Classification is a data mining (machine learning) technique used to predict group membership for data instances, microarray is the technology which allows researchers to gather information on various gene expression all at the same time and this techniques have been applied in many computer application. Gene selection for cancer classification is one of the most important topics in biomedical field. In this survey a common microarray classification techniques based on data mining methodology for perform both accuracy and stability measurement.

Keyword:

Data mining, Feature selection, stability, microarray, classification.

1. INTRODUCTION

Feature selection as the pre-processing step to machine learning and mining biological data helps to extract useful knowledge from massive datasets gathered in biology, past few years data mining techniques have been successfully applied.[2] Usually, biological data needs to be pre-processed before they can be used in a data mining algorithm. A typical feature selection process consists of four basic steps (shown in Fig. 1), namely, subset generation, subset evaluation, stopping criterion, and result validation. Subset generation is a search procedure that produces candidate feature subsets for evaluation based on a certain search strategy. Each candidate subset is evaluated and compared with the previous best one according to a certain evaluation criterion. If the new subset turns out to be better, it replaces the previous best subset. The process of subset generation and evaluation is repeated until a given stopping criterion is satisfied. Then, the selected best subset usually needs to be validated by prior knowledge or different tests via synthetic and real world data sets. Feature selection can be found in many areas of data mining such as classification, clustering, association rules, and regression

Micro array is a powerful technology for biological exploration which enables to simultaneously measure the level of activity of thousands of genes, the amount of mRNA for each gene in a given sample (or a pair of samples) is measured Many feature selection methods have been adopted for gene selection from microarray data, and have shown good classification performance of the selected genes However, a common problem with existing gene selection methods is that the selected genes by the same method often vary significantly with some variations of the samples in the same data set[4] . To make the matters worse, different methods or different parameter settings of the same method may also result in largely different subsets of genes for the same set of samples.

In this paper we survey stable gene feature selection method with micro array classification.

2. MICRO ARRAY GENE EXPRESSION

Gene expression profiling or microarray analysis has enabled the measurement of thousands of genes in a single RNA sample.[3] There are a variety of microarray platforms that have been developed to accomplish this and the basic idea for each is simple: a glass slide or membrane is spotted or "arrayed" with DNA fragments or nucleotides that represent specific gene coding regions. Purified RNA is then fluorescently- or radioactively labelled and hybridized to the slide/membrane. In some cases, hybridization is done simultaneously with reference RNA to facilitate comparison of data across multiple experiments. After thorough washing, the raw data is obtained by laser

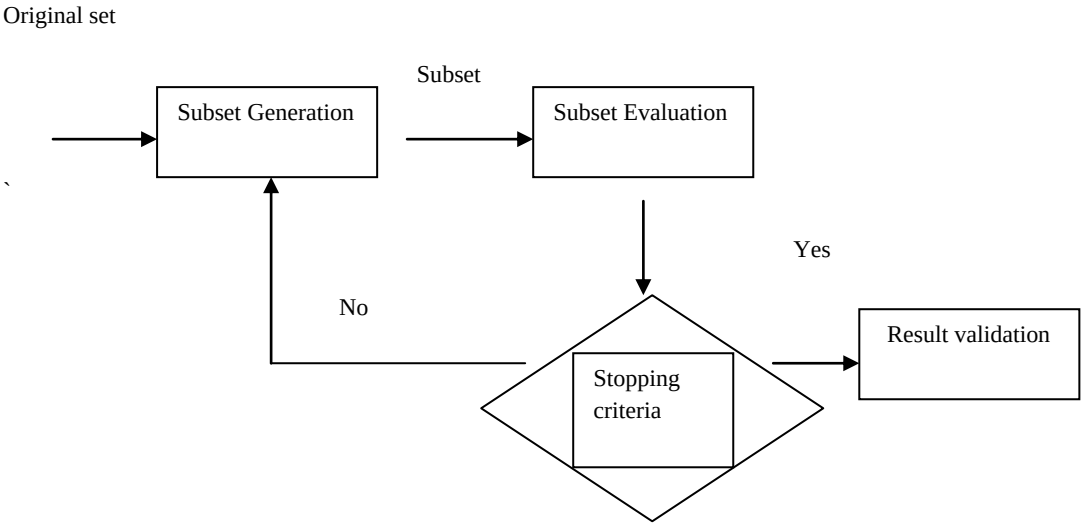

Figure1. Four steps of feature selection

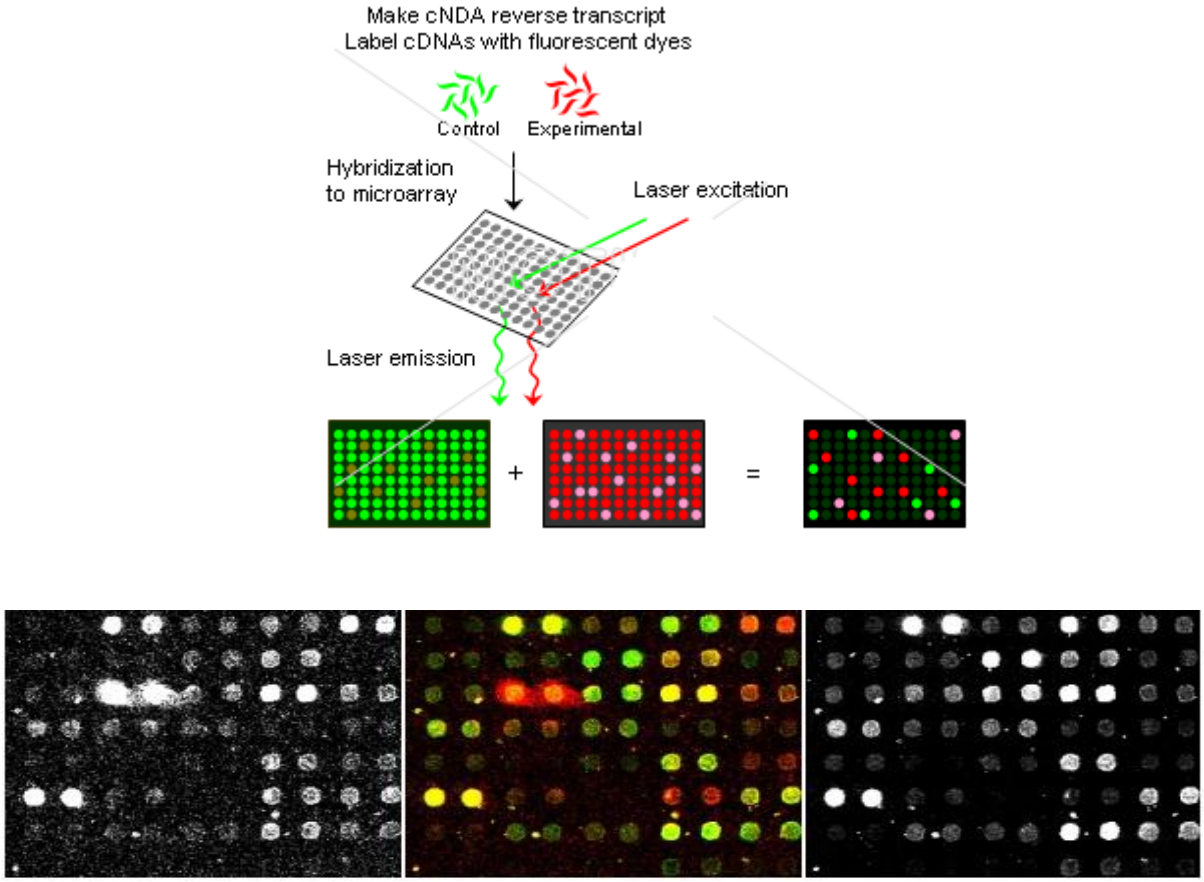

Figure 2 Micro array genes

scanning or auto radio graphic imaging . At this point, the data may then be entered into a database and analyzed by a number of statistical methods. The detected intensity distributions from a cDNA micro array for a region comprising round 80 probes.[14] The total number of probes on an array may range from a few dozens to tens of thousands. Left panel: grey-scale representation of the detected labels fluorescence at 635 nm (red), corresponding to mRNA sample A. Right panel: label fluorescence at 532 nm (green), corresponding to mRNA sample B. Spots that light up in only one of the two images correspond to genes that are only transcribed in one of the two samples. Middle panel: false-colour overlay image from the two intensity distributions. The spots are red, green, or yellow,

depending on whether the gene is transcribed only in sample A, sample B, or both.

Microarray type	Application
CGH	Tumour classification, risk assessment, and prognosis prediction
Expression Analysis	Drug development, drug response, and therapy development
Mutation/polymorphism analysis	Drug development, therapy development, and tracking diseases progression.

Table1.Microarray application

3. STABLE FEATURE SELECTION FOR MICRO ARRAY DATA:

3.1 stable and accurate feature selections

Stability is concern when the number of samples in a dataset is small and the dimensionality is high. The stable and performance measure of MRMR (minimum redundancy maximum relevance), two feature selection criteria are used by MRMR,MID(mutual information difference) and MIQ(mutual information quotient) the result is similar accuracy but it is MID more stable and they perform on eight dataset. Stability of MID and MIQ this compare the stability with different dataset. in general MID is more stable than MIQ the result is best stability and accuracy.(W.Buntine et al..Springer 2009 Binghamton University)

3.2 Various Reduction frame work for stable feature selection

A variance reduction approach for improve the stability of feature selection algorithm. The margin based instance weighting which weights training instances according to their influence to the estimation of feature relevance, more over the instance weighting algorithm is to be more effective than the recent ensemble feature selection method. Margin vector feature space, margin based instance weighting algorithm are used. Margin vector feature space is only considering one nearest neighbour form each class. To reduce noise or outliers we need multiple nearest neighbours for each class can be used to compute the margin vector. It verify that the instance weighting is an effective and efficient approach to reduce variance and improve the stability of feature selection algorithm.(2010 ieee international conference on datamining yuehan, lei yu)

3.3 Approach to active feature selection via selective sampling

a selective sampling approach to active feature selection in a filter model setting significant, time savings are observed using the raw performance of active feature selection and to investigate Relifs and other means characteristics for accuracy.(2004 Elsevier)

3.4 Stable feature selection via denser feature group

Here we point out the important stable feature selection in denser feature group based on kernel density estimation and treats features in each dense group as a coherent entity for feature selection DRAGS algorithm is developed under this based on microarray dataset. The feature groups which exhibit both high classification accuracy and stability, the various microarray dataset has verified that is stable for feature selection and DRAG algorithm which this leads to good classification, accuracy and stable.(2008 leiyu, chris ding)

3.5 Stable micro array via sample weighting

Sample weighting to improve the stability feature selection algorithm such as SVM-RFE which classification performance and the sample weighting to improve the stability of existing feature selection method for various sample weighting algorithm can also developed under this and the gene expression data based on the margin based sample weighting algorithm have more effective over stability.(2012 ieee/acm lei,yue and Michael)

4. CONCLUSION

This survey presents a collection of stable feature selection method under various performances of microarray. There is more feature selection algorithm available for classification but it is difficult to select one from data mining task. Therefore, on this survey the stability of feature selection method depend up on the specific choice of the similarity measure and the margin based sample weighting give better performance than other feature selection method.

5. FUTURE WORK

In future the stability of feature selection method for micro array data can be improved by modifying margin based sample weighting or margin vector feature space.

6. REFERENCES

[1] John, G.H., Kohavi, R., Pfleger, K.: Irrelevant Feature and The Subset Selection Problem. In: Proceedings of the Eleventh International Conference on Machine Learning, pp. 121–129 (1994)

[2] Liu, H., Yu, L.: Toward Integrating Feature Selection Algorithms for Classification and Clustering. IEEE Transactions on Knowledge and Data Engineering 17(4), 491– 502 (2005)

[3] Ding, C., Peng, H.: Minimum Redundancy Feature Selection from Microarray Gene Expression Data. In: Proceedings of the Computational Systems Bioinformatics conference (CSB 2003), pp. 523–529 (2003)

[4] Yu, L., Liu, H.: Efficient Feature Selection via Analysis of Relevance and Redundancy. Journal of Machine Learning Research 5, 1205–1224 (2004)

[5] Pepe, M.S., Etzioni, R., Feng, Z., et al.: Phases of Biomarker Development for Early Detection of Cancer. J. Natl. Cancer Inst. 93, 1054–1060 (2001)

[6] T. Li, C. Zhang, and M. Ogihara, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression," *Bioinformatics*, vol. 20, pp. 2429–2437, 2004

[7] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactionson Knowledge and Data Engineering (TKDE)*, vol. 17, no. 4,pp. 491–502, 2005.

[8] M. Wasikowski and X. Chen, "Combating the small sample class imbalance problem using feature selection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22,no. 10, pp. 1388–1400, 2010.

[9] C. A. Davis, F. Gerick, V. Hintermair, et al., "Reliablegene signatures for microarray classification: assessment ofstability and performance," *Bioinformatics*, vol. 22, pp. 2356–2363, 2006.

[10] C. A. Davis, F. Gerick, V. Hintermair, et al., "Reliable gene signatures for microarray classification: assessment of stability and performance," *Bioinformatics*, vol. 22, pp. 2356–2363, 2006.

[11] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," Science, vol. 286, pp. 531-537, 1999.

[12] T. Li, C. Zhang, and M. Ogihara, "A Comparative Study of Feature Selection and Multiclass Classification Methods for Tissue Classification Based on Gene Expression," Bioinformatics, vol. 20, pp. 2429-2437, 2004.

[13] Y. Saeys, I. Inza, and P. Larranaga, "A Review of Feature Selection Techniques in Bioinformatics," Bioinformatics, vol. 23, no. 19, pp. 2507-2517, 2007.

[14] H. Liu, J. Li, and L. Wong, "A Comparative Study on Feature Selection and Classification Methods Using Gene Expression Profiles and Proteomic Patterns," Genome Informatics, vol. 13, pp. 51-60, 2002

[15] P.A. Mundra and J.C. Rajapakse, "SVM-RFE with MRMR Filter for Gene Selection," IEEE Trans. Nano Bioscience, vol. 9, no. 1, pp. 31- 37, Mar. 2010

[16] I.H. Witten and E. Frank, Data Mining - Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishers, 2005.

[17] B.Y. Rubinstein, Simulation and the Monte Carlo Method. John Wiley & Sons, 1981

[18] Y. Tang, Y.Q. Zhang, and Z. Huang, "Development Two-Stage SVM-RFE Gene Selection Strategy for Microarray Expression Data Analysis," IEEE/ACM Trans. Computational Biology and Bioinformatics, vol. 4, no. 3, pp. 365-381, July 2007.

[19] Pawan Lingras, Chad West. Interval set Clustering of Web users with Rough k-Means, submitted to the Journal of Intelligent Information System in 2002.

[20] Yeung K.Y, Haynor D.R, Ruzzo W.L. *Validating clustering for gene expression data. Bioinformatics.* 2001

AUTHORS PROFILE

[1] G.Baskar received his Master's degree in Information Technology in K.S.Rangasamy College of Technology, Tiruchengode, Tamil Nadu India in 2008 and M.Phil Degree in Computer Science from Bharathiar University, Coimbatore, Tamil Nadu, India in 2010, and He is currently working towards the PhD degree in Department of Computer Science, Government Arts College, Coimbatore, Tamil Nadu, INDIA in 2011. His area of interest includes Data mining, bioinformatics

[2] P.Ponmuthuramalingam received his Masters Degree in Computer Science from Alagappa University, Karaikudi in 1988 and the Ph.D. in Computer Science from Bharathiar University, Coimbatore. He is working as Associate Professor and Head in Department of Computer Science, Government Arts College (Autonomous), Coimbatore. His research interest includes Text mining, Semantic Web, Network Security and Parallel Algorithms.