A Clustering and Selection Method using Wavelet Power Spectrum

S. Prabakaran R. Sahu S. Verma

Abstract—**In this paper, we consider the data mining problem of gene selection based on expression data and propose a method based on wavelet power spectrum. Wavelet power spectrum of the data has been analyzed and on the basis of the observations, a parameter was defined and used to select the features. Our proposed method has been applied on three types of cancer datasets. The results are in comparable with those of earlier works though our method is very simple and easy to and can be used in conjunction with other classification methods. Also, our technique is robust to noise.**

Index Terms—**Wavelet power spectrum/ relative percentage variation/ feature selection**

I. Introduction

Modern technologies like microarrays facilitate the generation of expression levels of thousands of genes simultaneously which are useful in solving the biological problems like disease classification. and identifying the genetic networks. But, .many of them contributes nothing to the problem domain and simply adds up the computation burden and lessens the accuracy of the solution which is not desirable for complex problems like disease classification and preparing more effective therapeutic solution like individual drug design. So, it is important and necessary to mine this data to select only genes contributing to the problem domain and to filter irrelevant data. Feature selection is the problem of identifying such genes or features [1] with significant information content to improve the generalization performance and inference of classification models [2] by overcoming the 'curse of dimensionality' which causes the risk of "over fitting". One important problem with feature selection methods is that both problem relevance and biological relevance of the features selected may not be achieved completely. Also, most of these methods do not fit for the wide range of datasets. They coupled with a particular classification method and time consuming. The purpose of this paper is to propose a feature selection method based on wavelet power spectrum which is capable of addressing these issues.

II. Background

Systematizations and surveys on feature selection algorithms have been presented in a variety of review articles like Blum and Langley [4], Kohavi and John [3] and Guyon [5]. A number of variable (or gene) selection methods like the support vector machine method [5], the genetic algorithm [6], the perceptron method [7], Bayesian variable selection [8,9,10,11], and the voting technique [12], mutual information-based gene and feature selection method [13], entropy based feature selection [15] and many artificial intelligent techniques like hill climbing, best first search [16], simulated annealing [17], backward elimination [18], forward selection and their combinations have been proposed. Specific to filter approach, Kira and Rendell's Relief algorithm [19] which selects features based on a threshold of weights assigned to each feature is a good example but it was tested on small set of features. A notable work on high dimensional microarray data was done by Golub et al [12] on the same dataset .using correlation measures. Califano et al [20] also worked on a high dimensional dataset of 6817 genes using a supervised learning algorithm. All these works revealed the fact that the result was better when used selected features instead of the whole data set.

Most of the methods of feature selection are complex and consume more time to converge. Another problem in these methods is redundancy of genes. Also, many of them do not fit for all data types in addition that they require more samples. Further, a very few model independent approaches for feature selection are available since most of the methods of feature selection are coupled with classification. In this paper, we propose a method of feature selection based on wavelet power spectrum which is found fit for a wide range of data sets and also works with smaller number of samples. It can be used in conjunction with other classification methods. The algorithm is very simple and requires comparatively less time to be executed. Unlike most of the other methods, it is relatively a very simple algorithm. In this work, apart from the novel use of wavelet power spectrum, we use the concept of distinct genes to tackle redundancy of genes. We observed that the features

Manuscript received April 21, 2006. S. Prabakaran is with the ABV-Indian Institute of Information Technology and Management, Gwalior, India (e-mail: pra_spin @ iiitm.ac.in).

R. Sahu is with the ABV-Indian Institute of Information Technology and Management, Gwalior, India (e-mail:rsahu @ iiitm.ac.in).

S. Verma is with the ABV-Indian Institute of Information Technology and Management, Gwalior, India (e-mail: sverma @ iiitm.ac.in).

selected by our method can be used in conjunction with more classification algorithms.

III. The method

In our approach of gene selection, we use the wavelet transforms of genes and the global spectral average of wavelet power spectrum over genes to select the genes useful for classification. A wavelet transform is a lossless linear transformation of a signal or data into coefficients on a basis of wavelet functions [22]. The use of wavelet transforms provides economical and informative mathematical representations of many objects of interest [23]. Also, the accessibility of wavelets has been made easier through many easily available software packages. Wavelet analysis is capable of providing analysis in a global fashion which is necessary in case of microarray data analysis. Surveys of wavelet applications in biological data and in data mining are presented at [24 -26] respectively. Mathematical details of wavelets may be referred at [27, 28, 29].

Local wavelet power spectrum is calculated by summing the squares of the coefficient values for each band. Global wavelet power spectrum [30] is the average of such local power spectra. The nature of genes in different diagnostic categories is different and in varying amount. So, it may be observed while analyzing the wavelet power spectrum that it may not be same in all diagnostic categories and based on this observation, a method to select important features relevant to each category against others can be devised. Hence, there is a possibility for class discovery and prediction by monitoring gene expression using wavelet power spectrum. It is obvious from Fig 1 that the power spectrum of gene2 is not the same in all the diagnostic categories of SRBCT data.

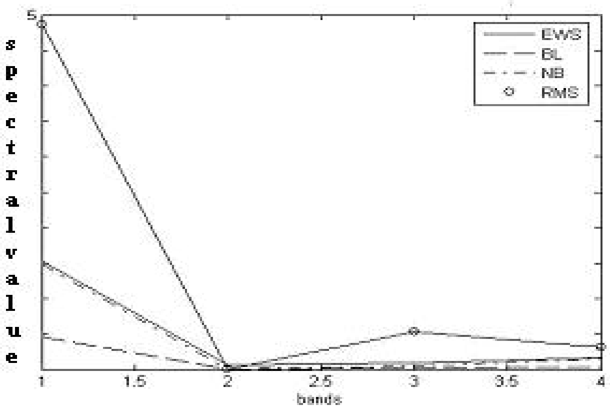

Fig 1. Wavelet power spectrum of gene 2 of SRBCT data set in different diagnostic categories. It is observed that gene 2 is dominant in EWS diagnostic category

Also, while analyzing the power spectrum of different genes, we observed that the spectrum of expression of a gene is dominant in one class against the whole group of other classes and this class is not the same for all genes (See Fig 2).

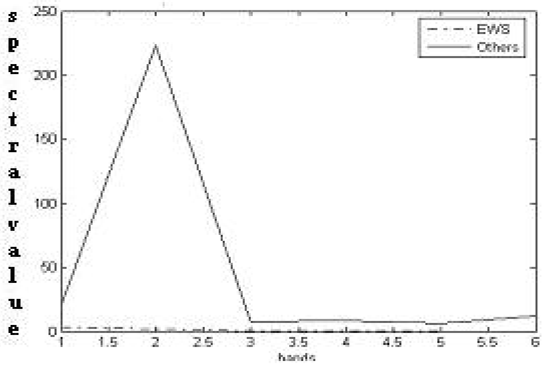

Fig 2. Wavelet power spectrum of gene 2 of SRBCT dataset in two sub clusters. It is observed that gene 2 is dominant in the cluster of samples of all other diagnostic categories than that in the cluster of EWS samples alone. This propert may be utilized for feature selection problem.

Fig 3 represents the power spectrum of another gene, Gene 1319 of the same dataset. A comparison of power spectrum of Gene2 and that of Gene 1319 of SRBCT dataset reflects the alternative trend present in genes .Gene 2 has a spectrum dominant in the group of data other than EWS family and Gene 1319 has a spectrum dominant in the EWS family

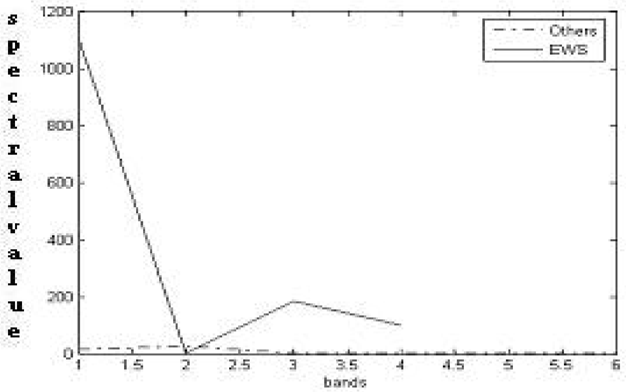

Fig 3. Wavelet power spectrum of gene1319 of SRBCT dataset. . It is observed that gene 2 is dominant in the cluster of of EWS samples alone than in the cluster of samples of all other diagnostic categories. This trend is just opposite to that observed in the case of gene2.

On the basis of the trends observed, we defined a parameter called relative percentage variation (RPV) to mine the data and to select the genes useful for distinguishing a diagnostic category from others. We calculated global average of the wavelet power spectra over genes in two subsets of a data, one containing single diagnostic category and the other containing remaining diagnostic categories.

The relative percentage variation (RPV) of the global average spectra of the genes against that of the other subset is calculated using the formula $RPV=\frac{(x_1-y_1)}{x_1}?\ 00\%$ where x_1 and y_1 are the global averages of genes in a particular diagnostic category and in the second subset. This clearly divides the data into two clusters. Cluster with genes with positive RPV were selected for the genes useful for classification. The genes selected for standard datasets were observed to be in tune with those reported in earlier works (12, 13, 14, 40) which used complex methods and most of these genes for classification. The same strategy can be applied to cluster the genes favourable for classification of any diagnostic category. Thus, the results obtained by our method are encouraging in both clustering genes and feature selection in the context of classification and hence found useful and may be examined further in solving other biological problems.

IV. Experimental results and discussion

A. *SRBCT dataset*

First, we focus on feature selection for the small, round blue cell tumors (SRBCT) of childhood. The dataset of SRBCT used for experimentation here is available at [31]. This dataset is composed of 2308 genes and 63 samples from four cancers which includes neuroblastoma (NB) (12 samples), rhabdomyosacoma (RMS)(23 samples), Burkitt lymphomas (BL) (8 samples) and Ewings family of tumors(EWS)(20 samples). Originally, Khan et al classified this dataset using artificial neural networks on gene expression profiles [31]. The feature selection and classification using this dataset has also been performed by Zhou et al using Gibb's sampler and SMC [14]. Our method was applied to select the genes important to classify EWS diagnostic category against others. It was also found reliable for feature selection of other diagnostic categories too.

Genes with index IDs 1319, 1645, 1954, 1831, 2303, 1980, 373 and 1626 were reported to be differentially expressed in EWS by our method (Table 1) as well as by Khan et al's work [31].

Rank	Index no.	Clone ID	RPV(%)
1	1319	866702	99.518
2	1645	52076	97.92
3	1954	814260	97.906
10	1831	208718	87.639
16	1980	841641	83.462
19	373	291756	81.311
20	1626	811000	81.219

Table1. Differentially expressed genes selected for classifying EWS diagnostic category of SRBCT data.

However, the order of these genes are little different. Gene 851 was not allocated to any class in [31]. Most of the other genes like gene 951 reported to be discriminating EWS have been selected as important genes for EWS but with some lower ranking.

Of all genes selected for EWS, neural-specific genes [32, 33] like TUBB5 (Gene 313), ANXA1 (Gene 1831), and NOE1 (Gene 1645) lend more credence to the proposed neural histogenesis of EWS [34].

B. *Acute leukemia data*

The experimental setup used for getting Acute leukemia data and other details can be found at [9]. The data set is publicly available at http://www-genome.wi.mit.edu/cgi-bin/cancer/publications. The microarray data consists of 7129 human genes and is split into a training set consisting of 38 samples (27 AML and 11 AML samples) and a test dataset of 34 samples (20 ALL and 14 AML samples). As a test case, important genes to classify AML versus ALL are selected on the basis of their relative percentage variations of expression levels between two classes and the results are displayed in Table 2.

Rank	Index No.	RPV %
1	5599	99.99822
5	1882	99.95379
11	5376	99.90413
12	6218	99.89492
14	6308	99.8568
17	2288	99.80596
18	2242	99.7741
19	2043	99.7637
20	6200	99.75038

Table 2. A list of genes selected useful for distinguishing AML and ALL samples.

Many genes reported in [9] are found in this list but in different order. Index number refers clone ID here. Among these selected genes, Genes with index numbers 2288, 1882,6200 and 2043 have been reported as important genes in discovering AML class in the original work at [9] Also, Genes with index numbers 5599, 2288, 5376 and 1882 have been reported to be important genes at [13] were genes were selected using mutual information.

Genes with index numbers 1882, 6218, 2288 and 6200 have been reported to be important genes selected using T-scores [13]. Gene 2242 has been reported as one of the important genes at [14]. Also, most of the other important genes reported to important are found to occupy almost the first 50 genes in this method. This clearly shows that this method of feature selection is worthy one and may be used in conjunction with different methods of classification.

C. *Breast cancer dataset*

Next, we examined our proposed method of feature selection on hereditary breast cancer data used in [29]. This dataset consists of twenty two breast tumor sample from 21 patients. Classification of each tumor sample into one of the classes based gene expression data was performed using a compound covariate predictor in [29]. In [14], the same classification was performed using SMC method and the genes were selected using a Gibb's sampler. The genes selected using our method to classify BRCA1 versus others is very close to those selected by Gibb's sampler in [14]. The genes with indices 10, 955, 2428, 2734, 585, 1288 and 1620 have been selected among top 20 genes by our method but with little difference in order (Table 3).

Rank	Index No.	RPV %
1	2272	97.25851
4	955	91.98577
8	1288	90.30327
15	585	88.15816
16	2248	88.11766
17	10	87.62514
18	1620	87.35974
19	2734	87.22749

Table 3. A list of genes selected useful for distinguishing BRCA1 vs. others

Some other genes presented in [14] are found within top 50 genes selected in our method. Among all these genes gene with index number 10 is reported as very important for all the methods in [8, 41]. It is observed in [14] that only with five or ten genes selected the classification was successful. This suggests that the genes selected by our method are worthwhile to use for classification of BRCA1 versus others since more of them are also found in the list mentioned in [14]. Gene 2272 has been identified as one of the top 20 strongest genes selected by mutual information [13]. Genes with index numbers 2734,2670,2893,1999 and 3009 which are also selected as the strongest genes in [13] are ranked between 26 and 45 by our method.

V. Conclusion

In this paper, we have treated the problem of feature selection of microarray gene expression data. We analyzed the wavelet power spectrum of genes and proposed a clustering and feature selection method useful for classification based on wavelet power spectrum. The top genes have been selected and listed. They have been compared with the results obtained in earlier works. The major advantages of this method of clustering and feature selection are multifold. The method is quite simple in comparison to other feature selection methods and for implementation it needs no special software since the accessibility of wavelets is made quite easier in already available software like matlab. Also, it can be used in conjunction with many established classification methods with lesser number of samples than that required for other methods. It is comparatively faster than other standard methods. The redundancy of genes which is a problem encountered in feature selection methods is counteracted by selecting distinct genes from the cluster of selected genes. The initial results of the idea of using wavelet power spectrum in feature selection using microarray data are encouraging and due to its simplicity, speed and effectiveness and fitness for a wide range of datasets, it may be further researched for applications in the area of genomic signal processing using microarrays.

References

[1] N.K. Kasabov, "*Evolving Connectionist Systems, Methods and Applications in Bioinformatics*", Brain Study and Intelligent Machines, Verlag Springer, 2002.

[2] F.L. Ramsey and D.W. Schafer, *The Statistical Sleuth, a course in methods of data analysis*, Duxbury Learning Publishing, 2002.

[3] R. Kohavi and G. John, "*Wrappers for feature selection*", Artificial Intelligence, Vol. 97(1-2), December 1997, pp. 273-324.

[4] Blum and P. Langley, "*Selection of relevant features and examples in machine learning*", Artificial Intelligence, Vol. 97(1-2), December 1997, pp.245-271.

[5] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, "*Gene selection for cancer classification using support vector machines*", Machine Learning, Vol. 46, 2002, pp. 389–422.

[6] L. Li, C.R. Weinberg, T.A. Darden and L.G. Pedersen, "*Gene selection for sample classification based on gene expression data, Study of sensitivity to choice of parameters of the GA/KNN method*", Bioinformatics Vol.17, 2001, pp. 1131–1142.

[7] S. Kim, E.R. Dougherty, J. Barrea, Y. Chen, M. Bittner and J.M. Trend, *"Strong feature sets from small samples*, J. Comput. Biol. Vol. 9, 2002, pp.127–146.

[8] K.E. Lee, N. Sha, E.R. Dougherty, M. Vannucc and B.K. Mallick, "*Gene selection, A Bayesian variable selection approach*", Bioinformatics, Vol. 19, 2003, pp.90–97.

[9] M. Smith and R. Kohn R, "*Nonparametric regression using Bayesian variable selection*", J. Econometrics, Vol. 75, 1997, pp.317–344.

[10] P. Yau, R. Kohn and S. Wood,, "*Bayesian variable selection and model averaging in high dimensional multinomial nonparametric regression*", J. Comput. Graph. Stat., Vol. 12(1), 2003, pp. 23-54.

[11] X. Zhou, X. Wang and E.R. Dougherty, "*Binarization of microarray data based on a mixture mode*", J. Mol. Cancer Therapy, Vol. 2, 2003,pp.679–684.

[12] 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, 1999, pp.531–537.

[13] X. Zhou, X. Wang and E.R Dougherty, "*Nonlinear probit gene selection and wavelet based feature selection*", Journal of Biological Systems, Vol. 12, No. 3, 2004, pp.371–386.

[14] X. Zhou, X. Wang and E.R Dougherty, "*A Bayesian approach to nonlinear probit gene selection and classification*", Journal of the Franklin Institute, Vol. 341, 2004, pp. 137-156.

[15] E. Xiang, M. Jordan and R. Karp, "*Feature selection for high dimensional genomic microarray data*", Proc. 8th Int. Conf. Machine Learning, Williams College, Massachusetts, 2001.

[16] R.Caruana. and D. Freitag, "*Greedy attribute selection*", International Conference on Machine Learning, 1994, pp. 28-36.

[17] J. Doak, *An evaluation of feature selection methods and their application to computer security,* Technical Report CSE-92-18. University of California at Davis, 1992.

[18] D.W. Aha, and R.L. Bankert, "*A comparative evaluation of sequential feature selection algorithms*", Artificial Intelligence and statistics, Springer-Verlag, New York, 1996.

[19] K. Kira and L. Rendell, "*A practical approach to feature selection*", Proceedings of the Ninth International Conference on Machine Learning, Morgan Kaufmann, 1992, pp.249-256.

[20] A. Califano, G. Stolovitzky and Y. Tu, "*Analysis of gene expression microarrays for phenotype classification*", Proceedings of the Annual Intelligent Systems in Molecular Biology, Vol.8, 2000, pp.75-85.

[21] D. Michie, "*Personal models of rationality*", Journal of Statistical Planning and Inference Vol. 21, 1990, pp. 381-399.

[22] S. Mallat, *A Wavelet Tour of Signal Processing*, Academic Press, San Diego, January 1998.

[23] F. Abramovich, T. Bailey and T. Sapatinas, "*Wavelet analysis and its statistical applications", JRSSD*, Vol. 48, 2000, pp.1–30.

[24] A. Aldroubi and M. Unser (editors), *Wavelets in Medicine and Biology*, CRC Press, Boca Raton, 1996.

[25] P. Lio, "*Wavelets in bioinformatics and computational biology, State of art and perspectives"*, Bioinformatics, Vol. 19, 2003, pp.2–9.

[26] T. Li et al, "*A survey on wavelet applications in data mining*", SIGKDD explorations, Vol. 4, Issue 2, 2002, pp. 49-68.

[27] G. Strang, "*Wavelets and dilation equations, A brief introduction", SIAM Review*, Vol.31 (4) , 1989, pp.614–627.

[28] I. Daubechies, *Ten Lectures on Wavelets*, Capital City Press, Montpelier, Vermont, 1992.

[29] C. K. Chui, *An Introduction to Wavelets*, Academic Press, Boston, 1992.

[30] T.A. Kestin, D. J. Karoly, J.I. Yano and N. A. Rayner, "*Time–frequency variability of ENSO and stochastic simulations*", J.Climate, Vol.11**,** 1998, pp**.** 2258 – 2272.

[31] J .Khan et al, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks", Nature medicine, volume 7, number 6, June 2001, pp. 673-679.

[32] V.L. Savchenko, J.A. McKanna,, I.R. Nikonenko and G.G. Skibo, "*Microglia and astrocytes in the adult rat brain, comparative immunocyto chemical analysis demonstrates the efficacy of lipocortin immuno reactivity"*, Neuroscience, Vol 96, 2000, pp.195–203.

[33] T. Nagano et al, "*Differentially expressed olfactomedin- related glycol proteins (Pancortins)", The brain*. Brain Res. Mol. Brain Res. Vol.53, 1998, pp.13–23.

[34] A.O. Cavazzana, J.S. Miser, J. Jefferson and T.J. Triche, "*Experimental evidence for a neural origin of Ewing's sarcoma of bone"*, Am. J. Pathol. Vol. 127, 1987, pp. 507–518.