International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 8, August 2017)

Percentage Change Method for Microarray Gene Expression Based Classification of Glioma

Supriya Patil[1], Gourish Naik[2], K. R. Pai[3], Rajendra Gad[4]
[1]*Research Scholar, Goa University, India*
[2]*Professor and Head, Electronics Dept., Goa University, India*
[3]*Professor and Head, Electronics and Telecommunication Dept., Goa, India*
[4]*Associate Professor, Electronics Dept., Goa University, India*

***Abstract*—With the invention of microarray technology there is considerable improvement in the survival rate of cancer patients. Microarray gene expression data has a very large dimension and small sample size. In this paper we have suggested number of methods to reduce the size of microarray data. In the first method, we propose to threshold and calculate the percentage change in gene expression values followed by feature extraction using discrete wavelet transform and classification using neural network. In second method we propose to calculate the mean and standard deviation of percentage change of every cancer sample and use them as two inputs to the neural network for classification. In the third method, we propose to perform the classification using two gene expression values having maximum percentage change.**

***Keywords*—Artificial Neural Network (ANN), Discrete Wavelet Transform (DWT), Principal Component Analysis (PCA), Resilient Back Propagation Algorithm (RPROP).**

I. INTRODUCTION

This Cancer sub-classification is one of the important research area in the biomedical field. The survival rate of the cancer patients can be enhanced by precise diagnosis of cancer sub-type using microarray technology. The microarray cancer gene expression data set contains a matrix of data, every column of which specifies one cancer sample and every row denotes the expression value of a specific gene of different samples.

Microarray cancer classification is usually carried out in two steps. First, the feature selection/feature extraction of microarray data and then classification of microarray data using different classifiers. Usually, by using feature selection or feature extraction, the size of the microarray data is reduced [1]. Feature extraction methods transform the signal into new domain using various methods like discrete cosine transform (DCT), Principal component analysis (PCA) [2], discrete wavelet transform (DWT) [3] etc.

The feature selection methods are basically classified as filter, wrapper, embedded, hybrid [4] and Ensemble method [5]. Feature selection methods chooses the most significant set of features for classification with filter methods [6], [7], wrapper method [8], [9], embedded method [10], [11], hybrid method [12], [13] and ensemble methods [14], [15].

The different classification algorithms used are, Support Vector Machines, K-means clustering, [9], [11], Naïve Bayes [16], Error Back Propagation algorithm [17] etc.

Each method proposed has its own merits and demerits. Even after development of all these techniques of reduction in the size of microarray data and classification Algorithms, there is opportunity to improve a method to get hundred percent accuracy classification.

In proposed work , we have developed a system for the classification of glioma grade III, grade IV datasets- GDS1975, GDS1976, downloaded from Gene Expression Omnibus Database [18], [19].

Shen, Qi [20] has employed Particle Swarm Optimization method for feature selection and SVM for classification. For GDS1976 dataset with 400 genes, the classification accuracy achieved is 92.67%.

Heba [21] has applied eight gene selection techniques like t-statistics, information gain etc. The different classifiers used are SVM, K-mean clustering, Random forest algorithms. The maximum classification accuracy obtained is 94.59%, 90.81% with GDS1975, GDS1976, data sets respectively with twenty genes and Random forest algorithm.

We propose to first use a percent change method, discrete wavelet transform for dimension reduction followed by classification using Resilient Back Propagation Algorithm.

Section II describes the feature extraction method, section III shows the details of the classification algorithm, implementation is explained in section IV, followed by result analysis in section V.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 8, August 2017)

II. Feature Extraction

Different methods like principal component analysis, discrete cosine transform, discrete wavelet transform etc., can be used to extract the features from normalized gene intensities. In the case of DWT, a time scale representation of the digital signal is obtained using digital filtering techniques. The DWT is computed by successive low pass and high pass filtering of the discrete time-domain signal. At each level the result of dyadic decimation of filtered data produces two sets of coefficients: approximation coefficients and detail coefficients.

The approximation coefficients sj(k) and detailed coefficients dj(t) are given as below:

$$sj(k) = \sum_m h(m-2k) * sj+1(m) \qquad (1)$$

$$dj(k) = \sum_m g(m-2k) * dj+1(m) \qquad (2)$$

Where,

h(n) =Impulse response of low pass filter
g(n) = Impulse response of high pass filter
j = level of decomposition
k = translation parameter.

Approximation coefficients represent the low frequency part of signal and detailed coefficients represent the high frequency part of signal. The wavelet coefficients basically represent the features of the signal and they can be used for classification purpose [24].

III. Classification

While dealing with the nonlinear data, artificial neural network based classifier offers significant advantage. For a particular application the multi-layer neural network can be trained using number of examples.

Resilient Back Propagation Algorithm (RPROP):

RPROP algorithm considers the sign of the gradient of the error instead of the magnitude of the gradient of the error. If the sign of slope of the error curve in two consecutive iterations remains same then the size of weight update is increased and vice-versa. The weight update is retained if slope of the error curve becomes zero. [23].

IV. Implementation

We propose to do the classification of glioma grade III, grade IV datasets- GDS1975, GDS1976. There are 26 samples of glioma grade III and 59 samples of glioma grade IV.

Most of the times, multiple copies of same gene are attached to the microarray chip. As it increases the dimension of microarray data, the duplicates are replaced by a single value that is average of all values of that individual gene in analogous sample. The process is repeated for every gene of a sample and for all samples. Hence in resultant samples there are no duplicate genes. Eliminating duplicates results in dimension reduction of data.

From every class the genes above certain threshold value are extracted. As the smaller values mostly indicate noise, they are neglected. For every gene in a class, if the intensity value of that gene exceeds the set threshold for one or more than one sample, that particular gene is considered for classification. Thersholding helps to further reduce the size of data.

Usually, the intensity values of many genes in a sample are very high and there is very small difference between the samples of both the class. This is why discriminating one class samples from other and removing hidden information becomes challenging.

Percentage change method helps to highlight the hidden information in the samples of both classes, In this method, the percentage change in expression value of every gene of each sample of one class with respect to average and root mean square expression value of same gene in other class is calculated and vice-versa. The process is repeated for all the samples of both the classes.

We have implemented the classification of normalized data in three different ways. In the first method, after normalization, the features are extracted using discrete wavelet transform Db4, Bior2.4, followed by the classification algorithm using Resilient back propagation algorithm.

In microarray data, expression values of large number of genes are small, which normally is due to cross hybridization and that of few genes are high. Also there is lot of similarity in samples of same class and within the samples of different classes. Each sample has the value of mean smaller than the value of standard deviation. This makes the signal unsuitable for statistical analysis.

In the second method, we have used percentage change method on the normalized data to make the samples suitable for statistical analysis. Then the average and standard deviation of every sample is computed and classification is performed using average and standard deviation of every sample as inputs.

International Journal of Emerging Technology and Advanced Engineering
Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 7, Issue 8, August 2017)

In the third method, the genes above a high value of threshold are extracted. The value of threshold selected is 1, 50,000 and 3, 00,000 for GDS1975 and GDS1976 datasets respectively. Selection of high threshold causes significant reduction in the size of microarray data, 5 genes for GDS1975 and 7 genes GDS1976 dataset. For genes selected above threshold, percentage change is calculated. The percentage change is directly given as input to the classifier. It gives 100% classification accuracy.

V. Result Analysis

The results for GDS1975 dataset with threshold= 1, 50,000, using wavelet coefficients as inputs are as shown in Table I.

TABLE I.
GDS1975 RESULTS WITH THRESHOLD =1, 50,000

Sr. No	Wavelet	Level	Algorithm	Accuracy
1.	Db4	1	RPROP	100%
2.	Bior2.4	1		

The results for GDS1976 dataset with threshold=3, 00,000, using wavelet coefficients as inputs are as shown in Table II.

TABLE II.
GDS1976 RESULTS WITH THRESHOLD =2, 00,000

Sr. No	Dataset	Algorithm	Accuracy
1.	GDS1975	RPROP	100%
2.	GDS1976		

The results for GDS1975 and GDS1976 dataset using average and standard deviation as inputs are as shown in Table III.

TABLE 3
GDS1975, GDS1976 RESULTS WITH AVERAGE AND STANDARD DEVIATION AS INPUT.

Sr. No	Wavelet	Level	Algorithm	Accuracy
1.	Db4	1	RPROP	100%
2.	Bior2.4	1		

The results for GDS1975 and GDS1976 dataset using two genes expression values with maximum percentage change as input are as shown in Table IV

TABLE IV.
GDS1975, GDS1976 RESULTS WITH PERCENT CHANGE AS INPUT

Sr. No	Dataset	No. of Genes	Algorithm	Accuracy
1.	GDS1975	2	RPROP	100%
2.	GDS1976	2		

VI. Conclusion

Percentage change method reveals the useful information in microarray data. We have compared the results obtained using wavelet transform coefficients, average and standard deviation, percentage change as inputs to the Resilient back propagation algorithm. Classifier gives 100% accuracy for all methods values of microarray samples. The best result is obtained with inputs which, contain only two gene expression values with maximum percentage change as input to classification algorithm as compared to Heba [20], Q.Shen [21].

REFERENCES

[1] Golub, Todd R., Donna K. Slonim, Pablo Tamayo, Christine Huard, Michelle Gaasen beek, Jill P. Mesirov, Hilary Coller et al. "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring."science 286, no. 5439 (1999): 531-537.

[2] Mahapatra, Rajat, Banshidhar Majhi, and Minakhi Rout. "Development and performance evaluation of improved classifiers of microarray data." In Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on, pp. 519-523. IEEE, 2012.

[3] Li, Shutao, Chen Liao, and James T. Kwok. "Wavelet-based feature extraction for microarray data classification." In Neural Networks, 2006. IJCNN'06. International Joint Conference on, pp. 5028-5033. IEEE, 2006.

[4] Y. Leung and Y. Hung, "A Multiple-Filter-Multiple-Wrapper Approach to Gene Selection and Microarray Data Classifica-tion," IEEEACM Trans Comput Biol Bioinforma., vol. 7, no. 1, pp. 108–117, Jan. 2010.

[5] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, R. Duque, H. Bersini, and A. Nowe, "A Survey on Filter Techniques for Feature Selection in Gene Ex-pression Microarray Analysis," IEEEACM Trans Comput Biol Bi-oinforma., vol. 9, no. 4, pp. 1106–1119, Jul. 2012.

[6] X. Li and M. Yin, "Multiobjective Binary Biogeography Based Optimization for Feature Selection Using Gene Expression Data," IEEE Trans. NanoBioscience, vol. 12, no. 4, pp. 343–353, Dec. 2013.

[7] B. Liao, Y. Jiang, W. Liang, W. Zhu, L. Cai, and Z. Cao, "Gene Selection Using Locality Sensitive Laplacian Score," IEEE/ACM Trans. Comput. Biol. Bioinform., vol. 11, no. 6, pp. 1146–1156, Nov. 2014.

[8] L. Yu, Y. Han, and M. E. Berens, "Stable Gene Selection from Microarray Data via Sample Weighting," IEEEACM Trans Comput Biol Bioinforma., vol. 9, no. 1, pp. 262–272, Jan. 2012.

[9] Q. Liu, Z. Zhao, Y. Li, X. Yu, and Y. Wang, "A Novel Method of Feature Selection based on SVM," J. Comput., vol. 8, no. 8, Aug. 2013.

[10] Y. Liang, C. Liu, X.-Z. Luan, K.-S. Leung, T.-M. Chan, Z.-B. Xu, and H. Zhang, "Sparse logistic regression with a L1/2 penalty for gene selection in cancer classification," BMC Bioinformatics, vol. 14, no. 1, p. 198, Jun. 2013.

[11] X. Cai, F. Nie, H. Huang, and C. Ding, "Multi-Class L2, 1-Norm Support Vector Machine," in 2011 IEEE 11th International Conference on Data Mining (ICDM), 2011, pp. 91–100.

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

[13] S. S. Shreem, S. Abdullah, M. Z. A. Nazri, and M. Alzaqebah, "Hybridizing ReliefF, mRMR filters and GA Wrapper Approaches for Gene Selection," J. Theor. Appl. Inf. Technol., vol. 46, no. 2, 2012.

[14] S. Zhang, H.-S. Wong, Y. Shen, and D. Xie, "A New Unsupervised Feature Ranking Method for Gene Expression Data Based on Consensus Affinity," IEEEACM Trans Comput Biol Bioinforma., vol. 9, no. 4, pp. 1257–1263, Jul. 2012.

[15] Z. Yu, H. Chen, J. You, H.-S. Wong, J. Liu, L. Li, and G. Han, "Double Selection Based Semi-Supervised Clustering Ensemble for Tumor Clustering from Gene Expression Profiles," IEEE/ACM Trans. Comput. Biol. Bioinform., vol. 11, no. 4, pp. 727–740, Jul. 2014

[16] X. Wang and O. Gotoh, "A Robust Gene Selection Method for Microarray-based Cancer Classification," Cancer Inform., vol. 9, pp. 15-30, Feb. 2010.

[17] S.-W. Chang, S. Abdul-Kareem, A. F. Merican, and R. B. Zain, "Oral cancer prognosis based on clinicopathologic and ge-nomic markers using a hybrid of feature selection and machine learning methods," BMC Bioinformatics, vol. 14, no. 1, p. 170, May 2013.

[18] http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS1976

[19] http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc=GDS1975

[20] Q. Shen, Z. Mei, and B.-X. Ye, "Simultaneous genes and training samples selection by modified particle swarm optimization for gene expression data classification," Comput. Biol. Med., vol. 39, no. 7, pp. 646–649, Jul. 2009.

[21] Abusamra, Heba. "A comparative study of feature selection and classification methods for gene expression data of glioma." Procedia Computer Science 23 (2013): 5-14.

[22] Q. Shen, Z. Mei, and B.-X. Ye, "Simultaneous genes and training samples selection by modified particle swarm optimization for gene expression data classification," Comput. Biol. Med., vol. 39, no. 7, pp. 646–649, Jul. 2009.

[23] Lat Soman, K. P. Insight into wavelets: from theory to practice. PHI Learning Pvt. Ltd., 2010.

[24] Riedmiller, Martin, and Heinrich Braun. "A direct adaptive method for faster backpropagation learning: The RPROP algorithm." Neural Networks, 1993, IEEE International Conference on. IEEE, 1993.