Int. J. Adv. Res. Sci. Technol. Volume 2, Issue1, 2013, pp 110-116.

IJARST

International Journal of Advanced Research in Science and Technology

journal homepage: www.ijarst.com

ISSN 2319 – 1783 (Print)

ISSN 2320 – 1126 (Online)

Mixed PCA and Wavelet Transform based Effective Feature Extraction for Efficient Tumor Classification using DNA Microarray Gene Expression Data

Jayakishan Meher*

Dept. of Computer Science and Engineering, Vikash College of Engineering for Women, Bargarh, Odisha, India

****Corresponding Author's Email: jk_meher@yahoo.co.in***

ARTICLEINFO

Article history:

Received 02 July. 2013

Accepted 16 Aug. 2013

Available online 25 Aug. 2013

Keywords:

Feature extraction,
gene expression,
tumor classification,
wavelet transform,
Principal component analysis,
neural network.

ABSTRACT

Cancer classification is an emerging research area in the field of bioinformatics. Gene expression profiles using microarray data play important role in accurate tumor diagnosis. Hence correlation between gene expression profiles to disease through microarray data and its analysis has been an intensive task in molecular biology. As the microarray data have thousands of genes and very few samples, it is crucial to develop techniques to effectively exploit the huge quantity of data produced. Thus efficient feature extraction and computational method development is indispensible for the analysis. In this paper a mixed feature extraction method by combining principal component analysis (PCA) and discrete wavelet transform (DWT) has been proposed to detect informative genes effectively. The PCA is a dimensionality reduction algorithm which aims to map high dimensional data to a lower dimensional space. The reduced data represents the most important variables underlying the original data. Further a feature extraction method based on the DWT is proposed. The approximation coefficients obtained by the decomposition at a particular level is used as the features for further study. Radial basis function neural network (RBFNN) classifier is used to efficiently predict the sample class which has a low complexity than other classifier. The potential of the proposed approach is evaluated through an exhaustive study by many benchmark datasets. The experimental results show that the proposed method can be a useful approach for cancer classification with low computational complexity and high accuracy.

Introduction:

Gene expression profiling or microarray analysis has enabled the measurement of thousands of genes in a single RNA sample [1]. This technique has been successfully exploited for classification and diagnostic of cancer. An important application of microarray data is to classify biological samples or predict clinical outcomes. Numerous learning algorithms and mining techniques are currently applied for identifying cancer using gene expression data. Microarray technology has been used as a basis to unravel the interrelationships among genes such as clustering of genes, temporal pattern of expressions, understanding the mechanism of disease at molecular level and defining of drug targets [2]. Among the above types diseases classification and analysis has gained a special interest. Especially tumor classification through the gene expression profiles has center of attraction in many research communities as it is important for subsequent diagnosis and treatment. Gene's expressions are stained at different conditions or different cellular stages to reveal the functions of genes as well as their regulatory interactions.

Gene expression of disease tissues may be used to gain a better understanding of many diseases, such as different types of cancers. Empirical microarray data produce large datasets having expression levels of thousands of genes with a very few numbers of samples which leads to a problem of curse of dimensionality. Due to this high dimension the accuracy of the classifier decreases as it attains the risk of overfitting. As the microarray data contains thousands of genes, hence a large number of genes are not informative for classification because they are either irrelevant or redundant. Hence to derive a subset of informative or discriminative genes from the entire gene set is necessary and a challenging task in microarray data analysis. The purpose of gene selection or dimension reduction is to simplify the classifier by retaining small set of relevant genes and to improve the accuracy of the classifier. For this purpose, researchers have applied a number of test statistics or discriminant criteria to find genes that are differentially expressed between the investigated classes. Various methods and techniques have been developed in recent past to perform the gene selection to reduce the dimensionality problem.

PCA and Linear Discriminant Analysis (LDA) also fall under a peculiar category of feature transformation where in the former uses a statistical signal criterion whereas the latter uses a classification model. The Partial Least Squares (PLS) method can also be categorized under the same roof of transformation and is compared to PCA where in the former uses a linear regression model whereas the latter stresses on the use of maximum variance calculated. The Locally Linear Embedding (LLE) technique is a manifold learning methodology and thus falls under the non-linear practices of dimensionality reduction [3]. Dimension Reduction of Microarray Data Based on Local Principal Component has been presented to improve the dimension [4]. Factor analysis and wavelet transform method has been used for tumor classification using gene expression data[5]. The filter method basically use a criterion relating to factors and select key genes for classification such as Pearson correlation coefficient method, t-statistics method [6], signal-to-noise ratio method [7], the partial least square method, independent component analysis [8], linear discriminant analysis and principal component analysis [9].All the methods transform the original gene space to another domain providing reduced uncorrelated discriminant components. These methods do not detect the localized features of microarray data. Hence Liu [10,11] proposed a wavelet basis function to perform the multi resolution analysis of the microarray data at different levels. The relevant genes of the microarray data can be measured by wavelet basis based on compactness and finite energy characteristic of the wavelet function. It does not depend on the training samples for the dimension reduction of the microarray data set. It also does not require a large matrix computation like the LDA, PCA and ICA, so simpler to implement.

The data set has been preprocessed by applying a feature selection algorithm in order to remove the noise and irrelevant features which affect the result of dimensionality reduction algorithm [12]. Improved direct LDA and its application to DNA microarray gene expression data has been discussed [13]. This enhanced DLDA method surpassed DLDA and certain other related techniques such as PCA and LDA technique and the OLDA technique. Several Machine learning and statistical techniques have been applied to classify the microarray data. Tan and Gilbert [14] used the three supervised learning methods such as C4.5 decision tree, bagged and boosted decision tree to predict the class label of the microarray data. Dettling [15] have proposed an ensemble method of bag boosting approach for the same purpose. Many authors have used successfully the support vector machine (SVM) for the classification of microarray data [16]. Khan et al. [17] used the neural networks to classify the subcategories of small round blue-cell tumors. Also O'Neill and song [18] used the neural networks to analyze the lymphoma data and showed very good accuracy. B Liu et al. [19] proposed an ensemble neural network with combination of different feature selection methods to classify the microarray data efficiently. Fisher's linear discriminant analysis (LDA) in combination with a genetic algorithm is used to study the spatial system of gene subsets [20].

As the gene expression data in microarray is highly rich and it is a challenging task to meet the requirement to extract the feature effectively and reduce the dimension of the microarray data. But the conventional neural networks require a lot of computation and consume more time to train. In this paper a mixed PCA and signal processing based wavelet transform have been proposed for effective feature extraction and introduced a new promising low complexity neural network known as radial basis function neural network (RBFNN) to efficiently classify the microarray data.

The rest of this paper is organized as follows: Section 2 presents the details of the dataset used for the study in the paper. Section 3 focuses on the proposed methods of feature extraction such as principal component analysis and wavelet transform and tumor classification using gene expression data using radial basis function neural network and Section 4 presents the Simulation of the simulation result and discussion of the proposed methods. Section 5 draws the conclusions of this paper.

Material Preparation:

In this section, the cancer gene expression data sets used for the study are described. These datasets are also summarized below.

A. SRBCT Dataset:

The dataset consists of four categories of small round blue cell tumors (SRBCT) with 83 samples from 2308 genes. The tumors are Burkitt lymphoma (BL), the Ewing family of tumors (EWS), neuroblastoma (NB) and rhabdomyosarcoma (RMS). There are 63 samples for training and 20 samples for testing. The training set consists of 8, 23, 12 and 20 samples of BL, EWS, NB and RMS respectively. The testing set consists of 3, 6, 6 and 5 samples of BL, EWS, NB and RMS respectively.

B. MLL Leukemia Dataset:

The dataset consists of three types of leukemias namely ALL, MLL and AML with 72 samples from 12582 genes. The training dataset consists of 57 samples (20 ALL, 17 MLL and 20 AML) and the test data set consists of 15 samples (4 ALL, 3 MLL and 8 AML).

C. Colon Dataset:

The dataset consists of 62 samples from 2000 genes. The training dataset consists of 42 samples where (30 class1, 12 class2) and the test data set consists of 20 samples (10 class1, 10 class2).

Proposed Methods:

A. Principal Component Analysis:

The problem of high dimensionality can be approached with the use of dimensionality reduction methods. Principal component analysis is commonly used methods for the analysis of high dimensional data. The PCA dimensionality reduction method [21,22] is a linear dimensionality reduction method. It works by projecting a number of correlated variables into a (smaller) number of uncorrelated variables called principal components. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. The algorithm solves for the eigenvalues and eigenvectors of a square symmetric matrix with sums of squares and cross products. The eigenvector associated with the largest eigenvalue has the same direction as the first principal component. The eigenvector associated with the second largest eigenvalue determines the direction of the second principal component. The sum of the eigenvalues equals the trace of the square matrix and the maximum number of eigenvectors equals the number of rows (or columns) of this matrix.

PCA is a well-known method of dimension reduction [23]. The basic idea of PCA is to reduce the dimensionality of a data set, while retaining as much as possible the variation present in the original predictor variables. This is achieved by transforming the p original variables $X= [x_1, x_2, \ldots, i_p]$ to a new set of K predictor variables, $T= [t_1, t_2, \ldots, t_K]$, which are linear combinations of the original variables. In mathematical terms, PCA sequentially maximizes the variance of a linear combination of the original predictor variables

$$u_k = \underset{u'u=1}{arg\ max}\ Var(Xu) \qquad (1)$$

subject to the constraint $u_i' S_x u_j = 0$, for all $1 <= i < j$ The orthogonal constraint ensures that the linear combinations are uncorrelated, i.e. $Cov(xu_i, xu_j)=0$, $i \neq j$

These linear combinations $t_i = Xui$ are known as the principal components (PCs) [24]. Geometrically, these linear combinations represent the selection of a new coordinate system obtained by rotating the original system. The new axes represent the directions with maximum variability and are ordered in terms of the amount of variation of the original data they account for. The first PC accounts for as much of the variability as possible, and each succeeding component accounts for as much of the remaining variability as possible. Computation of the principal components reduces to the solution of an eigenvalue-eigenvector problem. The projection vectors (or called the weighting vectors) u can be obtained by eigenvalue decomposition on the covariance matrix S_x

$$S_x u_i = \lambda_i u_i \qquad (2)$$

Where λ_i is the i-th eigenvalue in the descending order for $i=1,\ldots,K$, and u_i is the corresponding eigenvector. The eigenvalue λ_i measures the variance of the i-th PC and the eigenvector u_i provides the weights (loadings) for the linear transformation [25, 26]. The maximum number of components K is determined by the number of nonzero eigenvalues, which is the rank of S_x, and $K \leq min(n,p)$. The computational cost of PCA, determined by the number of original predictor variables p and the number of samples n, is in the order of $min(np^2+p^3 , pn^2+n^3)$. In other words the cost is $O(pn^2+n^3)$ when $p>n$.

B. Wavelet transform based feature extraction:

For wavelet analysis for gene expression data, a gene expression vector can be represented as a sum of wavelets at different time shifts and scales using discrete wavelet analysis (DWT). The DWT is capable of extracting the local features by separating the components of gene expression vector in both time and scale. Wavelet transform proposed by Grossman and Morlet [27] is an efficient time-frequency

representation method which transforms a signal in time domain to a time-frequency domain. The basic idea is that any signal can be decomposed into a series of dilations and compressions of a mother wavelet ($\psi(t)$). Hence the continuous wavelet transform of a signal is defined as:

$$CWT(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t) \psi \left(\frac{t-b}{a} \right) dt \qquad (3)$$

where $\psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi \left(\frac{t-b}{a} \right), a \in R^{+}, b \in R$

The resolution of the signal depends on the scaling parameter 'a' and the translation parameter 'b' determines the localization of the wavelet in time. The CWT can be realized in discrete form through the discrete wavelet transform. The DWT is capable of extracting the local features by separating the components of the signal in both time and scale. In the microarray data the gene expression profile is considered as a signal which can be represented as a sum of wavelets at different time shifts and scales using the discrete wavelet transform (DWT).

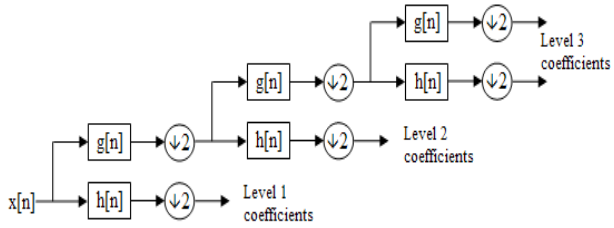

Fig: 1. Wavelet decomposition

The wavelets can be realized by iteration of filters with rescaling which was developed by Mallat [28] through wavelet filter banks. The resolution of the signal, which is a measure of the amount of detail information in the signal, is determined by the filtering operations, and the scale is determined by up sampling and down sampling operations. The approximation coefficients obtained by the decomposition at a particular level is used as the features for further study.

C. Classification using Radial Basis Function Neural Network:

Machine learning method is used for classification of tumor data. In this experiment training dataset is used to build the classifier and test dataset to evaluate the performance of proposed method based on datasets. For function approximation and pattern classification problems we are using the radial basis function network (RBFNN) which is a neural structure because of their simple topological structure and their ability to learn in an explicit manner. In the classical RBF network, there is an input layer, a hidden layer consisting of nonlinear node function, an output layer and a set of weights to connect the hidden layer and output layer. Due to its simple structure it reduces the computational task as compared to conventional multi layer perception (MLP) network. In RBFNN, the basis functions are usually chosen as Gaussian and the number of hidden units are fixed apriori using some properties of input data [29,30]. The structure of a RBF network is shown in Fig. 2.

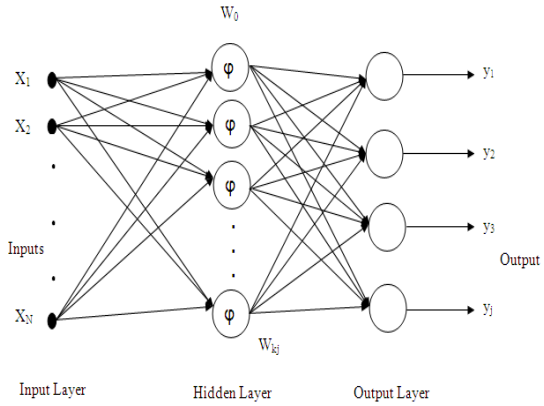

Fig: 2. The RBFNN based classifier

For an input feature vector x, the output of the jth output node is given as.

$$y_j = \sum_{k=1}^{N} w_{kj} \varphi_k = \sum_{k=1}^{N} w_{kj} e^{-\frac{\|x(n) - C_k\|}{2\sigma_k^2}} \qquad (4)$$

The error occurs in the learning process is reduced by updating the three parameters, the positions of centers (C_k), the width of the Gaussian function (σ_k) and the connecting weights (w) of RBFNN by a stochastic gradient approach as defined below:

$$w(n+1) = w(n) - \mu_w \frac{\partial}{\partial w} J(n) \qquad (5)$$

$$C_k(n+1) = C_k(n) - \mu_c \frac{\partial}{\partial C_k} J(n) \qquad (6)$$

$$\sigma_k(n+1) = \sigma_k(n) - \mu_\sigma \frac{\partial}{\partial \sigma_k} J(n) \qquad (7)$$

Where $J(n) = \frac{1}{2}|e(n)|^2$, e (n) =d (n) - y (n) is the error, d (k) is the target output and y (k) is the predicted output. μ_w μ_C And μ_σ are the learning parameters of the RBF network. The complete process of the proposed feature

extraction based classification process is presented in Fig. 3.

Simulation Result and Discussion:

The performance of the proposed method of feature extraction and classification is validated with a set of bench mark dataset such as Leukemia, SRBCT and MLL Leukemia and colon. All the datasets categorized into two groups: binary class and multi class to assess the performance of the proposed method. The Leukemia dataset is binary class and both SRBCT and MLL Leukemia are Multi class datasets. The feature selection process proposed in this paper has two steps. First the microarray data is decomposed by principal component analysis that optimally choose the discriminate feature set and then using Discrete wavelet transform into level 4 using db7 wavelet to get the approximation coefficients as the extracted feature set.

The performance of the proposed feature extraction method is analyzed with the low complexity neural network classifiers such as RBFNN. The leave one out cross validation (LOOCV) test is conducted by combining all the training and test samples for both the classifiers with all the three datasets and the results are listed in Table 1. For binary class the performance of RBFNN is comparable to MLP, but in case of multi class it outperforms the MLP. . The efficiency of the proposed method in predicting the class of the cancer microarray data using standard datasets is analyzed in table. The flowchart of the proposed feature extraction based tumor classification method is shown in Figure 3. Figure shows the step by step procedure that performs mixed PCA and wavelet transform method for feature extraction. The reduced data is subjected to neural network classifier based on radial basis function neural network.

The performance of dimensional reduction methods is shown in the Table 1. The performance of the proposed method is also compared with those obtained by the various methods and the results are shown in Table 2-4. The existing methods are used for cross validation test on the datasets. From Tables 2-4 it reveals that the proposed method is comparatively better the existing methods with the advantage of reduced computational load.

Table: 1 Dimension reduction methods for the dataset

Dataset	Original Dimension	PCA	DWT (Dubecies7) Level 4
SRBCT	83/2308	83/600	83/176
MLL Leukemia	72/12582	72/600	72/176
Colon	62/2000	62/600	62/176

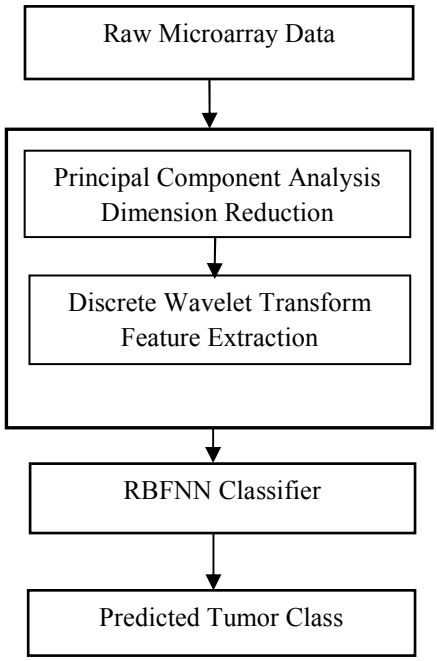

Figure: 3. Flow graph of the proposed feature extraction based tumor classification method

Table: 2. Comparison study of classification accuracy

Dataset	Method	Classification Accuracy
SRBCT	SVM	93.65%
	MLP	90.36%
	RBFNN	98.59%
MLL Leukemia	SVM	85%
	MLP	87.50%
	RBFNN	98.83%
Colon	SVM	91.5%
	MLP	93.54%
	RBFNN	96.33%

Table: 3. Prediction accuracy measures of SRBCT dataset

Methods	Classification accuracy
SLDA	100 %
BWNN	96.83 %
C4.5	91.18%
Bagboost	95.24 %
SVM	93.65 %
TPCR	100 %
Gradient LDA	100 %
FA+ NN	97.59 %
PCA+Wavelet+RBFNN	100%

Table: 4. Prediction accuracy measures of MLL Leukemia dataset

Methods	Classification accuracy
C4.5	73 %
Bagging C4.5	86.67 %
Adaboost C4.5	91.18 %
Combined feature selection + ensemble neural network	100 %
FA +NN	96.87%
PCA+Wavelet+RBFNN	100%

Conclusion:

The proposed mixed feature extraction method using the principal component analysis in conjunction with signal processing based wavelet transform has been used to effectively select the discriminative genes on microarray data and it not only enables to reduce the dimension of the dataset but also it reduces the computational complexity. A simple RBFNN based classifier enables to classify the microarray samples efficiently using the reduced extracted feature data. The comparison results elucidated that the proposed approach is faster and has better accuracy and reduced computational complexity.

Acknowledgement:

The author would like to thank the Management members and Principal for establishment of R & D Centre in the College and providing the required infrastructure and other supports to carry out the research work

References:

1. Asyali, M.H., Colak, D., Demirkaya, O., Inan, M.S.: Gene expression profile classification: A review. Current Bioinformatics 1, 55–73 (2006)
2. Xiong M., Jin L., Li W. and Boerwinkle E. Computational methods for gene expression-based tumor classification. BioTechniques, 2000, vol. 29, no. 6, pp. 1264–1268.
3. Nebu Varghese, Vinay Verghese, Gayathri. P and N. Jaisankar, "a survey of dimensionality reduction and classification methods" International Journal of Computer Science & Engineering Survey (IJCSES) Vol.3, No.3, pp 45-54, June 2012
4. Ali Anaissi#1, Paul J. Kennedy#2,Madhu Goyal, "Dimension Reduction of Microarray Data Based on Local Principal Component", World Academy of Science, Engineering and Technology 53 2011.
5. Meher J. K., Barik R.C., Panigrahi M.R., "Cascaded Factor Analysis and Wavelet Transform Method for Tumor Classification Using Gene Expression Data", IJITCS, 2012, Vol4, No.9, PP.73-79.
6. Baldi P. and Long A.D. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics, 2001, vol. 17, no. 6, pp. 509–519.
7. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., Lander, E. S. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring Science, 1999, 286(5439), pp.531-537.
8. Huang D.S. and Zheng C. H. Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. Bioinformatics, 2006, vol. 22, no. 15, pp. 1855–1862.
9. Yeung K.Y., Ruzzo W. L. Principal component analysis for clustering gene expression data. Bioinformatics, 2002, 17, pp.763–774.
10. Yihui Liu. Wavelet feature extraction for high-dimensional microarray data. Neurocomputing, 2009, Vol. 72, pp. 985-990.
11. Yihui Liu. Detect Key Gene Information in Classification of Microarray Data. EURASIP Journal on Advances in Signal Processing, 2007 pp.1-10.
12. A. Anaissi ,P. Kennedy and M. Goyal, A Framework for Very High Dimensional Data Reduction in the Microarray Domain . IEEEBITA, 2010.
13. Kuldip K. Paliwal , Alok Sharma, "Improved direct LDA and its application to DNA microarray gene expression data,". Pattern Recognition Letters (41) 2010, ScienceDirect 0167-8655
14. Tan AC, Gilbert D. Ensemble machine learning on gene expression data for cancer classification. Applied Bioinformatics, 2003, 2, pp.75-83.
15. Dettling M. Bag Boosting for tumor classification with gene expression data. Bioinformatics, 2004 vol. 20, no. 18, pp. 3583–3593.
16. Guyon I, Weston J, Barnhill and Vapnik V. Gene selection for cancer classification using support vector machines. Mach. Learn, 2002, 46, pp. 389- 422.
17. Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., Meltzer, P. S. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Medicine, 2001, 7(6), pp.673-679.
18. O'Neill MC and Song L. Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect. BMC Bioinformatics, 2003, 4:13.
19. Liu Bing, Cui Qinghua, Jiang Tianzi and Ma. Songde. A combinational feature selection and ensemble neural network method for classification of gene expression data. BMC Bioinformatics, 2004. 5:136, pp. 1-12.
20. Edmundo Bonilla Huerta, Beatrice Duval, Jin-KaoHao, "A hybrid LDA and genetic algorithm for gene selection and classification of microarray data," Neurocomputing (73) 2010, ScienceDirect 0925-2312.
21. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Elsevier (1990)
22. Pearson, K.: On lines and planes of closest fit to systems of points in space.Philosophical Magazine,2(6), 559–572 (1901)
23. Jolliffe, I.T. (1986). Principal Component Analysis. Springer, New York.

24. Massey, W.F. (1965) Principal components regression in exploratory statistical research. Journal of American Statistical Association, 60, 234-246.
25. S. Bicciato, A. Luchini, C. Di Bello, "Disjoint PCA Models For Marker Identification And Classification Of Cancer Types Using Gene Expression Data,"2002, IEEE 0-7803-7557-2.
26. Matthew Partridge, Rafael Calvo, "Fast Dimensionality Reduction and Simple PCA,"2006.
27. Grossmann A. and Morlet J. Decomposition of Hardy functions into square integrable wavelets of constant shape. SIAM Journal on athematical Analysis, 1984, vol. 15, no. 4, pp.723–736.
28. Mallat S. G. A theory for multiresolution signal decomposition: the wavelet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1989, vol. 11, no. 7, pp. 674–693.
29. Girija Chetty, Madhu Chetty, "Multiclass Microarray Gene Expression Classification based on Fusion of Correlation Features".
30. Dettling, M. and Buhlmann, P. (2003) Boosting for tumor classification with gene expression data. Bioinformatics, 19, 1061-1069.