An Approach to reduce the large feature space of Microarray Gene Expression Data by gene clustering for efficient sample classification

Sheela T^{#1}, Lalitha Rangarajan^{#2}

#1: Department of Studies in Computer Science, University of Mysore, Mysore,
Karnataka, India

#2: Department of Studies in Computer Science, University of Mysore, Mysore,
Karnataka, India,

ABSTRACT

Microarray Gene expression data has a huge number of genes as features and a very small number of samples. This paper proposes a method of dimensionality reduction for high-dimensional microarray gene expression data. The large dimensionality of the feature space in microarray dataset is reduced by assimilating the features into line segments. Before applying dimension reduction, genes with similar properties are grouped and significant genes from each group are selected by ranking the genes. Among the various traditional classifiers applied on gene expression data, we have used the most widely used, k-Nearest Neighbor (k-NN), Support Vector Machine (SVM), Naive Bayes (NB) and Decision Tree (DT) classifiers, in our work. Experiments conducted on binary class microarray data sets demonstrate that, the proposed method performs well and achieves better classification accuracy and can be a useful approach for cancer data classification.

Keywords: Dimension reduction, Feature space, Cluster analysis, significant genes, and Classifier performance.

Corresponding Author: Sheela T.

INTRODUCTION

Microarray measures expression data of thousands of genes simultaneously. These expression data are arranged into a matrix, where, rows are samples and columns represent number of genes. Each element of the matrix is the expression level of a gene under given condition and is given as a real number. Classification is an important data mining technique in the area of bioinformatics. By using microarray technology, computer researchers having no previous biological knowledge, are trying to classify cancer samples [6, 12]. By classifying microarray data into normal type of samples and cancerous type, it is possible to diagnose cancer and find treatment for that. Standard

machine learning techniques cannot perform well for cancer classification, due to the problem of high dimensionality and small sample size [17] of the datasets. Dimensionality reduction improves the classification performance by removing the redundant and irrelevant genes [3]. Also, reducing the number of genes will reduce the computational expenses and helps researchers to concentrate on the informative genes for biological validation.

Feature selection and Feature Extraction are the two methods of dimension reduction. In feature selection [34, 45], significant genes which help in differentiating the classes are identified. Filter model and wrapper model are the two techniques of feature selection [21]. In filter method, gene selection is independent of learning algorithm. Ranking is the most commonly used filter gene selection method in cancer classification problem. Ranking method uses an evaluation statistic for each gene and selects genes which are most discriminative between two groups of samples and arranges the genes in decreasing order of statistical score. In our work, we have chosen the most popular ranking method, student t-test [25] for ranking gene features.

With the help of microarray gene expression data, heterogeneous cancers can be classified, but the challenge is its high dimension. When the data cannot be processed because of its large size, divide and conquer concept can be used and the data is divided into subsets [29, 30] and then process these subsets. In microarray research, it has been proved that related genes have similar expression patterns across all samples [13, 40]. The huge dimension of the microarray gene expression data can be reduced by clustering the feature space into gene groups with similar expression patterns and informative genes selected from each of these clusters are used for sample classification.

Many of the earlier classification methods on microarray data give a poor classification result because of the noise occurring from irrelevant and redundant variables. An interesting approach to reduce or summarize variables would be to use symbolic data analysis (SDA) introduced by [5, 10, 11]. In this direction, Multi-dimensional data reduction is proposed in [22], where, regression line concept is used to reduce the feature values of a sample to only two features - slope and intercept. In our work, we have proposed dimensionality reduction method which reduces the set of features into symbolic features that are fewer and the classification is performed on the reduced dataset.

RELATED WORK

The method of grouping gene features for reduction before classification is used by many researchers for microarray data analysis. The basic approach in [19] is to cluster the genes, and then, based on the quality of the cluster, one or more representative genes are selected from each cluster. In Attribute Clustering Algorithm [2], mutual information is used to cluster genes and top-ranked genes are selected from each cluster for analysis. Distinct patterns in gene expression dataset are discovered [23] and gene clustering is done for reducing the data dimension. A two-stage classification method for microarray dataset is described in [43], where, in the first stage, top ranked genes are selected and passed to the second stage for classification. DRAGS (Dense Relevant Attribute Group Selector) method [44] groups features based on DGF (Dense Group Finder) algorithm, where, features in a group are nearer to the same density peak and one representative feature is selected from each group for classification. In two stage algorithm proposed in [32], t-test is used to identify the top most gene in each cluster and Sequential Forward Feature Selection is applied to these selected genes and then classification is performed. K-means

method is used to group gene feature set into clusters [18], and Particle Swarm Optimization (PSO) with Extreme Learning Machine (ELM) is used for gene selection.

DESCRIPTION OF THE METHODS USED

K-means Clustering

K-means method performs clustering in very few iterations. This algorithm divides the dataset into K disjoint subsets by taking the number of clusters (K) as input. Genes are grouped by calculating the distance between each gene and the center of each cluster, such that, the distance between genes within a cluster is minimum and between genes of different clusters is maximum. Since, the correct number of clusters cannot be estimated, the algorithm must be repeated for different values of K and clustering results are compared. In literature, genes are grouped by various diverse clustering methods. It has been shown that K-means clustering can generate stable gene groups, which are predictive [15, 44]. K-means algorithm uses Euclidian distance or Pearson correlation distance. In our work, we have used Pearson correlation coefficient distance in K-means clustering, as it is found to be effective for gene expression data.

T-test Ranking

We have used student t-test statistical method to identify significant genes in each group [32, 25]. T-test is a parametric method, it finds features where mean value is maximum between groups and variability is minimum within each group. Genes with higher t-scores are considered important, as they show significant difference between groups. The test is performed on each gene and the genes are sorted in descending order of t-statistic value so that the most significant genes can be selected. t score of each gene is calculated using equation (1).

$$t = \frac{(\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \quad (\text{Error! Bookmark not defined.})$$

Where, t is the t score, μ_1 μ_2 are sample mean, s_1 s_2 are sample standard deviation, n_1 n_2 represent number of samples in the two classes class1 and class2.

Proposed Feature Reduction Method

The top ranked genes in each group are summarized into four features by our proposed feature space reduction method as explained below:

Successive feature values $f_i, f_{i+1}, f_{i+2} \dots f_{i+m}$ are assimilated into a line segment, if the points $(i, f_i), (i+1, f_{i+1}), \dots (i+m, f_{i+m})$ are almost on a line. This can be inferred by comparison of slopes of adjacent points. For example, $(i, f_i), (i+1, f_{i+1}), (i+2, f_{i+2})$ are approximated by a line if the slope of line joining first two points is approximately equal to the slope of line joining later two points. That is, if $|(f_{i+1} - f_i) - (f_{i+2} - f_{i+1})| \leq \epsilon$. We conducted experiments on a range of thresholds and ϵ is set to 0.5. 'n' successive feature values are approximated by a single line segment, if the slopes of line joining adjacent points are nearly equal. Thus the features are transformed into fewer number of line segments.

A feature vector may have any number of line segments. We then compute average slope of all the line segments of a feature vector. The transformed symbolic features are the number of line segments and the average slope of all line segments. The process is repeated for all samples in

that group. Each feature set is represented by a feature descriptor of the form $V = (v_1, v_2, v_3, v_4)$, where v_1 is the number of line segments, v_2 is the average slope of these line segments, v_3 is the first feature of the sample and v_4 is the last feature of the sample. v_3 and v_4 are required to distinguish one sample from another, as the number of line segments and average slope may be same for two or more samples. To distinguish such samples we have included two more features v_3 and v_4 . Thus, the number of features is reduced to 4. Gene expression values in a cluster are more or less the same for all samples from the same class. The number of line segments and the average slope will be same for samples from the same class. Instead of comparing individual expression values, we need comparison of only 4 feature values after one transformation to line segment. The process is repeated for all clusters. The number of features in each sample for final classification is K (number of groups) $\times 4$. Example of line segments generated for feature values of a cluster of some sample in leukemia dataset is shown in Figure 1. Framework of feature reduction by our proposed method is shown in Figure 2.

Fig 1: Transformation of features into line segments

Fig 2: Framework of feature reduction by gene grouping and classification technique

Proposed Model

Algorithm: Feature reduction and classification

Input: Microarray dataset D of size $N \times m$

Output: A class label vector

Method:

Step 1: Features are grouped into K groups using K-means clustering method and set of gene groups $GS_1, GS_2, \dots, GS_K$ are generated.

Step 2: For each group GS_i (for $i = 1$ to K) repeat step 3 through step 6.

Step 3: Select significant genes using t-test where $p\text{-value} \leq 0.05$

Step 4: For every sample repeat step 5 and step 6.

Step 5: Sequence the genes in the order of selection from step 3 and fit lines for gene expression values.

Step 6: Compute the features (v_1, v_2, v_3, v_4)

Step 7: Classify samples using new reduced feature set.

Step 8: Stop.

DATA SETS AND EXPERIMENTS

The binary class microarray datasets used in our experimentation are summarized in Table 1. The number of samples and number of genes in each dataset is also given. The datasets are normalized to have zero mean and unit standard deviation.

Table 1. Description of the Datasets

<i>Dataset</i>	<i>Genes(m)</i>	<i>Samples(N)</i>	<i>Reference</i>
Leukemia	7129	72	[17]
Prostate	6033	102	[37]
CNS	7129	34	[28]
DLBCL	5470	77	[36]
Ovarian	4000	216	[24]
Breast	4348	97	[41]

We used the following parameters for the classifiers: In k-Nearest Neighbors (kNN) classification algorithm [14] Euclidean distance metric is used and experiments are conducted for $k=1, 3, 5, 7$ and best results are reported. For Support vector machine (SVM) classifier [9], a linear kernel is used. For Decision trees (DT) [31] method, 10 is the minimum parent size and 1 is the minimum leaf size. In the Naïve Bayes classification (NBC) algorithm [20] a kernel distribution was used with predictors. For evaluation we have used the leave-one-out cross-validation method, as microarray datasets have very few samples. Accuracy is used as the performance measure. For experimentation, we have used WEKA data mining tool [42].

RESULTS AND DISCUSSION

Results of the experiments conducted on the datasets are given in this section. Table 2 shows the results obtained for various cluster sizes. The first column has the result of classification accuracy for the original data without grouping or feature reduction and the next columns show the accuracy for the datasets grouped into different number of clusters ($K=2,3, \dots 10$) and then reduced using the proposed method.

Table 2. Classification accuracy (%) of datasets

Dataset		Original data	Number of Clusters								
			2	3	4	5	6	7	8	9	10
Leuk	kNN	91.67	97.22	94.44	98.61	95.83	97.22	98.61	95.83	93.06	98.61
	SVM	95.83	98.61	95.83	98.61	94.44	98.61	98.61	98.61	95.83	97.22
	NBC	87.5	97.22	93.06	91.67	97.22	93.06	97.22	95.83	97.22	88.89
	DT	94.44	91.67	90.28	98.61	91.67	90.28	91.67	91.67	94.44	94.44
Prostate	kNN	86.27	96.1	96.1	89.61	96.1	97.4	92.21	93.51	85.71	96.1
	SVM	92.16	89.61	97.4	92.21	97.4	99.02	97.4	90.91	93.51	93.51
	NBC	83.33	97.4	97.4	90.91	98.61	96.1	90.91	97.4	92.21	94.81
	DT	85.29	93.51	89.61	93.51	94.81	98.04	96.1	92.21	97.4	97.4
DLBCL	kNN	92.16	96.08	92.16	92.16	98.00	96.08	96.08	96.08	98.04	96.08
	SVM	88.24	95.10	91.18	92.16	95.10	95.10	95.10	91.18	90.20	95.10
	NBC	82.45	88.24	84.31	89.22	89.22	88.24	88.24	89.22	88.24	83.33
	DT	82.35	82.35	80.39	86.27	80.39	82.35	82.35	86.27	82.35	84.31
Ovarian	kNN	85.65	89.35	90.28	88.89	89.81	91.67	85.65	83.80	89.81	89.81
	SVM	91.2	98.15	90.28	92.59	98.15	91.20	96.30	95.37	98.15	92.59
	NBC	82.29	83.33	87.96	91.67	84.26	91.67	83.89	84.72	84.26	91.67
	DT	80.9	85.19	87.96	83.80	86.11	87.50	86.57	86.57	86.57	83.80
Breast Cancer	kNN	93.44	95.88	95.88	97.94	97.94	93.81	95.88	98.97	95.88	95.88
	SVM	93.81	96.91	93.81	93.81	92.78	96.91	93.81	97.94	93.81	96.91
	NBC	89.22	94.85	94.85	96.91	94.85	95.88	94.85	97.94	94.85	94.85
	DT	85.29	96.91	87.63	93.81	96.91	93.81	87.63	98.02	87.63	96.91
CNS	kNN	88.24	91.18	95.12	91.18	94.12	88.24	88.24	94.12	91.18	91.18
	SVM	85.29	88.24	94.12	85.29	82.35	91.18	91.18	88.24	94.12	85.29
	NBC	85.29	79.41	88.24	86.24	85.29	85.29	79.41	85.24	79.41	88.24
	DT	80.15	82.34	88.24	85.29	87.63	88.24	82.35	82.35	88.24	85.29

Validity of the obtained results

The number of optimal clusters in each dataset is verified using Silhouette scores [33]. Silhouette value for each point is a measure of how similar that point is to points in its own cluster compared to points in other clusters and it gives a view of the compactness and separation of clusters. Silhouette score is used to estimate the number of clusters for gene expression data [4, 38, 39] and higher silhouette value indicates a good cluster. Silhouette plot obtained for the experimental datasets for 2 to 10 clusters is shown in Figure 3. The figure reports the optimal number of clusters for the datasets.

Fig 3: Silhouette plots for all the experimental data sets for clusters = 2, 3, ... 10

From the above Table 2, it is evident that we get best classification accuracy for the optimal number of clusters for most of the experimental datasets. Classification accuracies for these

datasets reported in various literature is given in Table 3. The highest classification accuracy obtained by our proposed model for the selected datasets is also given in Table 4.

Table 3. Classification accuracies obtained from various methods

Reference	Dimension reduction method	Classifier	Dataset Name	Accuracy
[26]	Multi-Filter enhanced genetic ensemble(MF-GE)	DT	Leukemia	84.51
		kNN		95.48
		NBC		96.27
[27]	Support vector based t-statistic ranking (SVt-RFE)	SVM	Leukemia	98.41
			Prostate	97.18
			CNS	92.05
			Breast Cancer	92.47
[16]	Diverse Accurate Feature Selection (DAFS)	SVM	Leukemia	97.5
			Prostate	92.3
			DLBCL	98.0
[8]	Effective Range based Gene Selection (ERGS)	NBC	Leukemia	98.61
			Prostate	94.12
			DLBCL	94.79
		SVM	Leukemia	100
[35]	(Random kNN Feature Selection) RKNN-FS	RKNN	Prostate	93.14
			CNS	96.88
[7]	SVM-Recursive Feature Elimination (SVM-RFE)	3NN	Leukemia	98.0
		3NN	Prostate	95.0
		NB	CNS	93.0
		3NN	Leukemia	100
		SVM	Prostate	98.89
[1]	Fast Correlation Based Filter(FCBF)	Ensemble Learning	DLBCL	65.89
			Ovarian	100
			CNS	75.49
[1]	Fast Correlation Based Filter(FCBF)	Ensemble Learning	Leukemia	98.77
			Ovarian	99.40
			Breast Cancer	97.88

Table 4. Classification accuracies obtained by the proposed method

Datasets	Classifier accuracy by the proposed method				Maximum accuracy reported in literatures from Table 3
	kNN	SVM	NBC	DT	
Leukemia	98.61	98.61	97.22	98.61	100 [7,8]
Prostate	97.4	99.02	98.61	98.04	98.89 [7]
DLBCL	98.04	95.10	89.22	86.27	98.0 [16]
Ovarian	91.67	98.15	84.72	87.96	100 [7]
Breast Cancer	98.97	97.94	97.94	98.02	97.88 [1]
CNS	95.12	94.12	88.24	88.24	93.0 [35]

CONCLUSION

Proposed method is a useful approach for cancer data classification. The methods taken for comparison have high degree of complexity and the proposed method is computationally simple. The complexity of reduction of a sample in a cluster to 4 features is m , where m is the number of genes in the cluster. For a cluster, the complexity is $\Theta(N \times m)$ and this is repeated across K number of clusters. Thus the complexity of transforming all samples is $N \times K \times \text{Number of genes selected}$.

REFERENCE

- I Osareh A, Shadgar B. An efficient ensemble learning method for gene microarray classification. *BioMed Res Int*, 2013.doi: <http://dx.doi.org/10.1155/2013/478410>.
- II W HAU, ChanKCC, Wong AKC, Wang Y. Attribute clustering for grouping, selection, and classification of gene expression data. *IEEE/ACM Trans. ComputaionBiology and Bioinformatics*, 2005, vol 2, No2, pp.83–101.
- III Baldi P, BrunakS. *Bioinformatics: The Machine Learning Approach*, 2001,MIT Press(2nd ed).
- IV Bandyopadhyay S Mukhopadhyay A,Maulik U. An improved algorithm for clustering gene expression data.*Bioinformatics*, 2007, 23(21): 2859–65.
- V Bock HH, Diday E. *Analysis of Symbolic Data. Study in Classification, Data Analysis and Knowledge Organization*. Springer Verlag, 2000.
- VI Boulesteix AL, Strobl C, Augustin, T, Daumer M. Evaluating microarray-based classifiers: an overview. *Cancer Inform*,2008, 6:77–97.
- VII Carlos J, Alonso Gonzalez,et al. Microarray gene expression classification with few genes: Criteria to combine attribute selection and classification methods. *Expert Systems with Applications*,2012, 39; 7270-7280.
- VIIIChandraB, GuptaM. An efficient statistical feature selection approach for classification of gene expression data. *Journal of Biomedical Informatics*, 2011,vol.44,No.4,pp.529-555
- IX CortesC, Vapnik V. *Support-Vector Networks*. *Mach Learning*,1995, 20(3):273–297.
- X DidayE.An introduction to symbolic data analysis, Tutorial of 4th Conf. of IFCS, 1993,Paris.
- XI Diday E. An introduction to symbolic data analysis and sodas software.*Electro.J.Symb.Data Anal.*,2002,1-25.
- XII Dupoy A, Simon RM. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *J Natl Cancer Inst.*,2007, 9:147–57.
- XIIIEisen MB, et al. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 1998, 95, 14863–14868.
- XIV Friedman JH, Bentley JL, Finkel RA. An Algorithm for Finding Best Matches in Logarithmic Expected Time.*ACM Trans.Math.Softw.*,1977, 3(3):209–226.
- XV GeY, Sealfon SC. FlowPeaks: a fast unsupervised clustering for flow cytometry data via K-means and density peak finding. *Bioinformatics*,2012, 28(15):2052-8.

- XVI Giuliano Armano, et al. A New Gene Selection Method Based on random Subspace Ensemble for Microarray cancer Classification. *PRIB-2011*, LNBI 7036, pp.191-201
- XVII Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *SCIENCE*, 1999, Vol 286,531–7.
- XVIII Han F, SunW,Ling Q-H. A novel strategy for Gene selection of microarray data based on Gene-to-class Sensitivity Information.*PLoS ONE*, 2014, 9(5):e97530.
- XIX Jaeger J,SenguptaR,Ruzzo W L. Improved gene selection for classification of microarrays. *Pac. Symp.Biocomput*, 2003, 8, 53-64.
- XX JohnGH,LangleyP. Estimating Continuous Distributions in Bayesian Classifiers. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.*1995.
- XXI R Kohavi, George HJohn. Wrappers for feature subset selection.*Artificial intelligence*, 1997, Volume 97, Issues 1-2, Pages 273-324.
- XXII Lalitha Rangarajan, Nagabhusa P. Dimensionality reduction of multi dimensional temporal data through regression. *Pattern Recognition Letters*, 2004, Vol 25/8, pp.899- 910.
- XXIII TengL, ChanL. Discovering Distinct Patterns in Gene Expression Profiles. *Journal of Integrative Bioinformatics (JIB)*, 2008, Volume 5, Issue 2, Pages 221-233.
- XXIV Marci ESchaner, et al. Gene Expression Patterns in Ovarian Carcinomas. *MolBiol cell*, 2003, 14(11):4376-4386.
- XXV NguyenT, KhosraviA,Creighton D, NahavandiS. Hierarchical Gene Selection and Genetic Fuzzy System for Cancer Microarray Data Classification. *PLoS ONE*,2015, 10(3):e0120364.
- XXVI Pengi Yang, et al. A multifilter enhanced genetic ensemble system for gene selection and sample classification of microarray data. *BMC Bioinformatics*,2010, II (Suppl I):S5.
- XXVII Piyush KumarA, Mundrajagath,CRajapakse. Gene and sample selection for cancer classification with support vectors based t-statistic. *Neurocomputing*, 2010, 73, 2353-2362.
- XXVIII Pomeroy S, Tamayo P, Gaasenbeek M, et al. Prediction of central nervous system embryonal tumour outcome based of gene expression. *Nature*,2002, 415: 436–442. doi: 10.1038/415436a.
- XXIX Prasad S, Bruce LM, Ball J E. A multi-classifier and decision fusion framework for robust classification of mammographic masses..*ConfProc IEEE Eng Med Biol Soc.*, 2008, 3048-51. doi: 0.1109/IEMBS. 2008.4649846.
- XXX Qi Guo, Bo-Wei Chen, Feng Jiang, XiangyangJiand Sun-Yuan Kung. Efficientdivide-and-conquer classification based on feature-space decomposition. *IEEE Syst. J.*2015.
- XXXI Quinlan JR. Simplifying decision trees. *International Journal of Human-Computer Studies*,1999, 51 (2):497.

- XXXII RajaniBala, Ramesh Kumar Agrawal. Clustering in Conjunction With Wrapper Approach to Select Discriminatory Genes For Microarray Dataset Classification. Computing and Informatics, 2012, Vol. 31, 921–938.
- XXXIII Rousseeuw P. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comp. App. Math, 1987, vol. 20, pp. 53–65.
- XXXIV Y Sayes, I Inza, P Larrañaga. A review of feature selection techniques in bioinformatics. Bioinformatics, 2007, vol. 23, no. 19, pp. 2507–2517.
- XXXV Shengqiao Li, et al. Random KNN feature selection – a fast and stable alternative to Random Forests. BMC Bioinformatics, 2011, 12:450.
- XXXVI Shipp MA, Ross KN, Tamayo P, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med, 2002, 8: 68–74. doi: 10.1038/nm0102-68.
- XXXVII Singh D, et al. Gene expression correlates of clinical prostate cancer behavior. Cancer Cell, 2002, Vol. 1.
- XXXVIII Speer N, Fröhlich H, Spieth C, Zell A. Functional distances for genes based on go feature maps and their application to clustering. CIBCB-2005, 142-149.
- XXXIX Speer N, Fröhlich H, Spieth C, Zell A. Functional grouping of genes using spectral clustering and gene ontology. IJCNN-2005, 1:298-3031.
- XL Tamayo P, et al. Interpreting patterns of gene expression with self-organizing maps : methods and application to hematopoietic differentiation. Natl AcadSci USA, 1999, 96(6):2907-12.
- XLIV Van't Veer LJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature, 2002, 415, 530-536.
- XLII Weka: A multi-task machine learning software. <http://www.cs.waikato.ac.nz/ml/weka>.
- XLIII Wong TT, Hsu CH. Two stage classification methods for microarray data. Expert Syst Appl, 2008, 34:375-83.
- XLIV Yu L, Ding C, Loscalzo S. Stable feature selection via dense feature groups. In: Proceeding of the 14th ACM SIGKDD international conference on knowledge discovery and data mining. Las Vegas, USA, 2008, ACM. p. 803–11.
- XLV Hira ZM, Gillies DF. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. Advances in Bioinformatics, 2015, 1-13.