JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Classification of Binary Class Microarray data using Symbolic Classifier

[1]Dr. Sheela T, [2]Santhosh Kumar B N, [3]Prakasha Raje Urs M

[1]Associate Professor, [2]Assistant Professor, [3]Assistant Professor
[1]Computer Science,
[1]Maharani's Science College for Women, Mysore, Karnataka, INDIA

Abstract: Cancer classification is routinely done using gene expression data. With microarray technology, monitoring thousands of genes is an easy task. The reliable and precise classification of different tumour types is very important in cancer classification and drug discovery which is useful in providing better treatment. Microarray gene expression data analysis is extensively used for human cancer diagnosis and classification. Various methods of classification from the field of statistics and machine learning have been used to classify cancer microarray data. However, the large number of features with very few samples in the data is a challenge to the existing classification methods. In this work, our experiments are based on the symbolic classifier for predicting sample class labels. Experiments show that the proposed method will achieve high classification accuracies with very few genes.

IndexTerms **- Microarray Gene Expression data, Cancer Classification, High Dimensionality, Symbolic Classifier**

I. INTRODUCTION

The design of DNA microarray technology has made it easy to monitor thousands of genes simultaneous. The gene expression levels has the solution for the basic problems relating to the identification, prevention, drug discovery and cure of diseases [1]. Research has proved that this technology can be useful in the classification of cancers [2]. Several different methods have been proposed for classifying microarray gene expressions with good results. But there is need for further exploration.

In the microarray dataset, gene expression values are arranged in a matrix form, where, samples are rows and genes/features are columns. The elements of the matrix are real numbers and they represent the expression levels of genes under a specific condition. Classification of microarray data separates or distinguishes healthy samples from cancer samples, and is helpful in predicting response to therapy [3]. Microarray data analysis is required for early tumour and cancer discovery and it can help in cancer diagnosis and clinical treatment [4, 5, 6].

Standard machine learning techniques often fail to perform well for cancer classification, because of the huge number of genes as features with few samples in the microarray data [1]. Other characteristics of gene expression data, such as, a high level of noise, irrelevant and redundant data are the reasons for unreliable and low accuracy in analysis results. Selection of the relevant and significant genes will help in improving classification results [7, 8, 9, 10]. By retaining only significant features for classification, the performance, robustness and usefulness of classification algorithms can be improved.

In our work, student *t*-test gene ranking method is used to select significant genes. Experiments performed on three widely used binary class microarray datasets illustrate the efficiency of our proposed approach. A large number of standard classifiers are applied on microarray gene expression data, among them most widely used are k-Nearest Neighbor (k-NN) and Support vector Machine(SVM) classifiers. We have used these standard classifiers for comparison with our proposed class prediction method.

II. RELATED WORK

The work in [11] has addressed the problem of low signal-to-noise ratio in microarray data faced jointly with the high-data dimensionality problem by a method called GenSym. The basic idea is to take advantage of Symbolic Data Analysis capabilities with the use of interval representation to model uncertainty in microarray measurements, with the aim to design more accurate breast cancer management tools to help the physicians in their decision-making process. The feature selection algorithm InterSym [12] that handles symbolic interval data is used to derive a genetic signature. A preliminary computational study shows that the use of such strategy can improve and simplify significantly the cancer classification task by selecting a small number of relevant genes. A novel symbolic representation [13] is introduced, that can be used to cluster gene expression data. Also, here a procedure is presented for selecting a subset of biologically informative clusters by searching for overrepresented patterns in the data. The selection process is validated by running the algorithm on three different Datasets from Gene Expression Omnibus (GEO) Database. It has been shown in [14] that the discrete nature of symbolic representations is appealing because of the high levels of noise inherent in gene expression data. A popular example of a symbolic representation called Symbolic Aggregate approXimation (SAX) [15] has been applied to gene expression data through SLINGSHOTS in [16], which selects informative genes from gene expression data based on the symbolic representation.

III. METHODOLGY

The first step is to normalize the microarray gene expression data and then reduce this preprocessed data into smaller subset using t-test gene ranking method and most significant genes are selected. After gene selection, the proposed classification algorithm is applied to classify the reduced dataset.

3.1 Gene selection methods

T-test method: t-test statistical approach is a parametric method, which finds features where difference of mean is maximum between groups and within each group variability is minimum. In this method, genes with higher t-statistics are more significant [17]. The test is performed on each gene and the genes are arranged in decreasing order of t-statistic value so that the most significant genes can be selected. The t-score of each gene is calculated using equation (1).

$$t = \frac{(\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \qquad (1)$$

Where t is the t-score, μ_1 and μ_2 are sample means of class-1 and class-2, s_1 and s_2 are sample standard deviations of class-1 and class-2, n_1 and n_2 represent number of samples in class-1 and class-2 respectively.

3.2 Classifiers

The k-nearest neighbour (k-NN): This is one of the most simple and popular classifier introduced in [18]. The strategy of k-NN as a classifier is very simple and has two stages, k nearest neighbors are determined in the first stage and the class of sample is determined using those neighbors in the second stage. An advantage of this method is, it keeps all training instances. Here, all samples are in the vector form and a test sample is taken and distance between the test sample and each of the training sample in the vector form is computed. The training sample closest to the test sample is identified as its Nearest Neighbor and is recognized as the one most similar to the test sample and the class label of this most similar training sample is allocated to the test sample.

Support Vector Machine (SVM) Classifier: SVM is a powerful and effective classification algorithm, introduced by [19] for microarray data classification. The data in the large datasets are classified by SVM using a separating surface that is either linear or non-linear, in the input space of a data set. The separating surface depends on a subset of the original data, called support vectors. In a high dimensional data space, these support vectors construct a hyper plane or set of hyper planes, which are used for classification. The hyper plane that has the maximum distance to the nearest training data points of any class gives a good separation and is called functional margin. The generalization error of the classifier is minimum when the functional margin is maximum. In SVM method, the input data are transformed into an n-dimensional space by using kernel functions that are built around the SVM models, and a hyper plane is constructed in the n-dimensional space to partition the data.

3.3 Proposed Classification Method

Our study focuses on the performance of symbolic classifier on the high dimensional, binary class gene expression data.

3.3.1 Symbolic Representation For Microarray Gene Expression Data:

The recent developments in the area of symbolic data analysis have proved that the real life objects can be better described by the use of symbolic data, which are extensions of classical crisp data [13, 20]. Symbolic interval features are extensions of pure real data types, in the way that each feature may take an interval of values instead of a single value. Microarray datasets have considerable intra class variation. Using conventional data representation preserving these variations is difficult. Symbolic data analysis which has the ability to preserve the variations among the data more effectively.

Let $[S_1,S_2,S_3,\ldots S_n]$ be a set of n samples of a gene expression dataset of class C_j ; j=1,2,3,…N (N denotes the number of classes).
And $G_i = [g_{i1}, g_{i2}, g_{i3}, \ldots g_{im}]$ be the set of m features characterizing the Gene expression sample S_i of the class C_j ; j=1,2,3,…N (N denotes the number of classes).

Each k^{th} feature value of the class Cj is represented by the use of interval valued feature $[\bar{g_{jk}}, g_{jk}^{+}]$

Where $g_{j,k}^{+}$ and $\bar{g_{j,k}}$ are the maximum and the minimum of the k^{th} feature values obtained from all n samples of the class C_j. i.e., $g_{j,k}^{+} = \max(g_{1k}, g_{2k}, g_{3k}, \ldots g_{nk})$ and $\bar{g_{j,k}} = \min((g_{1k}, g_{2k}, g_{3k}, \ldots g_{nk})$

Hence, the interval $[\bar{g_{jk}}, g_{jk}^{+}]$ represents the upper and lower limits of a k^{th} feature value of gene expression data.

Now, the reference vector is formed for the class C_j by representing each feature $G_i = [g_{i1}, g_{i2}, g_{i3}, \ldots g_{in}]$ in the form of an interval and is given by $R_j = \{[\bar{g_{j1}}, g_{j1}^{+}], [\bar{g_{j2}}, g_{j2}^{+}] \ldots\ldots [\bar{g_{jm}}, g_{jm}^{+}]\}$

This symbolic feature vector is stored in the database as a representative of the class j. Similarly, symbolic feature vectors are computed for all individual classes (j=1,2,3, ..., N) and stored in the database for the future classification purpose. Thus, the database has N number of symbolic vectors each corresponding to a class.

3.3.2 Classification

Classification of a new sample test gene expression data G_t is to compare it with all the reference vectors R_j, j=1, 2, 3, ..., N in the database to obtain the $'Ac'$ acceptance count for each reference sample. The new test sample is said to belong to class with which it has a maximum acceptance count. Acceptance count Ac is given as

$$Ac = \sum_{k=1}^{m} C(g_{tk}, [\bar{g_{jk}}, g_{jk}^{+}])$$

Where ,

$$C(g_{tk},[\,g_{jk}^{-},g_{jk}^{+}]) = \begin{cases} 1 \; if \; (\; g_{tk} \geq g_{jk}^{-} \; and \; g_{tk} \leq \; g_{jk}^{+}\;) \\ 0 \qquad\qquad\qquad otherwise \end{cases}$$

IV. EXPERIMENS AND RESULTS

Microarray Gene Expression datasets:

(1) Leukemia Dataset [1] has two classes, Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML). There are 72 samples in the dataset and each sample has expression patterns of 7129 genes.

(2) The Colon Tumor data set [21] contains 62 samples collected from colon-cancer patients which consists of 40 tumor samples and 22 normal samples and expressions for 2000 genes.

(3) The Prostate Dataset [22] contains 102 samples out of which 52 are tumor samples and 50 are healthy samples. The pre-processed dataset has expression values for 5966 genes.

Leukemia dataset is downloaded from Broad Institute of MIT and Harvard website (www.broadinstitute.org). Both Colon and Prostate datasets are downloaded from Kent Ridge Bio Medical repository [23]. All these data sets are two class gene expression profiles and are publicly available.

We have used the following parameters for the classifiers used in this work: k=1 for kNN classification algorithm and a Distance metric used for all instances of kNN was Euclidean. For SVM classifier, linear kernel is used. Microarray datasets are characterised by relatively few samples, hence, we used the leave-one-out cross-validation (LOOCV) method for evaluation and classification accuracy is the performance metric.

Table 4.1: Results for Leukemia dataset

Number of Genes	kNN	SVM	Proposed classifier
10	90.28	93.06	95.83
20	93.05	94.44	94.44
30	94.44	94.44	94.44
40	95.83	94.44	97.22
50	97.22	97.22	**98.61**
60	95.83	97.22	98.61
70	95.83	97.22	97.22
80	97.22	**98.61**	95.83
90	**98.61**	98.61	95.83
100	98.61	98.61	95.83
120	97.22	98.61	97.22
140	98.61	98.61	97.22
160	97.22	98.61	95.83
180	97.22	98.61	97.22
200	95.83	98.61	97.22

Table 4.2: Results for Colon dataset

Number of Genes	kNN	SVM	Proposed classifier
10	79.03	85.26	83.87
20	80.64	87.48	83.87
30	**82.26**	89.87	85.48
40	77.42	**92.03**	**89.26**
50	80.64	88.48	82.64
60	74.19	85.48	80.64
70	77.42	83.87	80.03
80	77.42	85.48	77.42
90	77.42	84.50	79.03
100	77.42	82.26	77.42
120	80.64	80.64	77.42
140	80.64	81.03	77.42
160	80.64	80.64	77.58
180	79.03	80.64	77.58
200	79.03	80.64	77.58

Table 4.3: Results for Prostate dataset

Number of Genes	kNN	SVM	Proposed classifier
10	89.17	90.19	89.25
20	90.19	92.16	91.24
30	90.89	93.29	93.29
40	94.14	**97.06**	95.21
50	93.29	95.19	93.29
60	93.29	95.19	**97.19**
70	91.19	93.14	94.23
80	93.18	95.08	94.03
90	93.18	94.06	95.21
100	**95.18**	94.06	94.23
120	94.17	95.09	93.29
140	93.29	95.09	92.23
160	94.19	96.19	92.59
180	93.14	94.12	91.29
200	93.14	94.12	91.29

Tables 1 to 3 gives the comparison between the classification accuracy of kNN, SVM and the proposed class prediction methods, for number of genes from 10 to 200. In each table, the maximum accuracy obtained for the classifier for specified number of genes is given in bold font. Figure-1 shows the accuracy plots for the three datasets for both the classifiers and gene selection methods. Table 4 lists the relevant works on cancer classification of microarray datasets along with the gene selection method and classifier used in those works.

Table 4.4: Relevant works on cancer classification with microarray dataset. Number of genes is given in parenthesis.

Reference	Gene selection method	Classification technique	Accuracy(Number of genes)
[24]	NA	Support vector based t-statistic ranking(SVt-RFE)	Leukemia : 98.41(64) Colon : 91.14(83) Prostate : 97.18(21)
[25]	Multi-Filter enhanced genetic ensemble (MF-GE)	kNN	Leukemia : 95.48 Colon : 77.01
[26]	RKNN-FS (Random kNN Feature Selection)	RKNN	Leukemia : 98.0 Prostate : 95.0

[27]	DAFS (Diverse Accurate Feature Selection)	SVM	Leukemia : 97.5 Prostate : 92.3
[28]	ERGS (Effective range based Gene Selection)	SVM	Leukemia : 100(80) Colon : 83.87(100) Prostate : 93.14(60)
[29]	t-score	Two Gene Classifiers TGC-1 TGC-2	Prostate :89.00 : 90.00
[30]	IFSER(Improved Feature Selection based on Effective Range)	kNN	Leukemia : 97.22(90)
[31]	FGSA (Forward Gene Selection Algorithm)	NA	Leukemia : 98.64
Proposed Classifier	**t-test**	**NA**	**Leukemia : 98.61(50)** **Colon : 89.26(40)** **Prostate : 97.19(60)**

IV. DISCUSSION

Microarray data contains irrelevant and noise genes. Classification results can be improved by selecting significant genes and reducing the number of irrelevant genes. The results in our experiments show that the number of genes selected affects the classification accuracy, but a very few significant genes in the dataset does not necessarily guarantee the highest accuracy. It is necessary to retain a reasonable number of genes for classification. The results show that not all gene selection methods improve the classification performance. We must decide appropriate gene selection method and classification algorithm. The t-test method ranks the genes based on significance but does not consider redundancy of selected genes.

From the experimental results in Tables 1 to 3 the following observations are made: For Leukemia and Prostate datasets, the proposed method yields a maximum accuracy of 98.61 (also same for kNN and SVM classifiers) for far less number of genes. However for Colon dataset there is a drop in the maximum accuracy in the proposed method by about 3% when compared to SVM, but is much better when compared to kNN classification. For different number of selected genes, the proposed class prediction method performs well compared with the standard classifiers kNN and SVM.

V. CONCLUSION

With the proposed method, the accuracy is decreasing with the increase in number of genes. With appropriate gene selection methods our proposed class prediction method will give more accurate results with microarray gene expression datasets. It is comparable with traditional classifiers kNN and SVM for binary class Microarray gene expression data. But appropriate number of genes that can be selected for better result cannot be decided and it is possible to select number of genes which achieve highest classification accuracy.

REFERENCES

[1] T.R.Golub *et al*.,"Molecular classification of cancer: class discovery and class prediction by gene expression monitoring", SCIENCE(1999), Vol 286,531–537, 1999.
[2] S.Cho and H. Won," Machine learning in dna microarray analysis for cancer classification", First Asia Pacific bioinformatic conference on Bioinformatics 2003:189–98, 2003.
[3] A.L.Boulesteix *et al*., "Evaluating microarray-based classifiers: an overview", Cancer Inform ; 6:77–97, 2008.
[4] A.Dupuy and R.Simon,"Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting", J Natl Cancer Inst ;9:147–57, 2007.
[5] Lai C. M., and Huang H. P. *A gene selection algorithm using simplified swarm optimization with multi-filter ensemble technique*. Applied Soft Computing, 106994, 2020.
[6] Maniruzzaman M, et al. *Statistical characterization and classification of colon microarray gene expression data using multiple machine learning paradigms*. Comput Methods Prog Biomed;176:173–93, 2019.
[7] P.Baldi and S.Brunak,"Bioinformatics: The Machine Learning Approach", MIT Press(2nd ed) ,2001.
[8] Othman M.S., Kumaran S. R., and Yusuf L.M. *Gene Selection Using Hybrid Multi-Objective Cuckoo Search Algorithm with Evolutionary Operators for Cancer Microarray Data*. IEEE Access, 8, 186348-186361, 2020.
[9] Zhang X., He T., Ouyang L., Xu X., and Chen S. *A Survey of Gene Selection and Classification Techniques Based on Cancer Microarray Data Analysis*. IEEE 4th International Conference on Computer and Communications (ICCC) (pp. 1809-1813) IEEE, 2018.
[10] Hatim Z Almarzouki. *Deep-Learning-Based Cancer Profiles Classification Using Gene Expression Data Profile*. Journal of Healthcare Engineering, Article ID 4715998, 13 pages, https://doi.org/10.1155/2022/4715998, 2022.
[11] Hedjazi.L, Marie-Veronique Le Lann, Tatiana Kempowsky, Florence Dalenc, Joseph Aguilar-Martin, and Gilles Favre. *Symbolic Data Analysis to Defy Low Signal-to-Noise Ratio in Microarray Data for Breast Cancer Prognosis* J Comput Biol. 20(8): 610–620. 2013.
[12]Hedjazi L. Aguilar-Martin J. Le Lann M.-V., et al. *Similarity-margin based feature selection for symbolic interval data*. Pattern Recognit. Lett. 32:578–585. 2011.
[13] Jeremy D. Scheff,1 Richard R. Almon,2,,3 Debra C. DuBois,2,,3 William J. Jusko,3 and Ioannis P. Androulakis *A New Symbolic Representation for the Identification of Informative Genes in Replicated Microarray Experiments* OMICS : A JOURNAL OF INTEGRATIVE BIOLOGY Jun 2010; 14(3): 239–248. 2010.

[14]Androulakis, I.P. Yang E. Almon, R.R. *Analysis of time-series gene expression data: methods, challenges, and opportunities*. Annu Rev Biomed Eng., 9:205–228, 2007.
[15] Lin J. Keogh E. Wei L. Lonardi S. *Experiencing SAX: a novel symbolic representation of time series*. Data Mining Knowledge Discov. 15:107–144, 2007.
[16] E. Maguire T. Yarmush M.L. Berthiaume F. Androulakis I.P. Bioinformatics analysis of the early inflammatory response in a rat thermal injury model. BMC Bioinformatics. 2007;8:10
[17] T.Nguyen *et al*.,"Hierarchical Gene Selection and Genetic Fuzzy System for Cancer Microarray Data Classification", PLoS ONE 10(3):e0120364, 2015.
[18]E.Fix and J.L.Hodges,"Discriminatory Analysis-Nonparametric Discrimination :Consistency Properties. Techical Report, 21- 49-004. Report no.4 US Air Force School of Aviation Medicine, Randolph Field, 261-279, 1951.
[19] C.Cortes and V.Vapnik, "Support Vector Networks", Machine Learning, 1995; 20:3: 273-297, 1995.
[20]Hedjazi L. Aguilar-Martin J. Le Lann M.-V., et al. *Similarity-margin based feature selection for symbolic interval data*. Pattern Recognit. Lett. 32:578–585. 2011.
[21] U.Alon *et al*.,"Multiclass cancer diagnosis using tumor gene expression signatures", PNAS 96: 6745–6750, 1999.
[22] D.Singh *et al*., "Gene expression correlates of clinical prostate cancer behaviour". Cancer Cell 1: 203–209. doi: 10.1016/S1535-6108(02)00030-2, 2002.
[23]Kent Ridge Bio-medical Data Set Repository http://datam.i2r.a-star.edu.sg/datasets/krbd/index.html.
[24]Piyushkumar.A.Mundra and Jagath.C.Rajapakse, "Gene and sample selection for cancer classification with support vectors based t-statistic", Neurocomputing, 73(2010) 2353-2362, 2010.
[25] Pengi Yang *et al*., " A multi-filter enhanced genetic ensemble system for gene selection and sample classification of microarray data", BMC Bioinformatics, 11(Suppl 1):S5 doi: 10.1186/1471-2105-11-S1-S5, 2010.
[26] Li Shengqiao Li *et al*., " Random KNN feature selection – a fast and stable alternative to Random Forests", BMC Bioinformatics, 2011, 12-450.
[27]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, 2011
[28] B.Chandra and Manish Gupta," An efficient statistical feature selection approach for classification of gene expression data", Journal of Biomedical Informatics 44 ;529–535, 2011.
[29] Xiaosheng Wang, "Robust two-gene classifiers for cancer prediction", Genomics 99 (2012) 90–95, 2012.
[30] JianzhongWang *et al*., "An Improved Feature Selection Based on Effective Range for Classification", Hindawi Publishin0067Corporation , The Scientific World Journal, Article ID 972125, 8 pages ,2014.
[31] Dajun Du *et al*.,"A novel forward gene selection algorithm for microarray data", Neurocomputing133;446–458, 2014.