DOI: http://dx.doi.org/10.26483/ijarcs.v13i5.6914

Volume 13, No. 5, September-October 2022

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

EXPERIMENTAL STUDY ON PERFORMANCE OF SYMBOLIC CLASSIFIER WITH GENE SELECTION METHODS FOR MULTICLASS MICROARRAY GENE EXPRESSION DATA

Dr. Sheela T
Associate Professor of Computer Science
Maharani’s Science College for Women (Autonomous)
JLB Road, Mysore-6
Karnataka, INDIA

Abstract: Microarray is a useful technique for measuring expression data of thousands of genes simultaneously. The expression level of genes is known to contain the keys to address fundamental problems relating to the prevention and cure of diseases, biological evolution mechanisms and drug discovery. Previous research has demonstrated that this technology can be useful in the classification of cancers. Most proposed cancer classification methods work well only on binary class problems and not extensible to multi-class problems. This work is an attempt to classify high dimensional, multiclass Microarray Gene expression data using symbolic classifier.

Keywords: Microarray Gene Expression; cancer classification; binary class data; high dimensional data; multiclass data; symbolic classifier.

I. INTRODUCTION

Microarray technology has become an essential tool in functional genomics for monitoring the expression of many genes in parallel. The process of extracting the required knowledge from the microarray gene data remains an open challenge. In order to retrieve the required information, gene classification is vital. However, the task is complex because Gene expression Microarray data is usually of very high dimensions and a small number of samples [1, 2, 9]. This makes it very difficult for many existing classification algorithms to analyze this type of data. In addition, Gene expression Microarray data contain a high level of noise, irrelevant and redundant data. All these attribute to unreliable and low accuracy analysis results.

Some researchers proposed to do gene selection prior to cancer classification. Performing gene selection helps to reduce data size thus improving the running time. More importantly, gene selection removes a large number of irrelevant genes which improves the classification accuracy [3, 7, 17, 21]. Feature selection also helps biologists to focus on the selected genes to further validate their biological hypotheses [11]. Due to the important role it plays in cancer classification, we also study the effect of classical gene selection methods with symbolic classifier in this paper.

In the context of pattern recognition, genes are usually treated as features and the gene selection problem can be solved as a feature selection problem. Generally, the feature selection methods can be classified into three categories: the filter, the wrapper and the embedded methods [9, 10, 12, 23, 25]. The filter method employs intrinsic properties of a feature without considering its interaction with other features, and the selection procedure is independent of the classifier. While in the wrapper method, a classifier is usually built and employed as the evaluation criterion. If the criterion is derived from the intrinsic properties of the classifier, the corresponding feature selection method is named as the embedded method. In embedded algorithms [15, 20, 26], feature selection occurs by the internal mechanisms of the classification algorithm. Embedded approaches are said to solve at the same time feature selection and classification. In the first stage the dimensionality is reduced using a feature selection technique embedded within a classification model and in the second stage, a standard classification technique is applied to the resulting set of features. The selection step is followed by a predictive model learning step [8, 26, 27, 29, 30].

Microarray gene expression data is high dimensional, low sample-sized data. Many of the classification methods tend to give a poor classification result for this type of data. The main cause of this is noise occurring from irrelevant and redundant variables (dimensions). Therefore, there is a need to reduce or summarize variables. An interesting approach would be to use symbolic data analysis (SDA) popularized by [4, 5, 10]. Within this framework, interval data representation can be used to take into account the uncertainty and noise inherent in measurements [13]. Interval-valued data representation transforms the original data into a more manageable data in order to avoid the curse of dimensionality. 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.

II. RELATED WORK

The work in [14] 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 [13] 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 [16] 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 [3] 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) [19] has been applied to gene expression data through SLINGSHOTS in [22], which selects informative genes from gene expression data based on the symbolic representation.

III. METHODOLOGY

Our study focuses on the performance of symbolic classifier on the high dimensional, multi-class gene expression data. Three classical feature selection algorithms are also evaluated using symbolic classifier as the learning algorithm.

A. 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, 16]. 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,…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} , … 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 [g_{jk}^- , g_{jk}^+]

Where $g_{j,k}^+$ and $g_{j,k}^-$ 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} , … g_{nk}) and $g_{j,k}^-$ = min ((g_{1k} , g_{2k} , g_{3k} , … g_{nk})

Hence, the interval [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} , … g_{in}] in the form of an interval and is given by R_j = {[g_{j1}^- , g_{j1}^+] , [g_{j2}^- , g_{j2}^+] ……. [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.

B. 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

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

Where ,

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

IV. FEATURE SUBSET SELECTION ALGORITHMS

The goal of the feature selection is to select the smallest subset of features but carrying as much information about the class as possible. These methods return a subset of features based on an intrinsic determination of the feature set size. We used the well-known wrapper method-Sequential Forward Selection, a classical filter method-minimal Redundancy-Maximal-Relevance, and a popular embedded approach-Random Subset Feature Selection (RSFS).

A. Sequential Forward Selection (SFS)

Sequential Forward Selection [28] was chosen as the baseline method for feature selection, as it is well known and widely used in practice. And is regarded as one of the state-of-the-art feature selection algorithms.

The sequential forward scheme starts from an empty set, and sequentially includes a new feature into the feature subset so that the largest improvement on the evaluation criterion can be achieved. Once a feature is selected, it will not be removed from the subset. This wrapper method has been successfully used for gene selection in microarray gene expression data [20, 15].

B. Minimal Redundancy Maximum Relevance mRMR Feature Subset Selection

Minimal-redundancy-maximal-relevance (mRMR) is a filter-based feature selection approach proposed by [9]. It analyzes the mutual information between discretized features and class labels to maximize the feature relevance while simultaneously considering the mutual information among the discretized features in the selected feature set to minimize redundancy. The minimum redundancy maximum relevance (mRMR) criterion [24] computed both the redundancy between features and the relevance of each feature. Redundancy is computed by the mutual information between pairs of features, whereas relevance is measured by the mutual information between each feature and the class labels. The mRMR method has been used for gene selection [9].

C. Random Subset Feature Selection (RSFS)

The RSFS is based on the idea of Random Forests [6] a popular feature selection method for “small n, large p” problems and Random kNN [18] which is specially designed for classification of high dimensional datasets.

Random Subset Feature Selection (RSFS) is an embedded feature selection algorithm that aims to discover a set of features that perform better than an average feature of the available feature set. The set of "good" features is obtained by repetitively choosing a random subset of features from the set of all possible features and then classifying the data with a kNN classifier using these features.

V. EXPERIMENTATION

The experiments are conducted on 7 different multiclass microarray cancer datasets. Their characteristics are listed in Table I. These datasets are often used in the field of cancer classification problems from microarray data, which are available from http://www.gems-systems.org for non-commercial use. The 7 datasets are representative which have 2–11 classes, 50–203 samples and 2308–12600 genes after the data preparatory steps and are linearly normalized.

In the proposed work, we evaluated the performance accuracies of symbolic classifier, on the multi-class datasets. The experiment has been conducted on the seven multi-class datasets under varying training samples like 80%, 60% and 40%. It is observed that the classification accuracy improved when 80% of the data is used for training and 20% for testing. Table II shows the classification accuracy of three gene selection methods with different training samples.

Table I. Multiclass Gene Expression Datasets

Dataset name	*Number of samples*	*Number of genes*	*Number of classes*
D1: 11_Tumor	174	12533	11
D2: 9_tumor	60	6167	9
D3: Brain_Tumor1	90	5920	5
D4: Lung cancer	203	12600	5
D5: Brain_Tumor2	50	10367	4
D6: SRBCT	83	2308	4
D7: DLBCL	77	5469	2

Gene subset selection may in some cases improve the performance of the classifier since gene selection is not only concerned with reducing the number of genes but also eliminating the variables that produce noise or, are correlated with other already selected variables. To demonstrate the method, the entire sets of genes were used in predicting the output and for comparison purpose, we experimented using three classical gene selection methods. The result shows that the accuracy of microarray data classification which had feature selection implemented was better than without feature selection. This is shown in Table III. The differences are clearly shown in Figure 1. Classification accuracy can further be improved by using appropriate gene selection methods before classification.

Table II. Classification accuracy using symbolic classifier for the subsets given by the three methods (SFS, mRMR and RSFS) with varying training samples.

Dataset name	*SFS*			*mRMR*			*RSFS*		
	40%	*60%*	*80%*	*40%*	*60%*	*80%*	*40%*	*60%*	*80%*
D1: 11_Tumor	65.66	70	76.89	72.62	72.8	75.79	73	73	80.47
D2: 9_tumor	67.84	71.25	77.98	76.46	81.46	98.17	71.48	74.32	79.56
D3: Brain_Tumor1	66.67	69.78	74.19	77.8	77.8	80.5	69.4	72.22	75.66
D4: Lung cancer	50	71.60	77.56	73.78	75	75.31	74.56	79.54	81.45
D5: Brain_Tumor2	42.86	59.55	70	72.41	80.95	90	52.38	58.62	70
D6: SRBCT	65	70.35	78.12	75.9	78	79.66	81.45	81.45	84.31
D7: DLBCL	74.22	79.44	81.25	87.78	90.32	93.75	75	90.32	86.96

Table III. Accuracy with and without gene selection methods and time complexity of proposed classifier

Dataset name	*Accuracy of Symbolic classifier to the original datasets without any gene selection and time taken*		*Accuracy of Symbolic classifier after gene selection.*		
	Accuracy	*Time Taken (Minutes)*	*SFS+ Symbolic*	*mRMR+ Symbolic*	*RSFS+ Symbolic*
D1: 11_Tumor	74.79	0.0543	76.89	75.79	80.47
D2:9- Tumor	70	0.0340	77.98	98.17	79.56

D3:Brain Tumor 1	73.14	0.0280	74.19	80.5	75.66
D4: Lung Cancer	77.44	0.0585	77.56	75.31	81.45
D5:Brain Tumor 2	76.45	0.0491	70	90	70
D6: SRBCT	76.81	0.0199	78.12	79.66	84.31
D7: DLBCL	79.66	0.0287	81.25	93.75	86.96

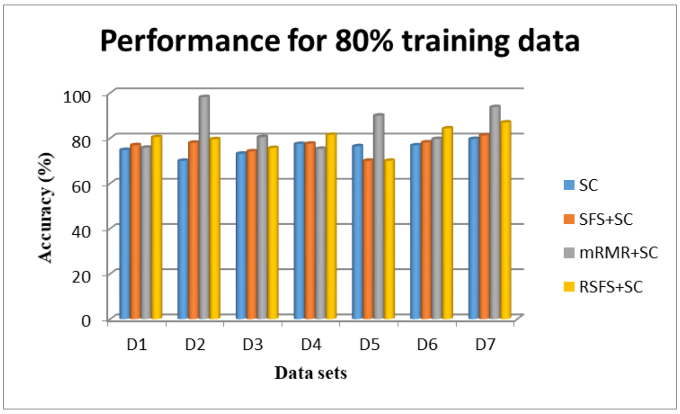

Figure 1. Comparison of symbolic classifier(SC) with and without gene selection methods.

VI. Conclusion

In gene expression microarray data, the capability of selecting few numbers of predictive and important genes, not only makes the data analysis efficient but also helps in their biological interpretation and understanding of the data. Feature selection reduces the number of features, removes irrelevant, noisy and redundant data, and results in acceptable classification accuracy.

In our work, we performed experiments with three classical gene selection methods, a filter, a wrapper and an embedded algorithm for feature selection in multiclass microarray data sets. Symbolic classifier served as an evaluator for the gene selection methods. Experimental results show that Interval valued symbolic classifier helped achieve good accuracy with less time complexity. Classification performance is enhanced due to removal of noisy and unreliable genes.

VII. References

[1] Ahmed, O., and Brifcani, A. (2019, April). Gene Expression Classification Based on Deep Learning. 4th Scientific International Conference Najaf (SICN) pp. 145-149, 2019.

[2] Alomari, O.A., Khader, A.T., Al-Betar, M.A., Abualigah L.M. MRMR BA: a hybrid gene selection algorithm for cancer classification. J Theor Appl Inf Technol , 95 (12):2610–8, 2017.

[3]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.

[4] Billard, L. and Diday, E.. Symbolic data analysis:Conceptual statistics and data mining. Wiley series in computational statistics. 2006.

[5] Bock, H.H and Diday, E. Analysis of Symbolic Data. Springer Verlag, 1999.

[6] Breiman L: Random Forests. Machine Learning. 45:5-32, 2001.

[7] Cahyaningrum, K., and Astuti, W. Microarray Gene Expression Classification for Cancer Detection using Artificial Neural Networks and Genetic Algorithm Hybrid Intelligence. International Conference on Data Science and Its Applications (ICoDSA) (pp. 1-7). IEEE, 2020.

[8] Christoph Bartenhagen, Hans-Ulrich Klein, Christian Ruckert, Xiaoyi Jiang and Martin Dugas. Comparative study of unsupervised dimension reduction techniques for the visualization of microarray gene expression data, BMC Bioinformatics, 11:567, 2010.

[9] Ding, C., Peng, H. Minimum redundancy feature selection from microarray gene expression data. In:Journal Bioinformatics and Computer Biology, pp.523-529, 2003.

[10] Diday. An introduction to symbolic data analysis and sodas software. Electro. J.Symb. Data Anal. 1-25, 2002.

[11] Golub T.R., Slonim D.K. and Tamayo. Classification of Cancer: Class discovery and Class Prediction by Gene Expression Monitoring. Science. 286, 315-333, 1999.

[12] 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.

[13]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.

[14] 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.

[15] Inza I., Larrañaga P., Blanco R., Cerrolaza A.J. Filter versus wrapper gene selection approaches in DNA microarray domains, Artif Intell Med, 31(2):91-103, 2002.

[16] 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.

[17] 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.

[18] Li S. Random KNN Modeling and Variable Selection for High Dimensional Data. PhD thesis. West Virginia University, 2009.

[19] 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.

[20] Liu Q, et al. Gene selection and classification for cancer microarray data based on machine learning and similarity measures. BMC Genomics 12(Suppl 5):S1, 2011.

[21] 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.

[22] 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

[23] 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.

[24] Peng,H.,Long,F.,Ding,C. Feature selection based on mutual information : Criteria of max-dependency, max-relevance and min-redundancy. IEEE Trans Pattern Anal Machine Intell. 27(8), 1226-1238. 2005.

[25] T.Ragunthar, S.Selvakumar. Classification of Gene Expression Data with Optimized Feature Selection. International Journal of Recent Technology and Engineering (IJRTE). ISSN: 2277-3878, Volume-8 Issue-2, July2019.

[26]Saeys Y., Inza I., Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics, 23(19), 2007.

[27] Statnikov A., Aliferis C.F., Tsamardinos I., Hardin D., Levy, S. A Comprehensive Evaluation of Multicategory Classification Methods for Microarray Gene Expression Cancer Diagnosis. Bioinformatics 21(5), 631–643, 2005.

[28] Whitney A.W. A Direct Method of Nonparametric Measurement Selection. IEEE Trans. Comput., 20, 1100–1103. doi: 10.1109/T-C.1971.223410. 1971.

[29] Xing E., Jordan M., Karp R. Feature selection for high-dimensional genomic microarray data. Proceedings of the 18th International Conference on Machine Learning, 2001.

[30] 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.