International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT2390131

Multilevel Feature Selection Method for Improving Classification of Microarray Gene Expression Data

Dr. Sheela T.*, Prakasha Raje Urs M., Santhosh Kumar B. N.
Department of Computer Science, Maharani's Science College for Women, Mysore, Karnataka, India

Article Info

Publication Issue :
Volume 9, Issue 1
January-February-2023

Page Number : 176-183

Article History
Accepted: 01 Feb 2023
Published: 11 Feb 2023

ABSTRACT

Microarray gene expression profiles provide valuable answers to a variety of problems, and contributes to advances in clinical medicine. Gene expression data typically has a high dimension and a small sample size. Gene selection from microarray gene expression data is a challenge due to high dimensionality of the data. The number of samples in the microarray dataset is much smaller compared to the number of genes as features. To extract useful gene information from cancer microarray data and reduce dimensionality, selection of significant genes is necessary. An effective method of gene feature selection helps in dimensionality reduction and improves the classification performance. Experimental results suggest that appropriate combination of filter gene selection methods is more effective than individual techniques for microarray data classification. In this paper, we propose a two-layered feature selection method. In the first layer, t-test statistical method is used to remove the features that have little correlation with the classification results. In the second layer, line segment approximation method is used to transform the feature subset into a less dimensional feature space. Four well known classifiers kNN, SVM, NBC, DT were used to verify the performance of the proposed feature selection algorithm on binary class microarray data. The experimental results show that the proposed method can effectively select relevant gene subsets, and achieves higher classification accuracy.

Keywords: Microarray gene expression data, significant genes, feature selection, line segment approximation

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 datasets are usually of very high dimensions and a small number of samples [1, 2, 3]. This makes it very difficult for many existing classification algorithms to analyse 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 [4, 5, 6]. Feature selection also helps biologists to focus on the selected genes to further validate their biological hypotheses [7].

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 [8, 9, 10]. 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 [11, 12, 13], 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 [14, 15, 16, 17].

Ensemble techniques represent a relatively new class of methods for feature selection. The idea of ensemble in biological data analysis originates from combining multiple classifiers for improving sample classification accuracy. Similar to the technique of ensemble classification [18], ensemble method is adopted for feature selection [19] to increase the stability of feature selection algorithms. It is a method of aggregating outputs of more than one feature selector. The underlying principle behind these methods is that genes that are considered significant by different measures may have genuine biological relevance compared to genes selected by a single measure [20]. It has been adopted and increasingly used in feature selection of high-dimensional data [21]. Ensemble techniques might be used to improve the robustness of feature selection techniques.

The proposed multilevel feature selection method is an ensemble technique, consisting of two main levels. In the first level, the statistical feature selection method t-test which selects the discriminative features based on importance of ranking and features correlations is used. The second level is the feature approximation method, which reduces the size of the feature space by the proposed line segment approximation method. Then the reduced feature set is given to classifier for sample classification.

II. LITERATURE REVIEW

The ensemble idea in supervised learning has been investigated since the late seventies. [22]suggests combining two linear regression models. Where, the first linear regression model is fitted to the original data and the second linear model to the residuals. There are two categories of ensemble techniques multi-expert systems and multistage systems. Multiexpert system is one in which different classifiers work in parallel and each one will give its own decision and final decision is made using a combiner. AdaBoost (Adaptive Boosting), a multi-expert system, was first introduced in [23], is a popular ensemble algorithm that improves the simple boosting algorithm via an iterative process. Some other examples for multiexpert systems are voting [24] mixture of experts [25] and stacked generalization

[26]. Multistage system uses serial approach and is called cascading system. In this system the next classifier is consulted for sample classification only when the previous classifier is not confident on its decision and rejects the sample. By designing a multistage ensemble system with small number of classifiers, we can get good accuracy with less computation cost, memory usage and time. Cascading uses the benefits of multistage properties and does not consult all classifiers on all instances and thus reduces classification time. Cascading, the multistage method of information fusion is discussed in [27, 28]. Where, a sequence of classifiers are ordered under some conditions and the next classifier is only considered for patterns refused by the previous classifiers. The advantage of a cascading system is that, an earlier classifier handles major cases and a complex classifier in the next stage is only utilized with a small possibility hence not increasing the complexity greatly. A two stage classifier architecture for text document classification is used in [29] to automatically handle rejections, where, documents can be either classified or rejected at the first stage and the rejected documents are automatically classified at the second stage. A two-stage cascading scheme for iris recognition is presented by [30] and results prove that the cascading classification system outperforms single classifier. A registration method of cascading two fingerprint registration schemes is designed by [31] and a series of experiments validate the effectiveness of the multi-stage strategy. In [32], a decision boundary for the binary class datasets is obtained using the statistical method of confidence interval.

III. PROPOSED MODEL

Figure 1 shows the stages in our proposed model.

A. Data normalization

Before performing a data mining activity it is required to prepare or represent the data in a form which will be appropriate for mining. Normally the feature value ranges widely in different databases. Thus training the model with these values may mislead the model to a different direction [33]. Due to this reason features are scaled to a specified range and the process of scaling is termed as data normalization. In this work min-max normalization is used and the feature values are scaled within 0 to 1.

C. Proposed line segment approximation method

The top ranked genes are now summarized into lesser number of features by our proposed line segment approximation method as explained below:

Successive feature values f_i, f_{i+1}, f_{i+2} … f_{i+m} are assimilated into a line segment, if the points (i, f_i), (i+1, f_{i+1}), . . . (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.

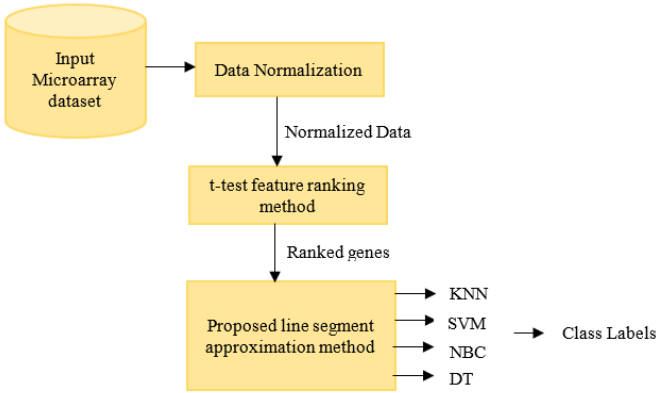


Figure 1: Proposed framework for feature subset selection and classification

B. T-test Ranking

We have used student t-test statistical method to identify significant genes in each group [34, 35]. 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. A t score of each gene is calculated using equation (1).

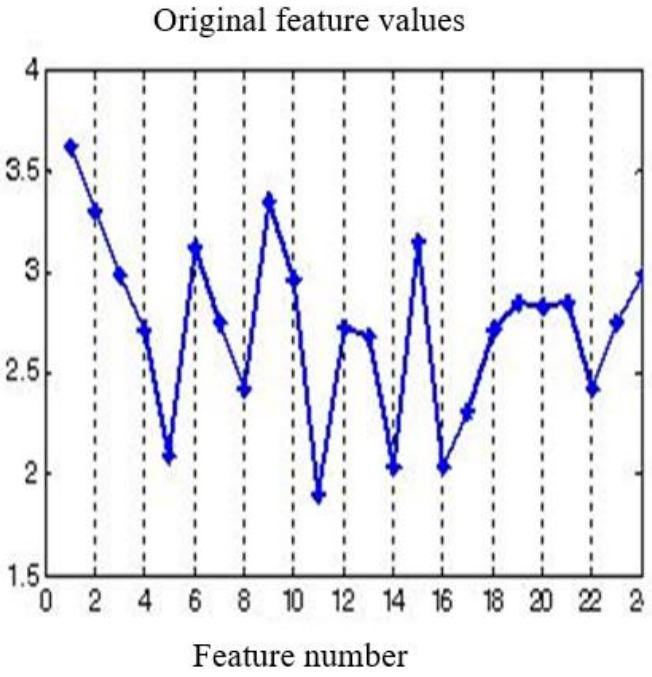
$$t = \frac{(\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)}} \quad (1)$$

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.

That is, if $|(f_{i+1} - f_i) - (f_{i+2} - f_{i+1})| \leq \varepsilon$. 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. Each feature set is represented by a feature descriptor of the form V= (v1,v2, v3,v4), where v1 is the number of line segments, v2 is the average slope of these line segments, v3 is the first feature of the sample and v4 is the last feature of the sample. 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 v3 and v4. Thus, the number of features is reduced to 4. The number of line segments
and the average slope will be same for samples from the same class. Instead of computing individual expression values, we need comparison of only 4 feature values after one transformation to line segment. The process is repeated for all top ranked genes. The number of features in each sample for final classification is 4. Example of line segments generated for feature values of some sample in leukemia dataset is shown in Figure 2. From Figure 2, it is evident that, in the range of feature number 0 to 4, four features are collinear and are approximated to two features. Similarly, seven features in the range of feature number 16 to 22 are approximated to four features.

We used the following parameters for the classifiers:
In k-Nearest Neighbors (kNN) classification algorithm[36] 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[37], a linear kernel is used. For Decision trees (DT) [38] method, 10 is the minimum parent size and 1 is the minimum leaf size. In the Naïve Bayes classification(NBC) algorithm [39] a kernel distribution was used with predictors.



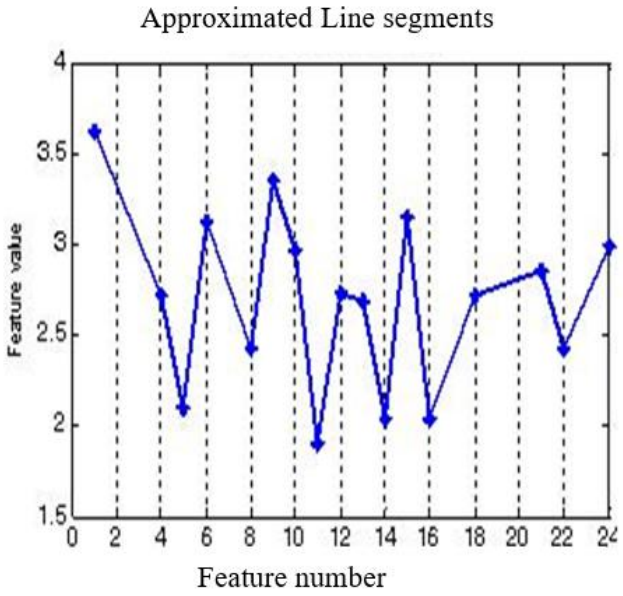


Figure 2 : Approximation of line segments

IV. 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 I

Description of the binary class microarray Datasets

Dataset	Genes(m)	Samples(N)
Leukemia	7129	72
Prostate	6033	102
CNS	7129	34
DLBCL	5470	77
Ovarian	4000	216
Breast	4348	97

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 [40].

V. RESULTS AND DISCUSSION

Results of the experiments conducted on six binary class microarray datasets is given in this section. Table II below, shows the classification accuracy for the datasets with various classifiers. The first column has the result of classification accuracy for the original data without feature selection and the next column shows the accuracy obtained with the proposed the multilevel feature selection method. From table II it is evident that compared to the original data, accuracy is increased by a maximum of 8% after applying the proposed line segment approximation method.

VI. CONCLUSION

This paper presents a multilevel gene selection algorithm to effectively classify the binary class microarray gene expression data. When the results are compared with those of similar algorithms, the proposed method yielded a higher level of accuracy. The time and
storage cost of the algorithm is very appealing, making it optimal for big data. The method can further be improved by using efficient gene selection method.

TABLE II. Accuracy obtained for the original data and for the reduced data

Dataset	Classifier	Accuracy (%)	
		Original data	t-test + proposed method
Leukemia	kNN	91.67	98.61
	SVM	95.83	98.61
	NBC	87.5	91.67
	DT	94.44	98.61
Prostate	kNN	86.27	96.1
	SVM	92.16	97.4
	NBC	83.33	97.4
	DT	85.29	89.61
DLBCL	kNN	92.16	96.08
	SVM	88.24	95.10
	NBC	82.45	88.24
	DT	82.35	86.35
Ovarian	kNN	85.65	89.35

	SVM	91.2	98.15
	NBC	82.29	83.33
	DT	80.9	85.19
Breast Cancer	kNN	93.44	95.88
	SVM	93.81	96.91
	NBC	89.22	94.85
	DT	85.29	96.91
CNS	kNN	88.24	95.12
	SVM	85.29	94.12
	NBC	85.29	88.24
	DT	80.15	88.24

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

[4]. 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.

[5]. 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.

[6]. 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.

[7]. 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.

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

[9]. 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.

[10]. 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.

[11]. 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.

[12]. 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.

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

[14]. 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.

[15]. 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.

[16]. Xing E., Jordan M., Karp R. Feature selection for high-dimensional genomic microarray data.

Proceedings of the 18th International Conference on Machine Learning, 2001.

[17]. 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.

[18]. Dietterich TG2000 Dietterich TG. Ensemble methods in machine learning. In: Proceedings of Multiple Classifier System.vol. 1857.Springer; 2000. pp. 1–15.

[19]. Saeys Y, Thomas Abeel, Yves Van de Peer. Robust feature selection using ensemble feature selection techniques. In Proceedings of the 25th European Conference on Machine Learning and Knowledge Discovery in Databases, Part II, Springer-Verlag, Berlin, Heidelberg, pp. 313–325 (2008).

[20]. Y.H., Xiao Y., Segal M.R. :Identifying differentially expressed genes from microarray experiments via statistic synthesis. Bioinformatics. 21(7):1084–1093 (2005)

[21]. 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.

[22]. JW (1977) Exploratory data analysis. Addison-wesley series in behavioral science, First Edition.

[23]. Y, Schapire RE (1996) Experiments with a new boosting algorithm. In: Machine Learning: Proceedings of the 13th international conference , pp325-332

[24]. 1998 Kittler, J., Hatef, M. Duin, R. P. W., & Matas, J. (1998). On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20, 226-239.

[25]. 1991 Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991) Adaptive mixtures of local experts. Neural Computation, 3, 79-87.

[26]. D. H. (1992) Stacked generalization. Neural Networks, 5, 241-259.

[27]. P 1992 P. Pudil, J. Novovicova, S.Blaha and J. Kittler. Multistage Pattern Recognition with Rejection Option. Proceedings of the 11th International Conference on Pattern Recognition, Vol.B, pp. 92 - 95, 1992.

[28]. 2000 C. Kaynak and E. Alpaydin. MultiStage Cascading of Multiple Classifiers: One Man's Noise is Another Man's Data. Proc. 17th International Conf. on Machine Learning, 2000.

[29]. G., Pillai, I., & Roli, F. (2004). A Two-Stage Classifier with Reject Option for Text Categorisation. In

[30]. Structural, Syntactic, and Statistical Pattern Recognition (pp. 771–779). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-27868-9_84.

[31]. Sun 2004] Zhenan Sun, Yunhong Wang, Tieniu Tan and Jiali Cui. Cascading Statistical And Structural Classifiers For Iris Recognition. Proceedings of IEEE International Conference on Image Processing, 2004, pp.1261 - 1264.

[32]. Qi, Zhongchao shi, Xuying Zhao and Yangsheng Wang. Cascading a Couple of Registration Methods for a High Accurate Fingerprint Verification System. Proceedings of Sinobiometrics'04, LNCS 3338, Beijing, China, Dec. 2004

[33]. and Dr.Lalitha Rangarajan. An Approach to reduce the large feature space of Microarray Gene Expression data by Gene Clustering for efficient sample classification. International Journal of Computer Applications, Issue 8, Volume 2, March-April 2018. (UGC No: 64190, ISSN : 2250 1797)

[34]. Dash, Rasmita, Misra, Bijan Biahri , 2016. Pipelining the ranking techniques for microarray data classification: a case study. Appl.soft Comput, 48, 298-316.

[35]. Rajani Bala, 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.

[36]. Nguyen T, Khosravi A,Creighton D, Nahavandi S. Hierarchical Gene Selection and Genetic Fuzzy System for Cancer Microarray Data Classification. PLoS ONE, 2015, 10(3):e0120364.

[37]. J H, Bentley J L, Finkel R A. An Algorithm for Finding Best Matches in Logarithmic Expected Time. ACM Trans.Math.Softw., 1977, 3(3):209–226.

[38]. Cortes C, Vapnik V. Support-Vector Networks. Mach Learning, 1995, 20(3):273–297.

[39]. Quinlan J R. Simplifying decision trees. International Journal of Human-Computer Studies,1999, 51 (2):497.

[40]. G H, Langley P. 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.

[41]. A multi-task machine learning software. http://www.cs.waikato.ac.nz/ml/weka.

Cite this article as :

Dr. Sheela T., Prakasha Raje Urs M., Santhosh Kumar B. N., "Multilevel Feature Selection Method for Improving Classification of Microarray Gene Expression Data", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 1, pp.176-183, January-February-2023. Available at doi : https://doi.org/10.32628/CSEIT2390131
Journal URL : https://ijsrcseit.com/CSEIT2390131