J. Bio-Sci. 28: 69-80, 2020 ISSN 1023-8654
http://www.banglajol.info/index.php/JBS/index
DOI: https://doi.org/10.3329/jbs.v28i0.44712

PERFORMANCE EVALUATION OF DIFFERENT MACHINE LEARNING ALGORITHMS IN PRESENCE OF OUTLIERS USING GENE EXPRESSION DATA

M Shahjaman[1*], MM Rashid[1], MI Asifuzzaman[1], H Akter[1], SMS Islam[3] and MNH Mollah[2]

[1]*Department of Statistics, Begum Rokeya University, Rangpur, Bangladesh*
[2]*Bioinformatics Lab., Department of Statistics, University of Rajshahi, Bangladesh*
[3]*Institutitute of Biological Sciences, University of Rajshahi, Bangladesh*

Abstract

Classification of samples into one or more populations is one of the main objectives of gene expression data (GED) analysis. Many machine learning algorithms were employed in several studies to perform this task. However, these studies did not consider the outliers problem. GEDs are often contaminated by outliers due to several steps involve in the data generating process from hybridization of DNA samples to image analysis. Most of the algorithms produce higher false positives and lower accuracies in presence of outliers, particularly for lower number of replicates in the biological conditions. Therefore, in this paper, a comprehensive study has been carried out among five popular machine learning algorithms (SVM, RF, Naïve Bayes, k-NN and LDA) using both simulated and real gene expression datasets, in absence and presence of outliers. Three different rates of outliers (5%, 10% and 50%) and six performance indices (TPR, FPR, TNR, FNR, FDR and AUC) were considered to investigate the performance of five machine learning algorithms. Both simulated and real GED analysis results revealed that SVM produced comparatively better performance than the other four algorithms (RF, Naïve Bayes, k-NN and LDA) for both small-and-large sample sizes.

Key words: Classification, DE gene, GED, Outliers, Robustness

Introduction

DNA microarrays are the tools used to measure the expression levels of large numbers of genes with small-number of biological replicates, simultaneously. A typical microarray experiments generates matrix (X) of $p \times n$ dimension of expression levels, where the rows (p) represent genes and columns (n) represent samples.

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1} & x_{p2} & \cdots & x_{pn} \end{bmatrix}$$

Usually the rows (p) contain 1000 to 100000 genes and columns contain (n) 2 to 100 samples. As a result for analyzing of this kind of datasets researchers often suffer from *curse of dimensionality*. *Curse of dimensionality* means "large p, small n" problem. Classification is a supervised machine learning

*Author for correspondence: shahjaman_brur@yahoo.com

approach, which assimilate the knowledge of sample class label information into the analysis. Classification of samples into one or more populations based on gene expression data (GED) has already become attracted to the research communities. The goal of classification is to classify the new sample into one of two or more population based on label datasets. Label datasets whose categories are known in advance is used to select a subset of expressed genes which have the most discriminative power between the classes to be predicted and construct a model, called classifier. In cancer research, DNA microarray technology has been extensively used, which enables classification of tissue samples based only on GED without prior knowledge (Golub et al. 1999, Dudoitet al. 2002). In earliest days, DNA microarray technology was used to identify cancer patient, in cancer research and many studies. Conventional methods for cancer diagnosis are invasive, subjective, and labor intensive. Recently, with advances in microarray technologies, gene expressions of biological samples from patients have been proven promising and feasible as a non-invasive, objective and accurate molecular diagnostic method in oncology. Cancer classification based on gene expression dataset is important for subsequent diagnosis and treatment. A number of classification methods available for microarray gene expression data analysis. Among them Fisher's linear discriminant analysis (LDA) (Fisher 1936, Carles 1989) is the oldest and popular. There are two variants of discriminant analysis: One is Diagonal Linear Discriminant Analysis (DLDA) which provides optimal discrimination when class densities have the same diagonal variance-covariance matrix. Other is the weighted voting algorithm introduced by Golub et al. (1999) which has become relatively popular and comes out to be a variant of DLDA. K Nearest Neighbors (k-NN) also has been used widely in GED analysis, probably due to its simplicity. It searches the *k* closest features in the training set and assigns to the class that appears most frequently. Other classification methods that are applied in gene expression data: logistic regression (Hosmer and Lemeshow 1989), decision tree (DT) (Fayyad and Irani 1992), generalized partial least squares (GPLS) (Ding and Gentleman 2003), adaBoost (Freund and Schapire 2003), logitBOOST (Dettling and Buhlmann 2003), artificial neural network (ANN) (Zuruda 1992), support vector machine (SVM) (Vapnik 1998), naive Bayes (Friedman et al. 1997), random forest (Ho 1995), two regularization methods: L1 (LASSO) and L2 (Ridge)-regularization (Hoerl et al. 1970, Tibshirani 1996) and so on. There are many comparative studies has been performed by the researcher's among these machine learning algorithms to select the appropriate algorithm. However, they did not consider the problems of outliers in their studies. GEDs are often contaminated by outliers due to several steps involved in the data generating process from hybridization of DNA samples to image analysis (Shahjaman et al. 2017a, 2017b and 2019). Therefore, in this paper, a comprehensive study has been carried out among five popular machine learning algorithms (SVM, RF, Naïve Bayes, k-NN and LDA) using both simulated and real gene expression datasets, in absence and presence of outliers.

Materials and Methods

Performance evaluation

In order to assess the performance of different machine learning classifiers for binary classification test such as Normal or Cancer, receiving operating characteristics (ROC) curve, area under the ROC curve (AUC) and all the measures associated with this curve were used. Suppose TP, TN, FP and FN denotes the number of true positive, number of true negative, number of false positive and number of false negative, respectively. Based on these parameters we calculate the following measures of performance:

True positive rate (TPR) = TP/ (TP + FN)
True negative rate (TNR) = TN/ (TN + FP)
False positive rate (FPR) = FP/ (FP + TN)
False negative rate (FNR) = FN/ (FN + TP)
False discovery rate (FDR) = FP/ (TP + FP)

A method was declared as good performer if it produces larger values of TPR, TNR, AUC and smaller values of FPR, FNR and FDR.

Data Sources

Simulated gene expression datasets

The simulated gene expression datasets were generated using Table 1. Three different patterns of gene groups were generated as described in this table. The patterns p1 and p2 represents the up-regulated and down-regulated gene groups and the pattern p3 represents the equally expressed (EE) group. The gene expression profiles of 1000 genes with normal and cancer conditions were generated from Table 1. 100 datasets were generated from this table for both small-and-large sample cases, respectively. For small-and-large sample sizes, 3 and 15 replicates were considered for normal and cancer conditions. The values of the parameter c and σ2 were set to 2 and 0.1, respectively. These 1000 genes were distributed into three patterns: p1 = 50, p2 = 50 and p3 = 900. 100 datasets were randomly divided into two independent datasets to construct the training and test dataset such that training and test datasets consist equal number of replicates in both groups.

Table 1. Simulated gene expression data generating model for two groups

Gene groups	Samples/Individuals		Gaussian noise
	n_1	n_2	
p_1	-c	c	+ N(0, σ^2)
p_2	c	-c	
p_3	c	c	

Real gene expression dataset

Discovery and cataloguing of gene expression in normal and disease conditions has been facilitated by generation of large gene expression datasets. Many of these sets have been deposited in public databases such as GEO (Barrett et al. 2013) or ArrayExpress (Kolesnikov et al. 2015). In this paper, the performance of five classifiers were evaluated as mentioned earlier in the real colon cancer dataset, which consists of 22

control and 40 colon cancer samples. Initially the gene expression profiles of this dataset were 6,500 those were analyzed with an Affymetrix technology. After selecting the highest minimal intensity across the samples, 2000 genes have been retained (Alon et al. 1999).

Results and Discussion

The performance of five popular machine learning algorithms (SVM, Naïve Bayes (NBC), LDA, K-NN and Random forest (RF)) was demonstrated for both small-and-large sample cases. Three R packages for the five machine learning algorithms were used in this study such as *MASS*, *kknn* and *e1071*. All the performance measures including AUCs were computed using R package ROC in comprehensive R archive network (cran).

Performance evaluation using simulated gene expression profiles

To investigate the performance of five popular machine learning algorithms (SVM, NBC, LDA, K-NN and RF), 100 datasets were generated from Table 1 for both small and large sample cases, respectively. To construct training and test datasets for classification performance of the five classifiers, each of 100 simulated datasets were partitioned as training and test datasets such that both datasets consist of same number of replicates in each condition. The DE genes were selected using t-test to train and boost the performance of the five classifiers. To investigate the performance of all the classifiers in presence of outliers, 10%, 30% and 50% genes were randomly corrupted by a single outlier in the training datasets. The outlying value was considered as ten times larger than the maximum value for each condition. Then the average values of different performance measures such as true positive rate (TPR), false positive rate (FPR), true negative rate (TNR), false negative rate (FNR), false discovery rate (FDR) and area under the receiving operating characteristics (ROC) curve (AUC) were computed based on 100 estimated DE genes by five methods for both small-and-large sample cases. A method is rewarded as best performer if it produces larger values of TPR, TNR, AUC and the smaller values of FPR, FNR and FDR. Table 2 and Table 3 summarize the average values of different performance measures estimated by five classifiers, for small-and-large sample cases, respectively. From Table 2 revealed that for small sample case in absence of outliers, SVM produced better performance than the other four classifiers NBC, LDA, K-NN and RF. For example, SVM produced AUC=0.84 which is larger than 0.78, 0.80, 0.78, and 0.80 for the four competitors, NBC, LDA, K-NN and RF. However, in presence of outliers, the performances of all the classifiers are deteriorated. In this case SVM, NBC and RF produced better results than the other two classifiers LDA and K-NN. SVM produced slightly better than the other methods. For large-sample case in absence of outliers all the five classifiers produce similar performance. But in presence of 10%, 30% and 50% outliers, SVM, NBC and RF produce better performance than other two classifiers LDA and K-NN. The ROC curve and boxplots are presented in Fig. 1 and Fig. 2, respectively also supported the results of Table 2.

Table 2. Performance evaluation based on simulated dataset for small-sample case

In Absence of Outliers							In Presence of 10% Outliers						
Methods	TPR	TNR	FPR	FNR	FDR	AUC	Methods	TPR	TNR	FPR	FNR	FDR	AUC
SVM	0.95	0.84	0.16	0.05	0.13	0.89	SVM	0.87	0.75	0.25	0.13	0.17	0.75
NBC	0.91	0.84	0.16	0.09	0.12	0.83	NBC	0.87	0.72	0.28	0.13	0.19	0.74
LDA	0.94	0.79	0.21	0.06	0.15	0.81	LDA	0.88	0.61	0.39	0.12	0.26	0.64
KNN	0.96	0.70	0.30	0.04	0.20	0.78	KNN	0.80	0.66	0.34	0.20	0.25	0.63
RF	0.93	0.87	0.13	0.07	0.10	0.91	RF	0.89	0.67	0.33	0.11	0.22	0.73
In Presence of 30% Outliers							In Presence of 50% Outliers						
Methods	TPR	TNR	FPR	FNR	FDR	AUC	Methods	TPR	TNR	FPR	FNR	FDR	AUC
SVM	0.87	0.70	0.30	0.13	0.21	0.72	SVM	0.66	0.81	0.19	0.34	0.15	0.64
NBC	0.86	0.64	0.36	0.14	0.22	0.65	NBC	0.91	0.50	0.50	0.09	0.31	0.56
LDA	0.91	0.50	0.50	0.09	0.32	0.62	LDA	0.86	0.44	0.56	0.14	0.37	0.51
KNN	0.83	0.54	0.46	0.17	0.34	0.57	KNN	0.84	0.39	0.61	0.16	0.37	0.44
RF	0.91	0.58	0.42	0.09	0.26	0.68	RF	0.85	0.49	0.51	0.15	0.32	0.54

Table 3. Performance evaluation based on simulated dataset for large-sample case

In Absence of Outliers							In Presence of 10% Outliers						
Methods	TPR	TNR	FPR	FNR	FDR	AUC	Methods	TPR	TNR	FPR	FNR	FDR	AUC
SVM	0.99	0.99	0.00	0.01	0.00	0.99	SVM	0.98	0.97	0.03	0.02	0.03	0.99
NBC	0.99	0.99	0.00	0.01	0.00	0.99	NBC	0.99	0.99	0.01	0.01	0.01	0.99
LDA	0.99	0.99	0.00	0.01	0.00	0.99	LDA	0.97	0.84	0.16	0.03	0.13	0.90
KNN	0.99	0.99	0.00	0.01	0.00	0.99	KNN	0.94	0.86	0.14	0.06	0.12	0.89
RF	0.99	0.99	0.01	0.01	0.01	0.99	RF	0.98	0.93	0.07	0.02	0.07	0.98
In Presence of 30% Outliers							In Presence of 50% Outliers						
Methods	TPR	TNR	FPR	FNR	FDR	AUC	Methods	TPR	TNR	FPR	FNR	FDR	AUC
SVM	0.92	0.96	0.04	0.08	0.03	0.97	SVM	0.82	0.73	0.27	0.18	0.21	0.77
NBC	0.98	0.97	0.03	0.02	0.03	0.94	NBC	0.88	0.77	0.23	0.12	0.18	0.75
LDA	0.88	0.77	0.23	0.12	0.19	0.80	LDA	0.78	0.57	0.43	0.22	0.35	0.62
KNN	0.86	0.83	0.17	0.14	0.16	0.83	KNN	0.79	0.49	0.51	0.21	0.38	0.57
RF	0.95	0.92	0.08	0.05	0.07	0.97	RF	0.86	0.63	0.37	0.14	0.27	0.76

Table 4. Performance evaluation based on colon cancer dataset

Methods	TPR	TNR	FPR	FNR	FDR	AUC
SVM	0.87	0.89	0.12	0.13	0.07	0.88
NBC	0.91	0.80	0.20	0.09	0.10	0.87
LDA	0.79	0.74	0.26	0.22	0.14	0.75
KNN	0.84	0.85	0.16	0.16	0.09	0.85
RF	0.87	0.88	0.12	0.13	0.07	0.87

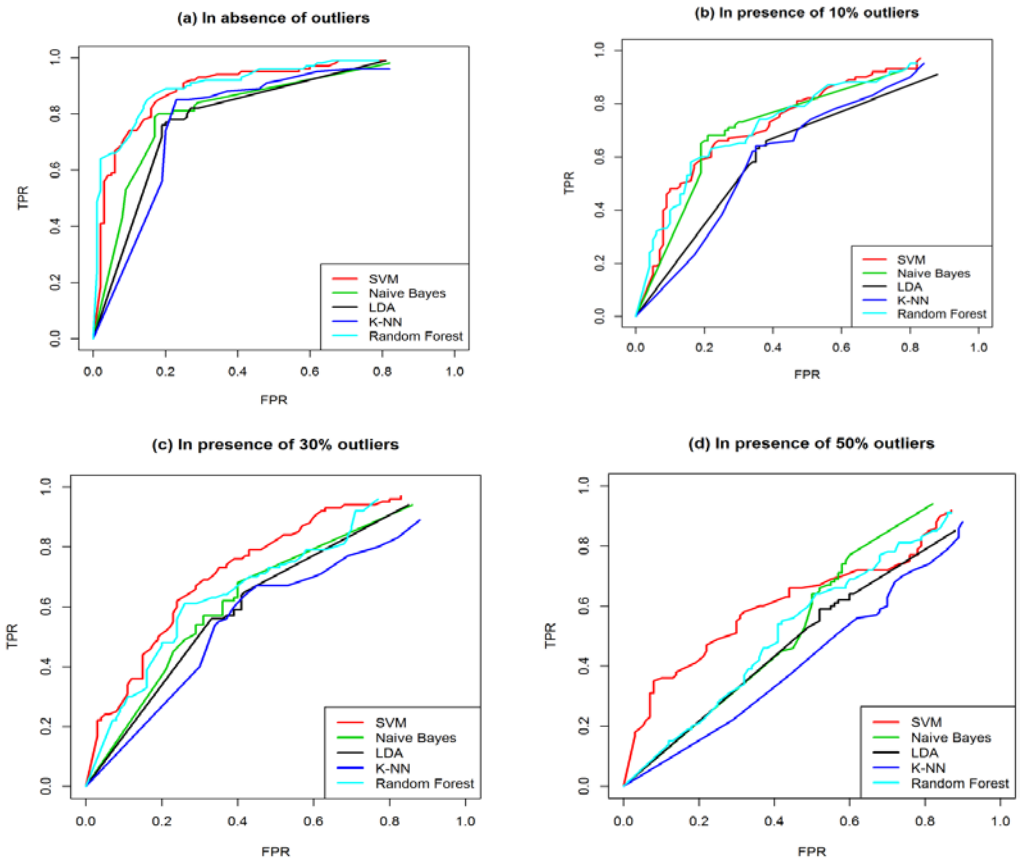


Fig. 1 (a-d). Performance evaluation using test ROC curve produced by five methods based on simulated dataset for small-sample case. (a) in absence of outliers, (b) in presence of 10% outliers, (c) in presence of 30% outliers, and (d) in presence of 50% outliers.

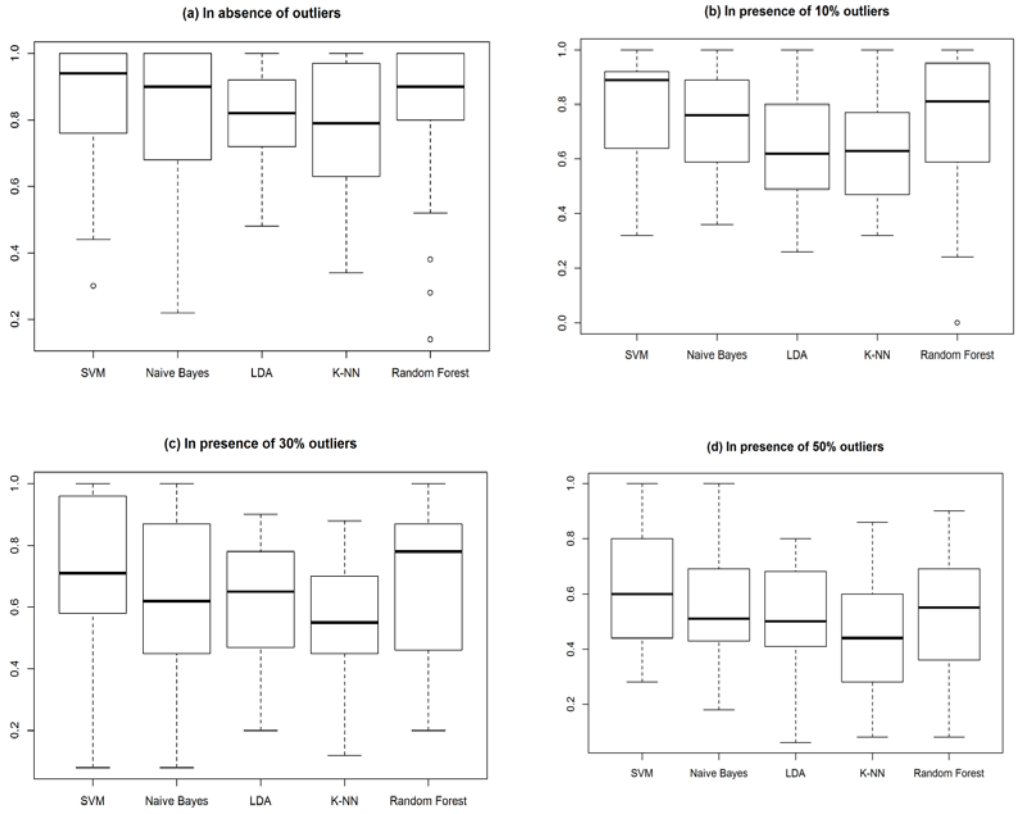


Fig. 2 (a-d). Performance evaluation using test AUC values produced by five methods based on 100 simulated datasets for small-sample. (a) in absence of outliersm, (b) in presence of 10% outliers, (c) in presence of 30% outliers, and (d) in presence of 50% outliers.

Performance evaluation using real gene expression profiles

To investigate the performance of five machine learning algorithms (SVM, NBC, LDA, K-NN and RF) in the colon cancer dataset, the whole dataset was randomly divided into two independent datasets (training and test) such that number of tumor samples in each dataset was same. For the sake of convenience, top 40 features were selected using the *t*-test based on the training dataset. Fig. 3a depicts the heatmap of hierarchical clustering (HC) using Ward's method for randomly divided training dataset. This figure divulged that there are two gene clusters. To confirm this PCA clustering was performed. The number of clusters in the dataset was determined using elbow method (Fig. 3b). This method declared two clusters in the dataset. Fig. 3c illustrates clustering by first two PCs. This figure also confirmed that there are two gene groups in the 40 DE genes. The up-regulated DE gene group consists of 18 genes and down-regulated DE gene group consists of 22 genes. The average values of different performance measures were computed based on test datasets such as TPR, FPR, TNR, FNR, FDR and AUC using 5-fold cross validation. The performance indices were summarized in Table. 4. This table revealed that SVM and Random forest (RF) produced better results compared to the NBC, K-NN and LDA. Fig. 3d shows the boxplot of estimated test AUCs by five

methods using 5-fold cross validation. To unmask the biological functions of top 40 DE genes identified by t-test, WebGestalt2 software package (Zhang et al. 2005) was applied. The GO (Gene Ontology) and KEGG pathway analysis results were obtained using this database. The GO analysis results explored that these genes are involved in different biological processes such as mRNA metabolic process, RNA splicing, mRNA processing, regulation of endodeoxyribonuclease activity etc (see Fig.4). The KEGG analysis revealed that these genes are significantly enriched in purine metabolism, cell adhesion molecules (CAMs), spliceosome, DNA replication pathways etc. (see Table 5). The last column of the tables represents the adjusted p-values. The p-values were obtained from the hypergeometric test and adjusted by Benjamini-Hichberg method. In addition protein-protein interaction (PPI) network using 18 and 22 DE gene was also performed, which was shown in Fig.5 using STRING web server database (Szklarczyk et al. 2017). STRING ingresses protein or gene association knowledge from databases and construct physical interactions according their common biological pathways. The medium confidence score (400) was set to construct the PPI.

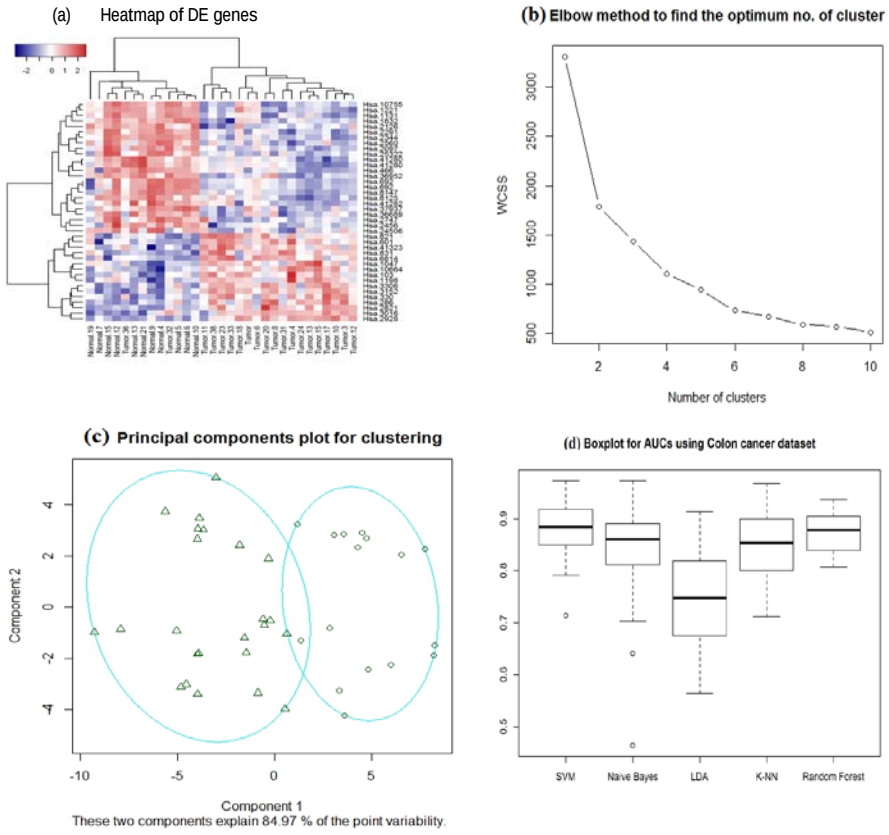


Fig. 3 (a-d). Heatmap and cluster analysis using PCA for DE genes identified by the t-test based on colon cancer dataset. (a) Heapmap of 40 DE genes, (b) Scree plot using 40 most variant genes, (c) Clustering by PCA, and (d) Test AUCs by five methods.

Table 5. KEGG pathways for the 40 DE genes identified by t-test

KEEG ID	Pathway Name	p-value
hsa00230	Purine metabolism	0.040865
hsa04514	Cell adhesion molecules (CAMs)	0.028863
hsa03040	Spliceosome	0.000086246
hsa05412	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	0.12692
hsa00970	Aminoacyl-tRNA biosynthesis	0.0065354
hsa05134	Legionellosis	0.098382
hsa04940	Type I diabetes mellitus	0.077717
hsa00620	Pyruvate metabolism	0.070732
hsa03030	DNA replication	0.06546
hsa00030	Pentose phosphate pathway	0.054834

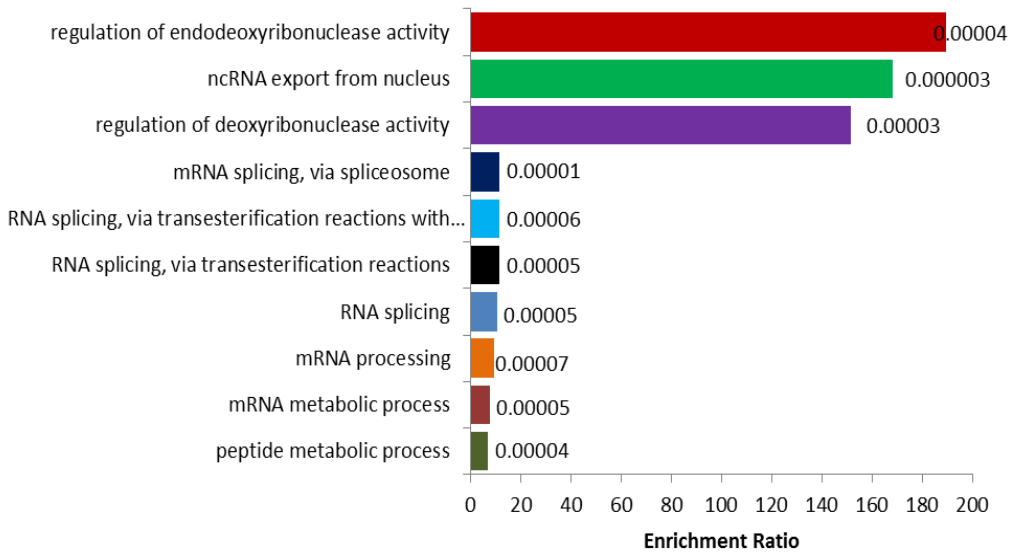
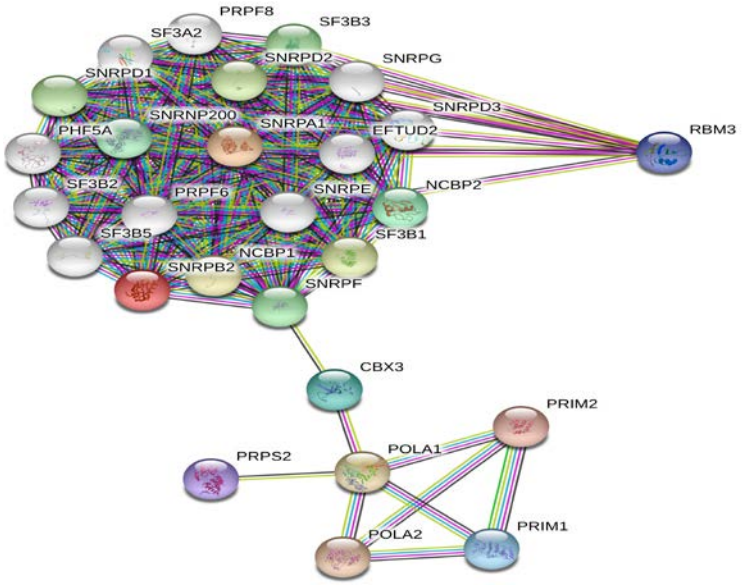


Fig. 4. Bar chart of biological process categories for 40 DE genes identified by t-test based on colon cancer dataset.

(a) PPI-Network for up-regulated DE gene



(b) PPI-Network for down-regulated DE gene

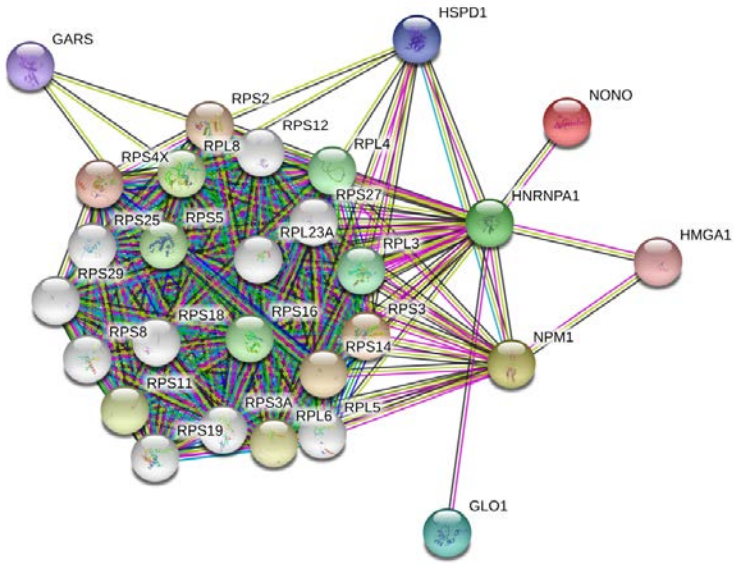


Fig. 5 (a-b). PPI-Network for DE genes identified by the t-test based on colon cancer dataset. (a) PPI-Network for up-regulated DE gene, and (b) PPI-Network for down-regulated DE gene.

Conclusion

Sample classification into one of two populations has been extensively used in gene expression data analysis. Without correct classification of disease samples it is very difficult to provide proper treatment and therapies. There are many machines learning algorithms have been developed to accomplish this task and many comparative studies have been performed among them to select the appropriate algorithm. However, these studies did not consider the problem of outliers. Most of the algorithms are sensitive to outliers and produce higher false positives and lower accuracies in presence of outliers. Therefore, in this paper, a broad comparative study has been carried out among five popular machine learning algorithms (SVM, RF, Naïve Bayes, k-NN and LDA) using both simulated and real gene expression datasets, in absence and presence of outliers. The results obtained from simulated and real GED analysis revealed that SVM, Naïve Bayes, RF produced comparatively better performance than the other two algorithms (k-NN and LDA) for both small-and-large sample sizes.

References

Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D and Levine AJ (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences, 96(12): 6745-6750.

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A (2013). NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res., 41: 991-995.

Dettling M and Buhlmann P (2003). Boosting for tumor classification with gene expression data, Bioinformatics, 19: 1061-1069.

Ding B and Gentleman R (2003). Classification Using Generalized Partial Least Squares, Journal of Computational and Graphical Statistics, 14: 280-298.

Dudoit S, Yang YH, Callow MJ and Speed TP (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, Statistica Sinica, 12: 111-139.

Fayyad UM and Irani KB (1992). The attribute-selection problem in decision tree generation, *In*: Proceedings of the Tenth National Conference on Artificial Intelligence, pp. 104-110.

Fisher R (1936). The use of multiple measurements in taxomonic problems. Annual of Eugenics, 7: 179-188.

Freund Y and Schapire RE (1997). A decision-theoretic generalization of on-line learning and an application to boosting, J Comput System Sci., 55: 119-139.

Friedman N, Geiger D and Goldszmidt M (1997). Bayesian network classifiers, Machine Learning, 29: 131-163.

Golub T, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD and Lander ES (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286: 531-537.

Ho TK (1995). Random Decision Forests, Proceedings of the 3rd international conference on document analysis and recognition, Montreal, QC, pp. 278-282.

Hoerl AE and Kennard RW (1970). Ridge regression: biased estimation for non-orthogonal problems, Technometrics, 12: 55-67.

Hosmer DW and Lemeshow S (1989). Applied Logistic Regression, Wiley Series in Probability and Mathematical Statistics, Wiley, New York.

Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, Dylag M, Kurbatova N, Brandizi M, Burdett T, Megy K, Pilicheva E, Rustici G, Tikhonov A, Parkinson H, Petryszak R, Sarkans U and Brazma A (2015) Array Express update–simplifying data submissions. Nucleic Acids Res., 43: 1113–1116.

Shahjaman M, Kumar N and Mollah MNM (2019). Performance Improvement of Gene Selection Methods using Outlier Modification Rule. Current Bioinformatics, 14: 1-13.

Shahjaman M, Kumar N, Begum AA, Islam SMS and Mollah MNH (2017b). Biomarker Genes Selection Methods: A Comparative Study in Presence of Outliers, Journal of Bio-Science, 25: 9-16.

Shahjaman M, Kumar N, Mollah MMH, Ahmed MS, Begum AA, Islam SMS and Mollah MNH (2017a). Robust Significance Analysis of Microarrays by Minimum β-Divergence Method, BioMed Research International, 1-18.

Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, Jensen LJ, Von Mering C (2017). The STRING database in 2017: Quality-controlled protein-protein association networks, made broadly accessible. Nucleic Acids Res.; 45: 362-8.

Tibshirani R (1996). Regression shrinkage and selection via the LASSO, Journal of the Royal Statistical Society (Series B), 58: 267-288.

Vapnik V (1998). Statistical Learning Theory. Wiley, Chichester, GB.

Zhang B, Kirov S, and Snoddy J (2005). Web Gestalt: an integrated system for exploring gene sets in various biological contexts, Nucleic Acids Research, 33(2): 741-748.

Zuruda JM (1992). Introduction to Artificial Neural Systems, PWS Publishing Company, Boston, NY.

(Manuscript received on August 17, 2019 and revised on October 25, 2019)