Original Article

Using Classification and K-means Methods to Predict Breast Cancer Recurrence in Gene Expression Data

Abstract

Background: Breast cancer is a type of cancer that starts in the breast tissue and affects about 10% of women at different stages of their lives. In this study, we applied a new method to predict recurrence in biological networks made from gene expression data. **Method:** The method includes the steps such as data collection, clustering, determining differentiating genes, and classification. The eight techniques consist of random forest, support vector machine and neural network, randomforest + k-means, hidden markov model, joint mutual information, neural network + k-means and suportvector machine + k-menas were implemented on 12172 genes and 200 samples. **Results:** Thirty genes were considered as differentiating genes which used for the classification. The results showed that random forest + k-means get better performance than other techniques. The two techniques including neural network + k-means and random forest + k-means performed better than other techniques in identifying high risk cases. **Conclusion:** Thirty of 12,172 genes are considered for classification that the use of clustering has improved the classification techniques performance.

Keywords: *Classification, gene, K-means*

Mohammadreza Sehhati[1], Mohammad Amin Tabatabaiefar[2,3], Ali Haji Gholami[4], Mohammad Sattari[5]

[1]Medical Image and Signal Processing Research Center, Department of Bioinformatics,School of Advanced Technologies in Medicine, Isfahan University of Medical Sciences, Isfahan, Iran, [2]Department of Genetics and Molecular Biology, School of Medicine, Isfahan University of Medical Sciences, Isfahan, Iran, [3]Pediatric Inherited Diseases Research Center, Research Institute for Primordial Prevention of Non Communicable Disease, Isfahan University of Medical Sciences, Isfahan, Iran, [4]Department of Hematology-Oncology, Isfahan University of Medical Sciences, Isfahan, Iran, [5]Health Information Technology Research Center, Isfahan University of Medical Sciences, Isfahan, Iran

Submitted: 06-May-2021 **Accepted:** 20-Dec-2021 **Published:** 12-May-2022

Introduction

Breast cancer is a type of cancer that starts in the breast tissue and affects about 10% of women at different stages of their lives.[1] Despite many advances in the early diagnosis and appropriate treatment of the breast cancer, the mortality rate of it is high. In the disease, malignant cells pass through the immune system without defensive or aggressive reactions. In recent years, with the development of diagnostic methods, the early detection of breast cancer is more possible. With the detection of tumors in the early stages, a significant reduction in breast cancer mortality is seen in recent years.[1,2] One of the important issues for a better treatment of breast cancer is dividing patients into specific subgroups and then choosing a specific treatment method for each subgroup. Furthermore, the introduction of marker genes in each subgroup with the aim of identifying the mechanism of cancer progression and choosing an effective treatment is important.

In recent years, gene expression measuring techniques such as microarray have been used in the diagnosis of diseases. This subject shows that the expression of many genes is related to clinical parameters.[3] Studies have shown that microarray technology makes it possible to study tumor behavior in the living tissue and evaluate the diagnosis method and drug resistance.[4,5] Microarray technology measures and displays the expression levels of thousands of genes. The gene expression is the amount of activity of the gene in the experimental.

Sample or in other words, the amount of transcription of that gene.[6] Hence, by having the expression levels related to the genes of a sample, the cellular states of that sample can be described. Many studies such as gene regulation,[7] prognosis and diagnosis,[8] cancer classification,[9] discovery of vital signs,[10] and discovery of new drugs[11] applied microarray data. After initial processing of the microarray data, we will have a numeric array with thousands of rows (genes) and tens of columns (samples). High dimensions,

Address for correspondence:
Dr. Mohammad Sattari,
Isfahan University of Medical Sciences, Salamat Boulevard, School of Management and Information, Room 324, Isfahan, Iran.
E-mail: msattarimng.mui@gmail.com

Access this article online

Website: www.jmssjournal.net

DOI: 10.4103/jmss.jmss_117_21

Quick Response Code:

For reprints contact: WKHLRPMedknow_reprints@wolterskluwer.com

How to cite this article: Sehhati M, Tabatabaiefar MA, Gholami AH, Sattari M. Using classification and K-means methods to predict breast cancer recurrence in gene expression data. J Med Sign Sens 2022;12:122-6.

the small number of samples and inherent variability in laboratory and biological processes have posed serious challenges in the analysis of microarray data. These challenges first increase the computational cost and complexity of the classifiers. Second, they reduce the generalizability and stability of classifiers for predicting new samples.[12] Third, because of the high number of traits compared to the samples, it is very likely that unrelated genes will appear relevant. Fourth, it is difficult to interpret the function of the genes that cause the disease. From a biological point of view, only a small set of genes are related to the disease. As a result, data on the majority of genes actually play the role of a background noise that can obliterate the effect of a small subset. Hence, focusing on a smaller set of gene expression data leads to a better interpretation of the role of information-containing genes. Thus, the first important step in analyzing microarray data is reducing the number of genes, or in other words, selecting differentiating genes for classification. Many studies performed to predict cancer recurrence using the microarray gene expression data.[12] In 2002, Van't Veer *et al*. studied microarray data with 117 breast cancer patients. Lymph nodes in these patients did not contain cancer cells.[13] The study divided dataset into training and test group. It used 98 training samples to analyzing the profile of gene expression levels and extracting the index genes. Then, it use remaining 19 samples for final validation of the model. In the study, they first measured the frequency of transcripts of approximately 25,000 genes in each tumor sample using the Hu25K Agilent microarray. Hence, based on the results of this study, a commercial assay device called mamma print developed and widely used to detect the recurrence of breast cancer using the proposed list of 70 NKI70 genes. Mamma print test evaluated in different countries and the results published in a large number of scientific articles. The positive results of these studies led to the approval of the US FDA in 2007.[14] The results confirm the superior value of mamma print test compared to conventional tests and a significant reduction in patients requiring chemotherapy. These results ensure that treatment decisions are evaluated and reviewed.[15] Another important challenge in analyzing gene expression data is clustering genes based on the cause-and-effect relationship between them. He takes into account gene overlap in different modules identified in different subgroups of the disease.[16,17] In this study, we seek to apply a new method to predict recurrence in biological networks made from gene expression data. The difference between this method and previous methods is that the previous methods used biological nature for predicting recurrence. However, the proposed method used clustering to add a trait to a set of gene traits. The use of clustering has been due to the nature of the method, which is finding similar points. The proposed method used combination of clustering and classification methods to predict breast cancer recurrence. Subjects and Methods: The methods have several steps such as data collection, clustering, determining differentiating genes, and classification.

Data collection

The dataset is in the form of a matrix that contains 200 samples and 12,172 traits (genes). Each gene is a 200-member vector that contains different values in different experiments. To normalize the data, the logarithm function is applied to the data. Then, the function normalized the gene expression values. The method used 200 samples including high-risk groups (samples that metastasized <5 years after diagnosis) and a low-risk group (samples that had no problems such as death or recurrence of cancer for up to 5 years from the diagnosis). Out of 200 samples, 141 samples are low risk and 59 samples are high risk.

Clustering

After dividing the samples into low-risk and high-risk groups, clustering was performed for each group separately using the k-means method. The initial value of the number of clusters was considered equal to 100. In this regard, 100 genes were randomly considered as the centers of the clusters and the remaining genes joined the clusters that have the most similarity based on Euclidean distance to the center of the cluster. After updating the centers of the clusters at each stage of transferring gene between clusters, the above steps are repeated until the change in the value of each cluster center in two consecutive stages is less than a threshold. Finally, clusters in the last iteration of the algorithm are considered as a subnet. Thus, in the end, we will have 100 clusters as 100 subnets for low and high risk group separately. The number of cluster members in the low risk category is between 2 and 200 and in the high risk category between 70 and 185.

Then, in each category, the average values of samples for each gene were calculated. Furthermore, for each of the clusters, the total average value of the samples was calculated and then it divided by the number of the cluster members. The calculated value considered as the average values of the cluster. Then, the average value of the corresponding cluster of each gene was added to the gene vector.

Determining differentiating genes

The clusters are used to extract differentiating genes. Hence, for each gene, two clusters related to low-risk group and high-risk group are considered. The distance between the centers of the two clusters is calculated. Then, we calculate the distance between the clusters of each gene and consider the cluster with the largest distance between the clusters as the distinguishing gene. The optimal number of clusters is obtained by calculating the accuracy of the techniques and determining the maximum accuracy.

Finally, 30 genes were considered as differentiating genes.

Classification

In the classification step, 30 differentiating genes obtained from the previous step is considered an attribute for each sample. The 10-fold cross validation method is applied to divide the data set into train and test set. Based on this, the existing 200 samples are divided into 10 groups. During 10 repetitions of the experiment, each time nine groups that make up 90% of the main data set as train data set. The remaining group considered as test dataset, which forms 10% of the main data set. A train data set is used to generate input patterns to construct a classification model, and an test data set is applied to check its performance. In this paper, random forest techniques,[18] support vector machine[19] and neural network[20] have been used as classification models. Random forest creates multiple trees with different properties. The best decision from the trees determines the index to be associated with the class.[18] Support vector machine seeks a linear relationship with a high confidence margin between the independent and dependent variables.[19] The software used to test the models used is RapidMiner version 9.

Results

Various criteria are used to check the methods. One of these criteria is accuracy, which the closer the accuracy of this criterion, the better the result.[21] This criterion is calculated based on the following formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Other criteria are precision and recall[22], the closer to one the better the performance.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

The target group is the high and low risk of breast cancer metastasis. The True Positive (TP) identifies the number of records that the group has correctly placed on the high risk of breast cancer metastasis. True Negative identify records that have been correctly identified as low risk class for breast cancer metastasis. False Positive (FP) identifies records that the group that are mistakenly place in the high risk class for breast cancer metastasis. False Negative (FN) identifies records that are mistakenly place in the low risk for breast cancer metastasis. Then, based on the confusion matrix, the efficiency criteria of the data mining method are calculated.

The target class comprises two modes: Low risk and high risk. The 70% of the samples belong to the low risk class and 30% of the samples belong to the high risk class. We compare the proposed methods including random forest + k-means, support vector machine + k-means, and neural networks + k-means with random forest,[18] support vector machine,[19] neural networks,[20] hidden markov models,[23] and joint mutual information.[12]

Figure 1 shows the accuracy of the techniques based on the number of differentiating genes. As it is, random forest + k-means method performed better than the other methods. The lowest amount of random forest + k-means accuracy is corresponded to 20 and 35 genes and the highest amount of it is related to 30 genes. Moreover, it can be said that the least value of support vector machine + k-means accuracy is corresponded to 35 genes and the highest value of it is related to 30 genes. The neural network + k-means have the highest possible oscillation, so that the distance between the least value and the greatest value is equal to 0.13. Furthermore, as shown in Figure 1, the use of clustering improves the performance of classification techniques such as random forest, support vector machine, and neural network. Finally, according to Figure 1, the number of selected genes will be equal to 30. The accuracy of all seven techniques with 30 selected genes is also in Table 1.

According to recall, the performance of all seven techniques in the high class was better than the low class [Table 2]. The support vector machine + k-means technique was able to identify 81% of the Low risk class, while random forest + k-means and neural network + k-means recognize 100% of cases in high risk class.

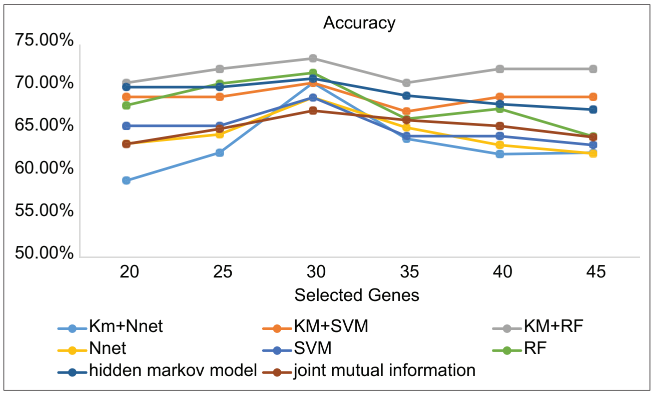

Figure 1: The performance of techniques

Table 1: Accuracy of different techniques

	Accuracy (%)
Random forest	71.67
Neural network	68.78
SVM	68.78
Randomforest+Kmeans	73.37
SVM+kmeans	70.49
Neural networks+kmeans	70.49
Hidden markov model[24]	70.76

SVM - Support vector machines

Table 2: Recall and precision of different techniques

Class	Random forest[18]		Neural network[20]		SVM[19]		Random forest + kmeans		Neural network + k-means		SVM+k-means		Hidden markov model[24]		Joint mutual information[12]	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Low risk (%)	65.28	30.56	63.28	57	66	53.25	72.41	15.79	70	63.16	81.58	5	64.23	59	62.25	45.23
High risk (%)	95.65	93.25	92.27	85.13	75.53	90	100	100	100	73.81	52.17	100	92.22	86.25	84.32	96.12

SVM - Support vector machines

Discussion

In this research, an attempt has been made to investigate the low-risk or high-risk status of different samples. First, 12,172 genes were examined, then the genes were clustered separately in the low-risk and high-risk groups. A new attribute called Cluster Mean was added to the attribute set. Considering this feature, out of 12,172 genes, 30 genes were selected that make more distinction between low-risk and high-risk groups. These 30 genes were selected as sample attributes. In fact, these genes can represent as a candidate set for classification.

The eight techniques consists of random forest, support vector machine and neural network, random forest + k-means, neural network + k-means, hidden markov model, joint mutual information and suport vector machine + k-menas were implemented. The results showed random forest + k-means has better performance than other techniques.

Given the accuracy of methods, it can be said that 30 selected genes can be a good representation for the data set. Also, the two techniques including neural network + k-means and random forest + k-means performed better than other techniques in identifying high risk cases. Given that the identification of such cases is very important,[24] so the correct diagnosis the cases can be very helpful in making appropriate decisions about cancer patients. However, the support vector machine + k-means predict low risk cases more accurate than other techniques. Another point is that the use of clustering has improved the classification techniques performance. This is because in clustering, similar points are placed side by side, which can improve classification.

Sehati *et al*. developed a markov model-based method for extracting selected genes from the breast cancer dataset. He concluded that 20 genes are a good representation of the selected dataset.[23] The number of input genes similar to this study was equal to 12,172, but the samples are more than this study and equal to 1271 samples. In fact, the number of samples is almost six times that of this study.

Financial support and sponsorship

None.

Conflicts of interest

There are no conflicts of interest.

References

1. Bagherian H, Haghjooy Javanmard S, Sharifi M, Sattari M. Using data mining techniques for predicting the survival rate of breast cancer patients: A review article. Tehran Univ Med J 2021;79:176-86.
2. Garcia-Murillas I, Schiavon G, Weigelt B, Ng C, Hrebien S, Cutts RJ, *et al.* Mutation tracking in circulating tumor DNA predicts relapse in early breast cancer. Sci Transl Med 2015;7:302ra133.

3. Yoo SM, Choi JH, Lee SY, Yoo NC. Applications of DNA microarray in disease diagnostics. J Microbiol Biotechnol 2009;19:635-46.
4. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, *et al.* Molecular portraits of human breast tumours. Nature 2000;406:747-52.
5. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, *et al.* Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 2001;98:10869-74.
6. Momenzadeh M, Sehhati M, Rabbani H. A novel feature selection method for microarray data classification based on hidden Markov model. J Biomed Inform 2019;95:103213.
7. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol 2013;31:46-53.
8. Bustamam A, Sarwinda D, Ardenaswari G. Texture and gene expression analysis of the mri brain in detection of alzheimer's disease. Journal of Artificial Intelligence and Soft Computing Research. 2018;8:111-20.
9. Kourou K, Rigas G, Papaloukas C, Mitsis M, Fotiadis DI. Cancer classification from time series microarray data through regulatory dynamic Bayesian networks. Comput Biol Med 2020;116:103577.
10. Ke W, Wu C, Wu Y, Xiong NN. A new filter feature selection based on criteria fusion for gene microarray data. IEEE Access 2018;6:61065-76.
11. Sato H, Ishida S, Toda K, Matsuda R, Hayashi Y, Shigetaka M, *et al.* New approaches to mechanism analysis for drug discovery using DNA microarray data combined with KeyMolnet. Curr Drug Discov Technol 2005;2:89-98.
12. Sehhati M, Mehridehnavi A, Rabbani H, Pourhossein M. Stable gene signature selection for prediction of breast cancer recurrence using joint mutual information. IEEE/ACM Trans Comput Biol Bioinform 2015;12:1440-8.
13. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, *et al.* Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530-6.
14. Slodkowska EA, Ross JS. MammaPrint 70-gene signature: Another milestone in personalized medical care for breast cancer patients. Expert Rev Mol Diagn 2009;9:417-22.
15. Wittner BS, Sgroi DC, Ryan PD, Bruinsma TJ, Glas AM, Male A, *et al.* Analysis of the MammaPrint breast cancer assay in a predominantly postmenopausal cohort. Clin Cancer Res 2008;14:2988-93.
16. Lu X, Zhu Z, Peng X, Miao Q, Luo Y, Chen X. InFun: a community detection method to detect overlapping gene communities in biological network. Signal, Image and Video Processing 2021;15:681-6.
17. Pio G, Ceci1 M, Prisciandaro F, Malerba D. Exploiting causality in gene network reconstruction based on graph embedding. Mach Learn 2020;109:1231-79.
18. Qi Y. Random forest for bioinformatics. In: Ensemble Machine Learning. Boston, MA: Springer; 2012. p. 307-23.
19. Noble WS. What is a support vector machine? Nat Biotechnol 2006;24:1565-7.
20. Albawi S, Mohammed TA, Al-Zawi S. Understanding of a convolutional neural network. In: 2017 International Conference on Engineering and Technology (ICET). Turkey: IEEE; 2017. p. 1-6.
21. Chan PK, Stolfo SJ. On the accuracy of meta-learning for scalable data mining. J Intell Inf Syst 1997;8:5-28.
22. Sajjadi MS, Bachem O, Lucic M, Bousquet O, Gelly S. Assessing generative models via precision and recall. Arxiv 2018;31.
23. Momenzadeh M, Sehhati M, Rabbani H. Using hidden Markov model to predict recurrence of breast cancer based on sequential patterns in gene expression profiles. J Biomed Inform 2020;111:103570.
24. Bick U, Engel C, Krug B, Heindel W, Fallenberg EM, Rhiem K, *et al.* High-risk breast cancer surveillance with MRI: 10-year experience from the German consortium for hereditary breast and ovarian cancer. Breast Cancer Res Treat 2019;175:217-28.