MEDICAL RECORDS-International Medical Journal

Research Article

Comparison of Feature Selection Methods in Breast Cancer Microarray Data

Melih Agraz

Giresun University, Faculty of Arts and Science, Department of Statistics, Giresun, Türkiye

Abstract

Aim: We aim to predict metastasis in breast cancer patients with tree-based conventional machine learning algorithms and to observe which feature selection methods is more effective in machine learning methods related to microarray breast cancer data reducing the number of features.

Material and Methods: Feature selection methods, least squares absolute shrinkage (LASSO), Boruta and maximum relevance-minimum redundancy (MRMR) and statistical preprocessing steps were first applied before the tree-based learning conventional machine learning methods like Decision-tree, Extremely randomized trees and Gradient Boosting Tree applied on the microarray breast cancer data.

Results: Microarray data with 54675 features (202 (101/101 breast cancer patients with/without metastases)) was first reduced to 235 features, then the feature selection algorithms were applied and the most important features were found with tree-based machine learning algorithms. It was observed that the highest recall and F-measure values were obtained from the XGBoost method and the highest precision value was received from the Extra-tree method. The 10 arrays out of 54675 with the highest variable importance were listed.

Conclusion: The most accurate results were obtained from the statistical preprocessed data for the XGBoost and Extra-trees machine learning algorithms. Statistical and microarray preprocessing steps would be enough in machine learning analysis of microarray data in breast cancer metastases predictions.

Keywords: Microarray, breast cancer, metastasis, machine learning, feature selection

INTRODUCTION

Cancer can be defined as a disease with uncontrolled cell growth, metastasizing and attacking other tissues (1). After non-melanoma skin cancer, breast cancer is the 2nd common cancer for women in worldwide (2) and it is known as the primary cancer among women (3). In addition to that this is also a big problem for patients who are diagnosed with cancer and recover, so Bahceli and Kucuk (4) showed that fear of cancer recurrence is high in women with breast cancer in Turkey. Metastasis is a process in which cancer cells disperse from the primary tumor site and spread from there to different parts of the body (5). The majority of breast cancer deaths are due to breast cancer metastases (6). Thereby, predicting whether metastasis would occur or not is important in terms of taking precautions.

DNA Microarray technology is an old but effective method and results obtained from microarray analysis are robust, since it has been possible to calculate the thousands of genes simultaneously with microarray technology (7). With DNA microarray technology, large microarray data of gene expression have begun to be produced and this data has been used for the discovery and classification of diseases (1). In addition, cancer studies with microarray data have been carried out for a long time; Dhanasekaran et al. (8) studied prostat cancer on microarray data and they successfully classified the metastatic prostate, normal prostate and localized prostate cancer. Chang et al.

CITATION
Agraz M. Comparison of Feature Selection Methods in Breast Cancer Microarray Data. Med Records. 2023;5(2):284-9. DOI:1037990/medr.1202671

Received: 11.11.2022 **Accepted:** 04.01.2023 **Published:** 23.03.2023

Corresponding Author: Melih Agraz, Giresun University, Faculty of Arts and Science, Department of Statistics, Giresun, Türkiye
E-mail: melih.agraz@giresun.edu.tr

(9) demonstrated the using cDNA microarrays to identify arrays involved in transformation in oral cancer. van't Veer et al. correctly predicted the output of the breast cancer disease for 65 patients out of 75 patients. In addition to these studies, many machine learning (ML) researches with microarrays to determine breast cancer have been published recently. Paksoy and Yangin (23) predicted the colon cancer on microarray data. Pirooznia et al. (10) first applied feature selection (FS) algorithms, such as correlation FS, support vector machine recursive feature elimination and chi squared methods, after that, they run ML models on selected featured data and compared the results. Cho and Won (11) predict and diagnose cancer on microarray data with ML algorithms after applying signal to noise ratio FS algorithms, correlation coefficients, Euclidean distance and information gain. Alagukumar and Kathirvalavakumar (12) applied FS algorithms, like Welch test, ANOVA, Wilcoxon test, Kruskal–Wallis, LIMMA, and F-test to extract the microarray genes and proposed classifier. Lonith (13) proposed principal component analysis used to decrease the number of features on microarray data for breast and liver cancer and particle swarm optimization to increase the ML algorithms accuracy. Mod et al (14) proposed some hybrid FS algorithms whale optimization algorithm, grey wolf optimization, gravitational search algorithm, cuckoo search algorithm, firefly algorithm, artificial bee colony optimization and particle swarm optimization for the breast cancer microarray data.

As briefly explained in the literature review, microarray data includes huge number of genes with very small observations (n<<p), so FS methods gain importance in microarray data. FS is one of the key steps of the ML algorithm, because the dataset that best expresses the output will come from the best features. In this study, we try to improve the metastasis prediction accuracy with the latest FS algorithms and proposed best arrays for the future studies and compare the FS methods on breast cancer microarray data. For this reason, we applied three different FS algorithms such as LASSO (15), boruta (16) and MRMR (17,18) and statistical method as preprocessing microarray analysis.

MATERIAL AND METHOD

In this study, as represented in flowchart of Figure 1, two different data GSE102484 and GSE20685 were first downloaded.The datasets can be found in NCBI Geo Databank. After that, we combined the data by using R programming language 4.1.1. Both datasets includes breast cancer patients with metastasis (label-1) and non-metastasis (label-0). We excluded outliers (3 standard deviations away from the mean) and missing observations and normalized the variables in the Microarray Preprocessing step. After that, we have 54675 features and 202 observations/patients (101 metastasis and 101 non-metastasis). Since the number of features are less than the observations (n<<p), we applied statistical data preprocessing (Statistical FS) analysis to reduce the number of features. In this part, we selected the features that were well distinguished by classes and we call it histogram differences method. This method is briefly explained under the Histogram Differences part. We selected the 235 features after the statistical and microarray preprocessing steps and this is our original feature pool. We plotted the heatmap to see the correlation between the patient and gene expression sequences, as can be seen in Figure 2. According to heat map, there is an associations between patients and arrays, so the microarray data with 235 features is applicable to the ML classification problem. After Microarray and Statistical data analysis preprocessing, we had more feature than observation (n:202<p:235). Even this is applicable for the ML analysis, we can still apply feature selection methods in order to continue with the data set that is less and better describes the output. For this reason, finally, we executed the least absolute shrinkage and selection operator (LASSO), boruta and maximum relevance-minimum redundancy (MRMR) FS algorithms to minimize the dataset and compared the results.

Machine Learning Algorithms

Machine learning (ML) is a research topic in statistics and computer sciences that makes inferences from data by imitating the way human's learning. With the increasing amount of data and developments in computer science, the interest in ML has increased in recent years in health sciences. Tree-based algorithms are used in many ML related studies (19,20). In this study, we use tree-based conventional ML algorithms, Decision Trees, Extremely Randomized Trees and Gradient Boosting Trees, since they give the variable importance of the model. Decision Trees (DT) is the first and simplest version of tree-based models that make decisions using leaves and nodes. Extremely Randomized Trees (Extra-Trees) (21), also known as Extra-Trees is a kind of ensemble machine learning method that is similar to random forest model. Gradient Boosting Tree (XGBoost) (22) is used in both regression and classification algorithms like the other algorithms that used in this study and it is an advanced learning ensemble method that uses progressively improved predictions to obtain a final prediction result.

Feature Selection (FS) Algorithms

Feature selection algorithms can be divided into 3 categories as Filter, Wrapper and Embedded. We tried to select different feature selection algorithms from the different part of the category such as LASSO (embedded), boruta (wrapper) and MRMR (filter).

Histogram Differences

This method simply selects the features that is well distinguished by classes and we call this method histogram differences method in this study. To do that, we followed the following steps;

I) min and max values of the features were calculated in the entire data set. It was observed that the min-max range of the features are between 1.9732-14.3942.

II) Range between 0 to 15 was chosen to keep the histograms in the standard range and 100 was selected as the number of bins.

III) Two histograms were extracted for each feature, one with a Y value of label-1 and the other with Y value of label-0.

IV) These two histograms for each feature were subtracted from each other, the absolute value was taken and divided by the number of samples.

V) The sum of the two histogram differences was converted into a score, and the distribution of scores for all features was examined. A certain threshold value was selected (3 in this study) and the features higher than 3 were selected.

Least absolute shrinkage and selection operator (LASSO)

LASSO (15) is a regression technic that can be used in both variable selection and regularization by using the following loss formula: (1)

$$\text{Loss}=\sum_{i=1}^{N}|y_i-\sum_j\beta_j x_{ij}|^2+\lambda\sum_j|\beta_j|$$

where λ is numerical value between 0 and 1, x is input, y is output. LASSO gives better results as feature selection method when there are few observations and too many variables. LASSO feature selection ensures that unrelated variables are removed from the model by making their coefficients zero (24,25).

Boruta

Boruta FS (16) is an ensemble feature selection algorithm which uses random forest classifier and it was first developed as a package of R-programming language. Boruta algorithm uses a Random Forest (26) classifier-based wrapper approach for robust feature selection method (27).

Maximum Relevance - Minimum Redundancy (MRMR)

MRMR was first defined by Peng et al. (17) as a embedded feature selection algorithm based on mutual informations with the following formula; (2)

$$max[\frac{1}{|S|}\sum_{x_i\in S}I(x_i;C)-\frac{1}{|S|^2}\sum_{x_i\in S}\sum_{x_j\in S}(I(x_i;x_j))]$$

where x_i is the mth feature in subset S. Since the MRMR method is fast and effective, it has been studied extensively. Zhao et al. (18) determined FCQ-MRMR method which uses the F-score to measure the relevance and calculate the correlation between features to measure redundancy as represented the formula below (3).

$$f_{FCQ}=F(X_i,y)/\frac{1}{|S|}\sum_{X_s\in S}\rho(X_s,X_i)$$

Where F is the F-score between ith feature and response variable and ρ is the correlation between the features. Ding and Peng (28) applied the MRMR method on microarray data and compared the feature selected results and baseline features. They showed that MRMR outperform then the baseline features for both continuous and discrete datasets.

Accuracy Measures

In binary classification problems, we can show the accuracy values of ML methods by various methods. Since our dataset is balanced and we want to specify the success of true metastasis prediction, we only used F1, recall and precision measures as formulated below (4-6).

$$\text{Precision}=\frac{TP}{TP+FP},$$
$$\text{Recall}=\frac{TP}{TP+FN},$$
$$\text{F1}-\text{Score}=\frac{2*TP}{2*TP+FP+FN},$$

Where FP (false positive) represents the wrong prediction classes when the actual class is positive (metastasis), TP (true positive) shows true predicted classes, and FN (false negative) is the wrong prediction when the actual one is negative.

RESULTS

In this study, feature selection (FS) methods on microarray breast cancer data were used and tree-based conventional ML methods were executed to see the model prediction accuracy and variable importance. As can be seen in the flowchart in Figure 1, 54675 features are very high compared to the observations for the ML applications.

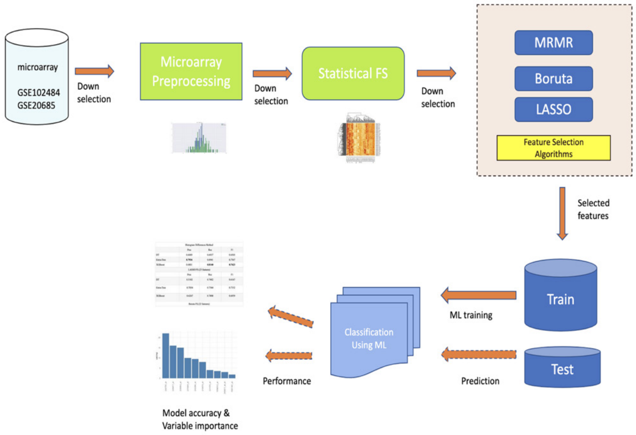

Figure 1. Flowchart of the methodology

Thereby, we applied microarray FS methods and statistical preprocessing method to the dataset and the number of features were reduced to 235 arrays. Afterwards, tree-based ML methods were applied to this dataset and it is observed that XGBoost is the best method with recall 0.8140 and F1-measure: 0.7423 on prediction metastasis, but we think that the accuracy values could be increased by selecting appropriate features from 235 datasets, since conventional ML algorithms work best with the less

Table 1. Cross validated accuracy measure results of tree-based machine learning algorithms on selected features

	Histogram Differences Method		
	Prec	Rec	F1
DT	0.6069	0.6937	0.6505
Extra-Tree	**0.7934**	0.6941	0.7367
XGBoost	0.6801	**0.8140**	**0.7423**
	LASSO FS (23 features)		
	Prec	Rec	F1
DT	0.5182	0.7482	0.6167
Extra-Tree	0.7054	0.7544	0.7332
XGBoost	0.6267	0.7498	0.6939
	Boruta FS (22 features)		
	Prec	Rec	F1
DT	0.6198	0.6212	0.6156
Extra-Tree	0.6738	0.6170	0.6469
XGBoost	0.7531	0.5643	0.6416
	MRMR FS (20 features)		
	Prec	Rec	F1
DT	0.4346	0.3810	0.4032
Extra-Tree	0.7509	0.5606	0.6428
XGBoost	0.6302	0.5608	0.617

Prec: precision; Rec: recall; F1:F1-score/measure

features. Therefore, we applied boruta, MRMR and LASSO feature selection methods to this dataset. Results are listed in Table 1 with precision, F1 and recall values, with the highest values shown in bold.

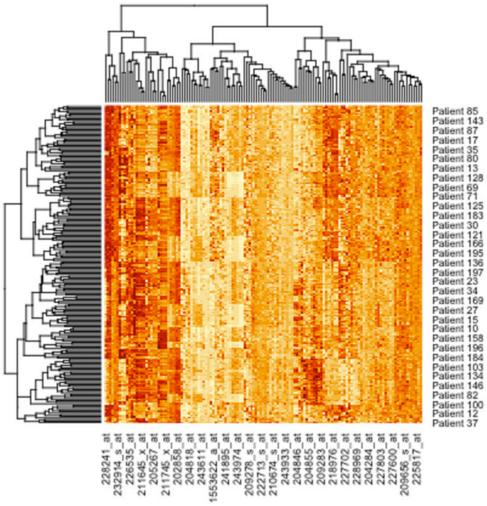

Figure 2. Heat map of the microarray data after preprocessing steps

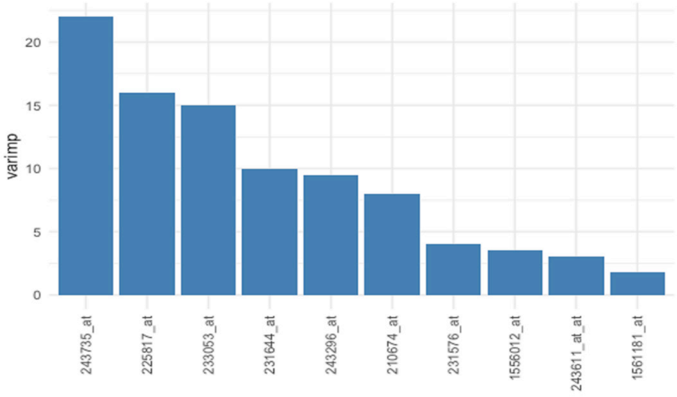

Figure 3. Variable importance of features selected from the Histogram Differences Method with XGBoost algoirthm

DISCUSSION

According to the Table 1, the highest value in precision was seen in Histogram Difference Method Extra-tree model as 0.7934, and the highest values in recall and F1 were obtained in XGBoost as 0.8140 and 0.7423, respectively. It is observed that the best results were coming from the

Histogram Difference method. So the top 10 arrays from this method is listed in Figure 3 by variable importance. Figure 3 lists the top 10 features provided by the XGBoost model. According to the Figure 3, the arrays of highest importance used in the model are listed; 243735_at, 225817_at, 233053_at, 231644_at, 243296_at, 210674_at, 231576_at, 1556012_at, 243611_at, 1561181_at as listed in Figure 3.

Limitations

This paper has some limitations. Performing the analysis with very few patients was not enough for us to use all machine learning methods. However, we were able to start analyzes with 202 patients, as the preprocessing such as merging and cleaning the data took too much time. Other limitation is, extracting missing data in microarray preprocessing step may have caused information loss, but there were too many missing observations for missing data imputation.

CONCLUSION

High number of features with few observations (n<<p) is a problem in microarray data. Thereby, statistical and microarray preprocessing steps can be used to reduce the dimension of feature, especially in ML studies. We can say that XGBoost is the most useful conventional ML algorithm in ML studies for metastasis prediction in breast cancer patients. For the future studies, cancer researchers can examine the arrays listed in Figure 3, in addition to that prediction can be done with Super Learner (19) which approaches the same level of accuracy as the best algorithm among the candidate learners in asymptotically and deep learning methods.

Financial disclosures: *The authors declared that this study hasn't received no financial support.*

Conflict of Interest: *The authors declare that they have no competing interest.*

Ethical approval: *Ethical approval is not required.*

Acknowledgements: *The author received no financial support for the research. Thanks to Genz Biotechnology Co, Umut Agyuz and Birol Kuyumcu for their valuable supports.*

REFERENCES

1. Abd-Elnaby M, Alfonse M, Roushdy M. Classification of breast cancer using microarray gene expression data: a survey. J Biomed Inform. 2021;117.
2. Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;68:394-424.
3. Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018;70:313.
4. Bahçeli PZ, Kucuk BY, Fear of cancer recurrence in women with breast cancer: A cross-sectional study after Mastectomy, Med Records. 2022;4:315-20.
5. Chaffer CL, Weinberg RA. A perspective on cancer cell metastasis. Science. 2011;25:331.
6. Scully OJ, Bay B, Yip G, Yu Y. Breast cancer metastasis. Cancer Genomics Proteomics. 2012:9;311-20.
7. Curtis RK, Oresic M, Vidal-Puig A. Breast cancer metastasis Pathways to the analysis of microarray data. Trends Biotechnol. 2005;23:429–35.
8. Dhanasekaran SM, Barrette TR, Ghosh D, et al. Delineation of prognostic biomarkers in prostate cancer. Nature. 2001;412:822–6.
9. Chang DD, Park NH, Denny CT, et al. Characterization of transformation related genes in oral cancer cells. Oncogene. 1998;16:1921-30.
10. Pirooznia M, Yang JY, Yang MQ, et al. A comparative study of different machine learning methods on microarray gene expression data. BMC Genomics. 2008;9:13.
11. Sung-Bae C, Hong-Hee W. Machine learning in dna microarray analysis for cancer classification. APBC. 2003;189-98.
12. Alagukumar S, Kathirvalavakumar T. Classifying Microarray Gene Expression Cancer Data Using Statistical Feature Selection and Machine Learning Methods. In: Saraswat, M., Sharma, H., Balachandran, K., Kim, J.H., Bansal, J.C. (eds) Congress on Intelligent Systems. Lecture Notes on Data Engineering and Communications Technologies, 2022;114.
13. Lohith RD, Chetty RN, Shaan MS, et al. Gene Expression Analysis using Particle Swarm Optimization and Machine Learning Algorithms for Diagnosing Liver & Breast Cancer, 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC), 2022;1176-81.
14. Mohd A, Besar N. Hybrid feature selection of breast cancer gene expression microarray data based on metaheuristic methods: a comprehensive review. Symmetry. 2022;14:1955.
15. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc. B: Stat Methodol. 1996;58:267-88.
16. Miron B, Witold R. Rudnicki. Feature selection with the boruta package. J Stat Softw. 2010;36:1-13.
17. Hanchuan P, Fuhui L, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell. 2005;27:1226-38.
18. Zhao Z, Anand R, Wang M. Maximum relevance and minimum redundancy feature selection methods for a marketing machine learning platform. In 2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA) 2019;442–52.
19. Van der Laan MJ, Polley EC, Hubbard AE. Super learner. Stat Appl Genet Mol Biol. 2007; 6(1).
20. Secilmis D, Agraz M, Purutcuglu V. Two New Nonparametric Models for Biological Networks, In Hemanchardan K. et al. (editors) Bayesian Reasoning and Gaussian Processes for Machine Learning Applications. 2022;CRC Press.
21. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn. 2006;63:3–42.

22. Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of Statistics. 2001:1189–232.

23. Paksoy N, Yangin HF. Artificial Intelligence-based colon cancer prediction by identifying genomic biomarkers. Med Records. 2022;4:196-202.

24. Güçkiran K, Cantürk İ, Özyilmaz L. DNA microarray gene expression data classification using SVM, MLP, and RF with feature selection methods relief and LASSO. Journal of Suleyman Demirel University Institute of Science and Technology. 2019;23:126-32.

25. Baha Ş. Importance of attribute selection for parkinson disease. Academic Platform J Engineering Sci. 2020;8:175-80.

26. Breiman L. Random forests. Machine Learning. 2001;45:5–32.

27. Lacalamita A, Piccinno E, Scalavino V, et al. A Gene-based machine learning classifier associated to the colorectal adenoma-carcinoma sequence. Biomedicines. 2021;9:1937.

28. Ding C, Peng H. Minimum redundancy feature selection from microarray gene expression data. J Bioinform Comput Biol. 2005;3:185-205.