Original Article Open Access

Archives Of Breast Cancer

DOI: 10.19187/abc.20185111-14

A Comparative Study of Multilayer Neural Network and C4.5 Decision Tree Models for Predicting the Risk of Breast Cancer

Soolmaz Sohrabi[a], Alireza Atashi*[b,c], Ali Dadashi[d], Sina Marashi[b]

[a] *Shahid Beheshti University of Medical Sciences, Department of Medical Informatics, Tehran, Iran*
[b] *Department of E-Health, Virtual School, Tehran University of Medical Sciences, Tehran, Iran*
[c] *Medical Informatics Department, Breast Cancer Research Center, Motamed cancer institute (ACECR), Tehran, Iran*
[d] *Mashhad University of Medical Sciences, Department Of Medical Informatics, Mashhad, Iran*

ARTICLE INFO

Received:
13 October 2017
Revised:
19 January 2018
Accepted:
26 January 2018

Key words:
Decision tree,
multilayer neural network,
breast cancer,
data analysis

ABSTRACT

Background: Diagnosing breast cancer at an early stage can have a great impact on cancer mortality. One of the fundamental problems in cancer treatment is the lack of a proper method for early detection, which may lead to diagnostic errors. Using data analysis techniques can significantly help in early diagnosis of the disease. The purpose of this study was to evaluate and compare the efficacy of two data mining techniques, i.e., multilayer neural network and C4.5, in early diagnosis of breast cancer.

Methods: A data set from Motamed Cancer Institute's breast cancer research clinic, Tehran, containing 2860 records related to breast cancer risk factors were used. Of the records, 1141 (40%) were related to malignant changes and breast cancer and 1719 (60%) to benign tumors. The data set was analyzed using perceptron neural network and decision tree algorithms, and was split into two a training data set (70%) and a testing data set (30%) using Rapid Miner 5.2.

Results: For neural networks, accuracy was 80.52%, precision 88.91%, and sensitivity 90.88%; and for decision tree, accuracy was 80.98%, precision 80.97%, and sensitivity 89.32%. Results indicated that both algorithms have acceptable capabilities for analyzing breast cancer data.

Conclusion: Although both models provided good results, neural network showed more reliable diagnosis for positive cases. Data set type and analysis method affect results. On the other hand, information about more powerful risk factors of breast cancer, such as genetic mutations, can provide models with high coverage.

Introduction

Breast cancer is a disease in which malignant cells originate from breast tissue and proliferate irregularly and increasingly while passing immune system without causing any defensive and aggressive immune response.[1, 2] The disease usually initiates as a solid mass in superior lateral region of breast and may expand to axillary lymph nodes and then to the whole body.[3] Although cancer is the result of a combination of genetic and environmental factors, the main cause of breast cancer is not clear. A number of risk factors are known for breast cancer,[4, 5] including genetic and racial factors, diet, obesity, hormones, radiation, menopause (after age 50), oral contraceptives use, hormone therapy, family history, and alcohol consumption.[5, 6] Thus, identification of all breast cancer risk factors along with taking right actions to increase public awareness about those factors can help in the prevention and early detection of the disease. Using artificial intelligence and soft calculations are among the methods which can facilitate diagnosis, identification, and decision making in cancers, especially breast cancer.[7,8]

Address for correspondence:
Alireza Atashi, PhD
Address: Cancer Informatics Department, Breast Cancer Research Center
No. 1270, Enghelab Avenue, Tehran, Iran
P.O.Box 14155-4364
Tel: +98 2166462002
Fax: +98 2166400730
E-mail: smatashi@yahoo.com

Researchers have been interested in artificial intelligence for developing prediction models in various scientific fields such as medical engineering. Medical prediction models help physicians in overcoming health care problems and decreasing medical errors.[9] Furthermore, data analysis models may result in better and more accurate diagnosis of conditions in clinical settings by detecting hidden patterns. Classification is one of the main functions of data analysis. Neural networks are the most applicable models of artificial intelligence in medicine because they provide accurate responses and decision trees and the process is simple to follow.[10] Since the classification of medical problems is inherently non-linear, prediction models based on linear statistical methods would not be precise. Furthermore, conventional statistical techniques are not suitable for analyzing large data sets.[11] Data analysis and its techniques, if used properly, can be more efficient in this regard. Considering that the use of modern technologies and software knowledge have increased in medicine during the last two decades, and given the fact that early diagnosis has a significant role in decreasing cancer mortality, applying data analysis techniques to breast cancer data sets and extracting useful results for improvement of accuracy in medical diagnosis is crucial.[11]

Given the importance of breast cancer and its early diagnosis as well as understanding the effective role of different data analysis methods in development of prediction models, it seems essential that the accuracy of these techniques be evaluated practically in various sites and the most efficient and effective models be identified. Thus, the objective of the present study was to compare the accuracy of two different models, namely, neural network data analysis and decision tree, in predicting the risk of breast cancer.

Methods

In this retrospective study, a data set from Breast Cancer Research Center of Motamed Cancer Institute, Tehran, Iran, were used, which contained information related to patients admitted to ACECR breast diseases clinic in Tehran from March 2007 to September 2015. Every record consisted of 14 fields of information on breast cancer risk factors and 1 field on the type of main tumor (malignant or benign). The data set consisted of 2860 records, of which 1141 (40%) were related to breast cancer patients and 1719 (60%) to benign breast tumors. Table 1 presents the evaluated risk factors.

In preprocessing stage, columns unrelated to disease risk factors or related to patients' demographic information were omitted. Then, for the purpose of increasing validity, efforts were made to omit records with more than 20% missed information and records having irrelevant information, although no such record was identified. Finally, missing values were replaced by the mean of that variable for 25 adjacent cases in SPSS 21 so that the number of the remaining records were 2860 (unchanged). Then, by random sampling, 70% and 30% of the data set were used for model training and model testing, respectively. In order to design a multilayer perceptron neural network with Rapid Miner 5.2, the number of nodes was considered 14 with a learning rate of 0.3 and 1 hidden layer. The number of nodes in hidden layer was 10, and the number of iteration was considered 1000. To design the tree with Rapid Miner 5.2, data productivity criteria, minimum branch size of 4, minimum leaf size of 2, minimum productivity of 0.1, and confidence of 0.25 were used, and models were evaluated using 70% of the training data and 30% of the test data.

Table . Breast cancer risk factors considered in the study

Risk factor	Type
1.The age at the time of diagnosis	Quantitative – discrete
2.The age of the first menstruation	Quantitative – discrete
3.Menopausal age	Quantitative – discrete
4.The age of the first pregnancy	Quantitative – discrete
5.History of breastfeeding	Qualitative – classified
6.History of taking OCP	Qualitative – classified
7.History of hormone therapy after menopause	Qualitative – classified
8.History of breast cancer	Qualitative – classified
9.Family history of breast cancer	Qualitative – classified
10.History of infertility	Qualitative – classified
11.Tobacco use	Qualitative – classified
12.Marital status	Qualitative – classified
13.Education	Qualitative – classified
14.Traumatic events in life	Qualitative – classified
15.Type of disease (malignant or benign)	Qualitative – classified

For investigation of the success rate and efficacy of these models, we used confusion matrix and ROC diagram as common techniques in diagnosis classification models.[12] For interpretation of classification and diagnosis of diseases and breast cancer patients using confusion matrix, there exists four states including true positive, true negative, false positive, and false negative, with each one having a special meaning in confusion matrix as follows:

True Positive (TP): the number of records that are positive and the algorithm has truly identified their class.

False Positive (FP): the number of records that are negative but the algorithm has falsely identified their class as positive.

True Negative (TN): the number of records that are negative and the algorithm has truly identified their class.

False Negative (FN): the number of records that are positive but the algorithm has falsely identified their class as positive.[13]

In this paper, the function of confusion matrix was developed using concepts above, and, for analyzing their functions, three main criteria of sensitivity, specificity, and accuracy in classification were used. Definitions and characteristics of these indices have been described in all resources for data analysis.[13]

As mentioned earlier, after being entered in Rapid Miner software, the data set was split into two sets (70% for training and 30% for testing models), and the multi-layer perceptron neural network (MLP) and the decision tree (C4.5) were trained and tested using those data sets. Results were provided by three criteria of accuracy, sensitivity and specificity.

Results

As described in previous section, after training and testing the models, the software reported results by three indices of sensitivity, specificity, and accuracy. Table 2 presents the results of the evaluation of models.

It can be seen from the table that there was no significant difference between sensitivity and accuracy indices. However, regarding specificity, neural network is significantly effective than decision tree.

Discussion

In this study, we analyzed a data set from ACECR Breast Cancer Research Center for diagnosis of breast cancer using multilayer perceptron neural network and decision tree algorithms. In comparison, the neural network was significantly effective in diagnosing negative cases. Early diagnosis of breast cancer is important from different aspects and can improve patients' survival. Considering the importance of the risk factors in breast cancer incidence, the efficacy of data analysis techniques in development of effective models for prediction and diagnosis is undeniable.[9] It is worth mentioning that the use of the general terms "neural network" and "decision tree" do not seem appropriate for other algorithms like C4.5 and multilayer neural networks. In other words, "neural network" and "decision tree" are general terms for techniques which contain various algorithms.

Researchers have used neural network and decision tree algorithms with other breast cancer data sets, and the results are different from the present study. For instance, in a work by Senturk and Kara using neural network and decision tree algorithms for analysis of Wisconsin sampling data set, the accuracy of both models was greater than that of the present study. The reason for this difference can be attributed to the difference in databases, methods, and missing data management.[14]

In another work, Rajesh and Anand used C4.5 algorithm for analysis of SEER data set for diagnosing breast cancer, which displayed greater accuracy compared with the present study. Again, the difference can be attributed to differences in data sets, data selection, and data classification methods.[15] In a work by Lakshmi *et al* evaluating the efficacy of data analysis algorithms, Wisconsin sampling data set was analyzed using C4.5 algorithm. The accuracy of this model was significantly higher than our study because of the difference in the evaluation method. Therefore, differences in data sets can produce different results in data analysis.[16] Kiani and Atashi used decision tree algorithm for prediction of breast cancer recurrence. Similarities of these two studies are using decision tree, using a real sample of patients, and using similar outcomes for evaluation of the models. The researchers showed 75% accuracy for decision tree model, which is lower than that for our study. The most important reason for this difference may be that Kiani and Atashi used a lower number of records and different set of dependent variables.[17] Also, it is possible that increasing the amount of training data to a specific level may improve model accuracy.[10] Furthermore, Tolooi *et al* used C5 decision tree for analysis of the same data set used in this study and obtained an accuracy of 95%.

Table 2 . The results of model testing by sensitivity, specificity, and accuracy of models

Model	Sensitivity	Specificity	Accuracy
Multilayer perceptron neural network	90.88%	88.91%	80.52%
C4.5 decision tree	89.32%	80.97%	80.98%

This shows that significant differences in data modeling can be found among decision tree algorithms.

Among limitation of this study, we can mention the missing data. Because of independency among variables and the lack of specific order in them, missing data were estimated using replacement methods, which may have affected the results. Another limitation was the use of only one of the various available methods in neural network and decision tree, so it was not possible to identify the most effective algorithm among these algorithms. However, the main strength of this study was using a real-world data set of patients consisting of a large number of records, which improves system training and is relatively better than other regional studies in this context.

Considering one of the main purposes of medical data analysis, which is to produce the best algorithm for data description, the results of analyses of data sets are unique based on the method applied in every study, so the results are only valid for that specific method. On the other hand, a more complete list of risk factors can provide a model with more extensive coverage. Moreover, results of the models may be affected by data preprocessing and missing data handling, and the method used for data evaluation. Researchers can use the results of the present study for future analyses of breast cancer risk factor data sets to generate models with higher efficacy and accuracy. It is suggested that future studies compare separate modeling results in decision tree with different numbers of iterations, investigate results with different neural network indices, and compare more algorithms-specially SVM, due to its promising results in medicine.

Conflict of Interest

The authors have none to declare.

References

1. Rezvani B. The relationship between TRU9I polymorphism in vitamin D receptor (VDR) gene and breast cancer. MSc thesis: Islamic Azad University, Damghan branch; 2013.
2. Setayeshi S, Akbari ME, Darghahi R, Haghighat khah HR. Breast Cancer and Technical Analysis of its Diagnostics. Tehran: Bitarafan; 2011.
3. American Cancer Society. Breast cancer facts & figures 2009 [Available from: https://www.cancer.org/ research/cancer-facts-statistics/breast-cancer-facts-figures.html.]
4. National Breast Cancer Foundation. Early Detection 2012 [Available from: http://www.nationalbreastcancer.org/ earlydetection-of-breast-cancer]
5. Khoury-Collado F, Bombard AT. Hereditary breast and ovarian cancer: what the primary care physician should know. Obstetrical & gynecological survey. 2004;59(7):537-42.
6. Claus EB, Risch N, Thompson WD. Autosomal dominant inheritance of early-onset breast cancer Implications for risk prediction. Cancer. 1994; 73:643- 651.
7. Shell J, Gregory WD. Efficient Cancer Detection Using Multiple Neural Networks. IEEE journal of translational engineering in health and medicine. 2017;5:1-7.
8. Udayakumar E, Santhi S, Vetrivelan P. An investigation of Bayes algorithm and neural networks for identifying the breast cancer. Indian journal of medical and paediatric oncology: official journal of Indian Society of Medical & Paediatric Oncology. 2017;38(3):340.
9. Azimian F, Tadaion-T GH, Jalali M. Breast Cancer Detection Using Data Mining Techniques. 4th Iranian Data Mining Conference2010.
10. Grzymala-Busse JW, Hu M. A comparison of several approaches to missing attribute values in data mining. InInternational Conference on Rough Sets and Current Trends in Computing 2000: 378-385. Springer, Berlin, Heidelberg.
11. Hota H. Diagnosis of breast cancer using intelligent techniques. International Journal of Emerging Science and Engineering (IJESE). 2013;1(3):45-53.
12. Fielding A. Cluster and classification techniques for the biosciences. 2007.
13. Larose DT. Data mining methods & models: John Wiley & Sons; 2006.
14. Senturk ZK, Kara R. Breast cancer diagnosis via data mining: performance analysis of seven different algorithms. Computer Science & Engineering. 2014;4(1):35.
15. Rajesh K, Anand S. Analysis of SEER dataset for breast cancer diagnosis using C4. 5 classification algorithm. International Journal of Advanced Research in Computer and Communication Engineering. 2012;1(2):2278-1021.
16. Lakshmi K, Krishna MV, Kumar SP. Performance comparison of data mining techniques for prediction and diagnosis of breast cancer disease survivability. Asian Journal of Computer Science & Information Technology. 2013;3(5):81 - 87.
17. Kiani B, Atashi A. A prognostic model based on data mining techniques to predict breast cancer recurrence. Journal of Health and Biomedical Informatics. 2014;1(1):26-31.