Journal of
Health & Medical Informatics

Research Article Open Access

Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence

Ahmad LG*, Eshlaghy AT, Poorebrahimi A, Ebrahimi M and Razavi AR

Department of Management Information Systems, Science and Research Branch, Islamic Azad University of Tehran-Iran, Iran

Abstract

Objective: The number and size of medical databases are increasing rapidly but most of these data are not analyzed for finding the valuable and hidden knowledge. Advanced data mining techniques can be used to discover hidden patterns and relationships. Models developed from these techniques are useful for medical practitioners to make right decisions. The present research studied the application of data mining techniques to develop predictive models for breast cancer recurrence in patients who were followed-up for two years.

Method: The patients were registered in the Iranian Center for Breast Cancer (ICBC) program from 1997 to 2008. The dataset contained 1189 records, 22 predictor variables, and one outcome variable. We implemented machine learning techniques, i.e., Decision Tree (C4.5), Support Vector Machine (SVM), and Artificial Neural Network (ANN) to develop the predictive models. The main goal of this paper is to compare the performance of these three well-known algorithms on our data through sensitivity, specificity, and accuracy.

Results and Conclusion: Our analysis shows that accuracy of DT, ANN and SVM are 0.936, 0.947 and 0.957 respectively. The SVM classification model predicts breast cancer recurrence with least error rate and highest accuracy. The predicted accuracy of the DT model is the lowest of all. The results are achieved using 10-fold cross-validation for measuring the unbiased prediction accuracy of each model.

Keywords:

Classification; Decision tree; Machine learning; Support vector machine; 10-Fold cross-validation

Introduction

Breast cancer (BC) is the most common cancer in women, affecting about 10% of all women at some stages of their life. In recent years, the incidence rate keeps increasing and data show that the survival rate is 88% after five years from diagnosis and 80% after 10 years from diagnosis [1]. Early prediction of breast cancer is one of the most crucial works in the follow-up process. Data mining methods can help to reduce the number of false positive and false negative decisions [2,3]. Consequently, new methods such as knowledge discovery in databases (KDD) has become a popular research tool for medical researchers who try to identify and exploit patterns and relationships among large number of variables, and predict the outcome of a disease using historical cases stored in datasets [4].

In this paper, using data mining techniques, authors developed models to predict the recurrence of breast cancer by analyzing data collected from ICBC registry. The next sections of this paper review related work, describe background of this study, evaluate three classification models (C4.5 DT, SVM, and ANN), explain the methodology used to conduct the prediction, present experimental results, and the last part of the paper is the conclusion. To estimate validation of the models, accuracy, sensitivity, and specificity were used as criteria, and were compared.

Literature review and previous works

A literature review showed that there have been several studies on the survival prediction problem using statistical approaches and artificial neural networks. However, we could only find a few studies related to medical diagnosis and recurrence using data mining approaches such as decision trees [5,6]. Delen et al. used artificial neural networks, decision trees and logistic regression to develop prediction models for breast cancer survival by analyzing a large dataset, the SEER cancer incidence database [6]. Lundin et al. used ANN and logistic regression models to predict 5, 10, and 15 -year breast cancer survival. They studied 951 breast cancer patients and used tumor size, axillary nodal status, histological type, mitotic count, nuclear pleomorphism, tubule formation, tumor necrosis, and age as input variables [7]. Pendharker patterns in breast cancer. In this study, they showed that data mining could be a valuable tool in identifying similarities (patterns) in breast cancer cases, which can be used for diagnosis, prognosis, and treatment purposes [4]. These studies are some examples of researches that apply data mining to medical fields for prediction of diseases.

Materials and Methods

In order to predict the 2-year recurrence rate of breast cancer, we used ICBC dataset in the National Cancer Institute of Tehran for the years 1997-2008. The ICBC is responsible for collecting incidence and survival data from the participating registries, and disseminating these datasets for the purpose of conducting analytical research projects. This dataset contained population characteristics and included 22 input variables. Our cases were collected from the total number of 1189 women that were diagnosed breast cancer. We preprocessed the data to remove unsuitable cases. After using data cleansing and data preparation strategies, the final dataset was constructed. Finally, 547 cases were analyzed after 642 records were excluded because of missing data. Patients with breast cancer recurrence were followed-up

***Corresponding author:** Leila Ghasem Ahmad, Department of Management Information Systems, Science and Research Branch, Islamic Azad University of Tehran-Iran, Iran, E-mail: lga_77@yahoo.com

Received January 28, 2013; **Accepted** April 18, 2013; **Published** April 24, 2013

Citation: Ahmad LG, Eshlaghy AT, Poorebrahimi A, Ebrahimi M, Razavi AR (2013) Using Three Machine Learning Techniques for Predicting Breast Cancer Recurrence. J Health Med Inform 4: 124. doi:10.4172/2157-7420.1000124

years. The independent variables that we used are shown in table 1. The dataset was cleaned by handling missing values, noise, identifying and correcting inconsistencies. Some fields, such as Her2, age of menarche, and Npositive, contained missing values. Since these variables are important in predicting recurrence, the records containing the missing data were removed from the dataset. Missing values for continuous variables were substituted using EM method [8].

Expectation maximization (EM)

The EM algorithm is a method for efficient estimation from incomplete data. In any incomplete dataset, there is indirect evidence about the likely values of the unobserved values. This evidence, when combined with some assumptions, comprises a predictive probability distribution for the missing values that should be averaged in the statistical analysis. The EM algorithm is a common technique for matching models to incomplete data. EM is important on the relationship between missing data and unknown parameters of a model. When the parameters are known, then it is possible to obtain impartial predictions for the missing values [8,9].

Data mining techniques

In this paper, we used DT, SVM, and ANN machine learning algorithms to predict the recurrence of breast cancer to find which method performs better. For DT, C4.5 algorithm which is based on the ID3 algorithm was used. Each tree node is either a leaf node or decision node. All decision nodes have splits, testing the values of some functions of data attributes. Each branch from the decision node corresponds to a different outcome of the test. Each leaf node has a class label attached to it. Weka software was implemented to analyze the data with C4.5. It is an open source data mining tool and offers many data mining algorithms including AdaBoost, Bagging, C4.5 and SVM. It is a collection of tools for data classification, regression, clustering, association rules, and visualization [10].

Support vector machine (SVM) is an emerging powerful machine learning technique to classify cases. SVM has been used in a range of problems and they have already been successful in pattern recognition in bioinformatics, cancer diagnosis [11], and more. Figure 1 shows SVM topology in hyperspace.

SVM is a maximum margin classification algorithm rooted in statistical learning theory. It is the method for classifying both linear and non-linear data. It uses a non-linear mapping technique to transform the original training data into a higher dimension. It performs classification tasks by maximizing the margin separating both classes while minimizing the classification errors [12].

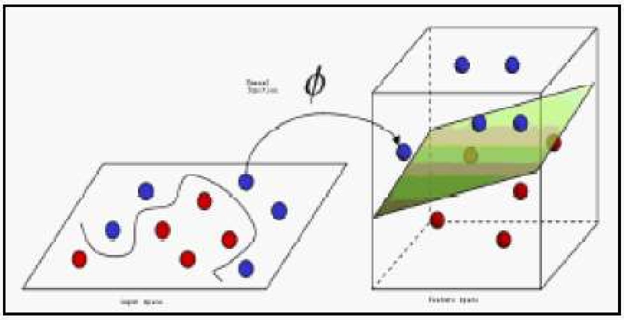

Figure 1: SVM Topology.

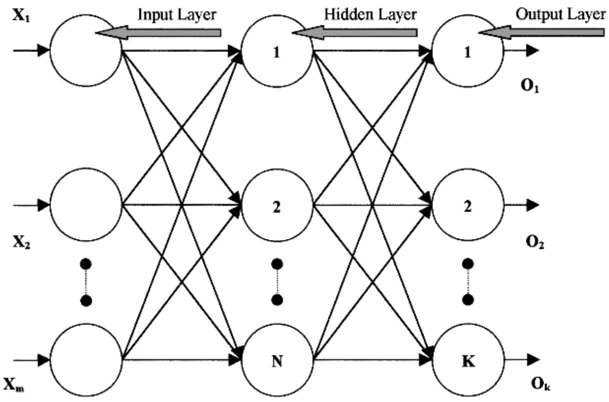

Figure 2: MLP feed forward Neural Network.

No	Variable Name	Definition
1.	Local Recurrence	Yes or No
2.	Age at Diagnosis	≤ 35, 35 to 44, 44-45, 55 ≥ years old
3.	Age at Menarche	≤ 12 to ≥ 12 years old
4.	Age at Menopause	≤ 50 to ≥ 50 years old
5.	Infertility	Yes or No
6.	Family History of Breast Cancer	Yes or No
7.	History of other Cancer (CA)	Yes or No
8.	Location	Upper outer Quadrant (UOQ), Upper inner Quadrant (UIQ), Lower outer Quadrant (LOQ), Lower inner Quadrant (LIQ), Central, Axilla, Upper half, Lateral half, Lower half
9.	Side	Left, Right, Bilateral
10.	Tumor Size	≤ 2cm to ≥ 5cm
11.	LN/Nexion	Lymph node involvement/number of removed nodes after surgery
12.	Metastasis	Bone, Liver, Lung, Brain, others
13.	NPositive	Number of Positive lymph node involvement
14.	B.Pathology	Results of Biopsy Pathology after
15.	Type of surgery	Mastectomy (Preservative or Bilateral)
16.	G (Grade)	1, 2 or 3
17.	Margin of Involvement	Free or ≥ 2cm
18.	Estrogen Receptor	Negative or Positive
19.	Progesterone Receptor	Negative or Positive
20.	Type of Chemotherapy	Adjuvant or Neoadjuvant
21.	Radiotherapy	Yes or No
22.	Hormone Therapy	Tamoxifen, Raloxifen, Femara, Aromasin or Megace
23.	Death	Realted to Breast Cancer or unrelated
24.	Her2	Negative or Positive

Table 1: Variables used for recurrence modeling of breast cancer.

Neural networks are a large number of interconnected nodes that perform summation and thresholding in loose analogy with the neurons of the brain. The multi-layer perceptron (MLP) model is capable of mapping set of input data into a set of appropriate output data. The primary task of neurons in input layer is the division of input signal X_i among neurons in hidden layer. The output of neurons in the output layer is determined in an identical fashion [13]. Figure 2 shows MLP feed forward Neural Network.

The back-propagation algorithm can be employed effectively to train neural networks; it is widely recognized for applications to layered feed-forward networks, or multi-layer perceptrons. MLP is the most commonly used algorithm and performs better than other ANN architectures for this type of classification problems [14].

The number of recurrence and non recurrence cases were been 117 and 430, respectively. In order to evaluate the validity of present results for making predictions regarding new data, 10-fold cross-validation was implemented in model building, evaluation, and comparison (We performed experiments using Weka, an open source data mining tool and the comparison is based on 10-fold cross-validation). In this method, each of the 10 subsets acts as an independent holdout test set for the model trained with the rest of the subsets. A pair of testing and training sets is called a ''fold".

Results and Discussion

To compare the models, the data from the Iranian Center for Breast Cancer dataset were analyzed. Table 2 shows the summaries of predictor variables. Table 3 shows sensitivity, specificity and accuracy for different classification techniques.

This paper has explored risk factors for predicting breast cancer by using data mining techniques. Each method has its own limitations and strengths specific to the type of application. Table 3 shows the accuracy, sensitivity, and specificity comparison of the data mining methods. Our results show that SVM outperforms both Decision Tree and MLP in all the parameters of sensitivity, specificity and accuracy. SVM is the best predictor of breast cancer recurrence. The results of decision tree C4.5 outperformed decision tree C4.5 and ANNs. Table 3 shows the accuracy, sensitivity, and specificity comparison of decision tree C4.5, SVM and ANNs.

There are some limitations in the current study. There were many cases lost in the follow-up and there were records with missing values that were omitted unfortunately. Some important variables such as S-phase fraction and DNA index were not included in the study because of their unavailability which may have decreased the performance of the models and also and there were some degree of missingness in our data.

However, these obtained results were based on a new database in Iranian Center for Breast Cancer, by comparison three different data mining methodology and also Weka toolkit.

Conclusion

There are different data mining techniques that can be used for the prediction of breast cancer recurrence. In this paper, researchers analyzed breast cancer data using three classification techniques to predict the recurrence of the cancer and then compared the results. The results indicated that SVM are the best classifier predictor with the test dataset, followed by ANN and DT. Further studies should be conducted to improve performance of these classification techniques by using more variables and choosing for a longer follow-up duration.

Variables	No. of missing records	Type of data	Mean	Std. Dev.	Range
Age at diagnosis	2	Numeric	46.44	11.09	18-87
Age at Menarche	38	Numeric	13.50	1.49	8-20
Age at Menopause	570	Numeric	47.14	5.56	32
Tumor Size	159	Numeric	2.97	1.57	0-15
LN involvement	58	Numeric	0.93	1.01	0- ≥9
Grade	284	Numeric	2.15	0.64	2
Nexion (Lymph node dissection)	360	Numeric	10.89	5.13	0-35
Her2	504	Numeric	0.39	0.48	+ or -

Table 2: Predictor Variables.

	Accuracy	Sensitivity	Specificity
Decision Tree (C4.5)	0.936	0.958	0.907
ANN (MLP)	0.947	0.956	0.928
SVM	0.957	0.971	0.945

Table 3: Comparison of data mining models.

References

1. (2003) Breast cancer facts and figures 2003-2004. American Cancer Society.
2. Karabatak M, Cevdet M (2009) An expert system for detection of breast cancer based on association rules and neural network. Expert Systems with Applications 36: 3465-3469.
3. Kovalerchuc B, Triantaphyllou E, Ruiz JF, Clayton J (1997) Fuzzy logic in computer-aided breast- cancer diagnosis: Analysis of lobulation. Artif Intell Med11: 75-85.
4. Pendharkar PC, Rodger JA, Yaverbaum GJ, Herman N, Benner M (1999) Association, statistical, mathematical and neural approaches for mining breast cancer patterns. Expert Systems with Applications 17: 223-232.
5. Zhou ZH, Jiang Y (2003) Medical diagnosis with C4.5 Rule preceded by artificial neural network ensemble. IEEE Trans Inf Technol Biomed 7: 37-42.
6. Delen D, Walker G, Kadam A (2005) Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine 34: 113-127.
7. Lundin M, Lundin J, Burke HB, Toikkanen S, Pylkkanen L, et al. (1999) Artificial neural networks applied to survival prediction in breast cancer. Oncology 57: 281-286.
8. Dempster AP, Laird NM, Rubin DB (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. J R Stat Soc Series B 39: 1-38.
9. Rubin DB, Schenker N (1991) Multiple Imputation in Health-Care Databases - an overview and some applications. Stat Med 10: 585-598.
10. Weka 3: Data Mining Software in Java.
11. Cristianini N, Shawe-taylor J (2000) An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, London: Cambridge University Press.
12. Joachims T (1998) Making large-scale support vector machine learning practical. Advances in Kernel Methods: Support Vector Learning. MIT Press, Cambridge, MA, 169-184.
13. Haykin S (1998) Neural networks: a comprehensive foundation. New Jersey: Prentice Hall, New Jersey, USA.
14. Hornik K, Stinchcombe M, White H (1990) Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. Neural Networks 3: 359-366.