Engineering Applications of Artificial Intelligence 26 (2013) 945–950

Contents lists available at SciVerse ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

A genetic algorithm model based on artificial neural network for prediction of the axillary lymph node status in breast cancer

R. Karakış [a,*], M. Tez [b], Y.A. Kılıç [c], Y. Kuru [d], İ. Güler [a]

[a] Department of Computer and Electronics Technology, Faculty of Technology, Gazi University, 06500 Teknikokullar, Ankara, Turkey
[b] Department of Surgery, Ankara Numune Educational and Research Hospital, Ankara, Turkey
[c] Department of Surgery, Hacettepe University Faculty of Medicine, Ankara, Turkey
[d] Department of Surgery, Ankara Oncology Hospital, Ankara, Turkey

ARTICLE INFO

Article history:
Received 29 February 2012
Received in revised form
16 October 2012
Accepted 17 October 2012
Available online 21 November 2012

Keywords:
Axillary lymph node
Breast cancer
Genetic algorithm
Multilayer perceptron
Logistic regression

ABSTRACT

Axillary Lymph Node (ALN) status is an extremely important factor to assess metastatic breast cancer. Surgical operations which may be necessary and cause some adverse effects are performed in determination ALN status. The purpose of this study is to predict ALN status by means of selecting breast cancer patient's basic clinical and histological feature(s) that can be obtained in each hospital. 270 breast cancer patients' data are collected from Ankara Numune Educational and Research Hospital and Ankara Oncology Educational and Research Hospital. These are classified using back propagation MultiLayer Perceptron (MLP), Logistic Regression (LR) and Genetic Algorithm (GA) based MLP models. Receiver Operating Characteristics (ROC) such as sensitivity, specificity, accuracy and area under of ROC (AUC) and regression are used to evaluate performances of the developed models. It is concluded from LR and GA based MLP, that menopause status and lymphatic invasion are the most significant features for determining ALN status. GA provides to select best features as MLP inputs. It also optimizes the weights of backpropagation algorithm in MLP. The values of regression and accuracy of the GA based MLP with 9 features (numerical age, categorical age, menopause status, tumor size, tumor type, tumor location, T staging, tumor grade and lymphatic invasion) are found as 0.96 and 98.0% with respectively. According to results, proposed GA based MLP classifier can be used to predict the ALN status of breast cancer without surgical operations.

1. Introduction

Breast cancer is one of the leading tumor-related causes of death in women (He et al., 2009). If cancer cells are present in Axillary Lymph Nodes (ALN), a risk for metastatic breast cancer rises. ALN may be removed by different types of lymph node dissection surgeries such as Sentinel Lymph Node Biopsy (SLNB) and Axillary Lymph Node Dissection (ALND). Research shows that more surgery operations may be necessary and cause some adverse events such as lymphoedema, restriction of arm and shoulder movement and numbness of upper arm skin (Karakis et al., 2011). If ALN status of breast cancer patients could be accurately predicted from basic clinical and histological features, surgical operations could be prevented (Patani et al., 2007; Karakis et al., 2011).

Previous researches use the statistical methods such as logistic regression, univariate and multivariate analysis in this area. It is showed that no single feature or combinations of features are sufficiently accurate to predict axillary status (Patani et al., 2007). For example, Harden et al. found that lymphvascular invasion and tumor size could be determined but tumor grade was not associated with ALN status (Harden et al., 2001). However, Farshid et al. did not find significance of tumor size and grade (Farshid et al., 2004). Hence, Artificial Neural Networks (ANN) have been proposed as a supplement or alternative to standard statistical techniques for prediction of axillary lymph node status in breast cancer patients (Tez et al., 2007; Karakis et al., 2011).

Artificial Neural Networks (ANNs) are a form of artificial intelligence that have been successfully applied to a variety of problems including pattern recognition, modeling, control and medical field (Papik et al., 1998; Haykin, 1999, 2000; Lisboaa and Taktak, 2006; Marchevsky, 2006; Paliwal and Kumar, 2009; Ferreira and Gil, 2012). However, there is a few artificial neural network studies in ALN status prediction (Naguib et al., 1996; Marchevsky et al., 1999; Seker et al., 2000, 2002, 2003; Mattfeldt et al., 2004; Lancashire et al., 2008; Karakis et al., 2011). In this field, the most striking results have been obtained by Marchevsky et al. (1999) and Lancashire et al. (2008).

* Corresponding author. Tel.: +09 03122028580.
E-mail address: rukiyekarakis@gazi.edu.tr (R. Karakış).

0952-1976/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.engappai.2012.10.013

Marchevsky et al. (1999) evaluated 19 prognostic features of 279 patients through probabilistic Neural Networks (NNs) based on genetic algorithms and logistic regression classifiers. Lancashire et al. (2008) applied the MLP to a gene microarray dataset that consists of 49 samples to identify gene signatures corresponding with estrogen receptor and ALN status in breast cancer. The high predictive accuracy values were obtained 89% and 100% respectively in these studies. However, the used gene protein data or flow-cytometric data are difficult to obtain in each hospital. Mattfeldt et al. (2004) used the histological (the age, tumor type, grade and size, skin infiltration, lymphangiosis carcinomatosa, pT4 category) and flow-cytometric (percentage of tumor cells in G2/M- and S-phases of the cell cycle, and ploidy index) features to predict ALN status. It also stated that basic features could be used as determiners. Nonetheless, the predictive accuracy of performed model was only 72%.

Previous study classified axillary lymph node status of 270 breast cancer patients through pattern recognition techniques such as MultiLayer Perceptron (MLP), Support Vector Machines (SVM), Linear Discriminant Analysis (LDA) and k-Nearest Neighbor (k-NN) (Karakis et al., 2011). MLP network structure that use Levenberg–Marquardt algorithm in back propagation training was consisted of a hidden layer with 4 neurons. MLP classifier found the correlation coefficient and the accuracy value of testing dataset as 0.872 and 94.4% respectively. However, none of the classifiers have the high accuracy value as features are selected randomly. Because, a general problem in MLP model selection is to obtain right parameters that fit a model with observed data. Hence, feature selection can be described as an optimization problem. The data without any feature selection might be redundant and may deteriorate. It also influences the efficiency of classification. Feature selection reduces computational cost, storage requirements, and training time. It also facilitates Back Propagation (BP) training process (Castillo et al., 2000; Lin et al., 2008).

Feature selection approaches can be categorized in two models. The filter model is based on statistical approaches and it is also quite fast technique. Nonetheless, the resulting feature subset may not be the best. The wrapper model uses some selection methods to choose feature subsets. It evaluates the result by means of the accuracy rates of classification algorithm calculates the accuracy rates. The wrapper model is widely used feature selection with Genetic Algorithm (GA) approaches in MLP (Lin et al., 2008). GA has been used with neural network to search for input features or to determine the number of nodes or connections in the network (Handels et al., 1999, Arifovica and Gencay, 2001; Palmes et al., 2005; Nicolás et al., 2006; Ambrogi et al., 2007; Benardos and Vosniakos (2007)).

This study has used only basic clinical and histological features of breast cancer patients. The data have been obtained from Ankara Numune Educational and Research Hospital and Ankara Oncology Educational and Research Hospital. The aims of our analysis are to identify potential feature(s) of ALN status and propose a classifier model by means of MLP, LR and GA based MLP.

The rest of this study is organized as follows. After the first section, material and method are described in Section 2. Section 3 presents results of developed models. It also includes a brief discussion of the obtained results. Finally Section 4 states the conclusion and future works.

2. Material and method

2.1. Dataset

The diagnosis of the breast cancer is based on the clinical, radiological, and pathological examinations. In clinical examination, patients' features including age, family health history, menstrual history, pregnancy age, previous breast illness history are gathered. Manual palpation of the breasts is examined by the expert doctor. If there is any suspected mass, radiological examination such as mammography, ultrasound is used. Magnetic Resonance Imaging (MRI) of the breasts is not a routine procedure. Conclusive diagnosis of breast cancer is examined biopsy (fine needle aspiration biopsy, core needle biopsy etc.). The World Health Organization (WHO) classification and the Tumor-Node-Metastases (TNM) staging system are used in pathological diagnosis (Patani et al., 2007; Kataja and Castiglione, 2009; Karakis et al., 2011).

In this study, clinical (numerical age-F1, categorical age-F2, menopause status-F3), radiological (tumor size-F4, tumor type-F5, tumor location-F6) and pathological features (T staging-F7, tumor grade-F8, lymphatic invasion-F9) are examined to determine of axillary lymph node status in breast cancer patients. They are encoded into numeric values given in Table 1.

The dataset consists of 270 samples of breast cancer patients who underwent modified radical mastectomy at Ankara Numune Educational and Research Hospital, and Ankara Oncology Educational and Research Hospital. 59% of patients have positive ALN. The dataset is divided into 80% training set and 20% testing set respectively. Matlab Neural Network Toolbox TM 6.0 and SPSS 13.0 statistical software are used for analysis.

2.2. Methods

This study proposes a hybrid genetic algorithm based approach, denoted as NeuralGenetic, to determine features as inputs and weights of backpropagation algorithm in MLP architectures. NeuralGenetic, LR and MLP models are compared through accuracy rates of testing dataset. The proposed classification model is shown in Fig. 1.

Table 1
Description of breast cancer features.

Features	Index	Properties	%
Axillary Lymph Node Status (ALNS)	–	0-No	59.3
		1-Yes	40.7
Age	F1	Numerical	100
Categorical age	F2	1-Age $<$ 50	53.3
		2-Age $\geq$ 50	46.7
Menopause status	F3	1-Premenopause	44.1
		2-Postmenopause	55.9
Tumor size	F4	Numerical	100
Tumor type	F5	1-Ductal	85.9
		2-Lobular	0.7
		3-Mixed	3.3
		4-Atypic medullar	4.0
		5-Private type	9.3
		6-Dcis	4.0
Tumor location	F6	1-Left/right UOQ*	57.8
		2-Left/right LOQ*	8.9
		3-Left/right UIQ*	24.1
		4-Left/right LIQ*	5.6
		5-Areol	3.7
T staging	F7	1-T1	25.9
		2-T2	62.6
		3-T3	11.5
Tumor grade	F8	1-G1	26.3
		2-G2	51.1
		3-G3	22.6
Lymphatic invasion	F9	0-No	70.4
		1-Yes	29.4

* UOQ=Upper Outer Quadrant, LOQ=Lower Outer Quadrant, UIQ=Upper Inner Quadrant, LIQ=Lower Inner Quadrant.