Expert Systems with Applications 36 (2009) 8247–8254

Contents lists available at ScienceDirect

Expert Systems with Applications

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

The comparisons of prognostic indexes using data mining techniques and Cox regression analysis in the breast cancer data

Mevlut Ture [a,*], Fusun Tokatli [b], Imran Kurt Omurlu [a]

[a] Trakya University, Medical Faculty, Department of Biostatistics, Edirne 22030, Turkey
[b] Trakya University, Medical Faculty, Department of Radiation Oncology, Edirne, Turkey

ARTICLE INFO

Keywords:
Decision tree
C&RT
CHAID
QUEST
ID3
C4.5
C5.0
Cox regression
Kaplan–Meier
Breast cancer
Disease-free survival
Random survival forests

ABSTRACT

The purpose of this study is to determine new prognostic indexes for the differentiation of subgroups of breast cancer patients with the techniques of decision tree algorithms (C&RT, CHAID, QUEST, ID3, C4.5 and C5.0) and Cox regression analysis for disease-free survival (DFS) in breast cancer patients. A retrospective analysis was performed in 381 breast cancer patients diagnosed. Age, menopausal status, age of menarche, family history of cancer, histologic tumor type, quadrant of tumor, tumor size, estrogen and progesterone receptor status, histologic and nuclear grading, axillary nodal status, pericapsular involvement of lymph nodes, lymphovascular and perineural invasion, adjuvant radiotherapy, chemotherapy and hormonal therapy were assessed. Based on these prognostic factors, new prognostic indexes for C&RT, CHAID, QUEST, ID3, C4.5 and C5.0 and Cox regression were obtained. Prognostic indexes showed a good degree of classification, which demonstrates that an improvement seems possible using standard risk factors. We obtained that C4.5 has a better performance than C&RT, CHAID, QUEST, ID3, C5.0 and Cox regression to determine risk groups using Random Survival Forests (RSF).

1. Introduction

The clinicopathologic characteristics of breast cancer patients are heterogeneous. Consequently, the survival times are different in subgroups of patients. Generally, 5-years recurrence-free survival is ranged from 65% to 80% in all population in breast cancer patients (Buchholz, Strom, & McNeese, 2003). The purposes of this study were to determine new prognostic indexes for the differentiation of subgroups of breast cancer patients with the various methods (decision trees and Cox regression analysis) and explore the interactions between clinical variables and their impact on survival.

Cheng et al. (2006) used Bayesian classification trees and Cox proportional hazard models were used to estimate the probability of local regional recurrence after mastectomy for individual breast cancer patients. Sauerbrei, Hübner, Schmoor, and Schumacher (1997) compared Cox regression analysis, Classification and Regression Tree (C&RT) and Nottingham Prognostic Index for determining new prognostic classification index in node negative breast cancer.

Decision tree algorithms allow for non-linear relations between predictive factors and outcomes and for mixed data types (numerical and categorical), isolates outliers, and incorporates a pruning process using cross-validation as an alternative to testing for unbiasedness with a second data set (Faderl et al., 2002).

Decision trees use recursive partitioning to assess the effect of specific variables on survival, thereby ultimately generating groups of patients with similar clinical features and survival times. The partitioning of patients into groups with differing survival times using clinical variables generates a tree-structured model that can be analyzed to assess its clinical utility. Therefore decision tree methods such as C&RT, Chi-squared Automatic Interaction Detector (CHAID), Quick, Unbiased, Efficient Statistical Tree (QUEST), Commercial version 5.0 (C5.0), Commercial version 4.5 (C4.5) and Interactive Dichotomizer version 3 (ID3) are more suitable than classical statistical methods.

In our previous study, we evaluated performance of C&RT, CHAID, QUEST, C4.5 and ID3 methods according to predictive values for disease-free survival (DFS) in breast cancer patients. We estimated DFS rates according to the decision tree method based on the C4.5 analysis. Then, according to multidimensional scaling method C4.5 performed slightly better than other methods in predicting risk factors for recurrence (Ture, Tokatli, & Kurt, 2008). In this study, we analyzed the simultaneous relationship among risk factors for breast cancer by C&RT, CHAID, QUEST, C4.5, ID3, C5.0 and Cox regression analysis. We purpose to determine new prognostic indexes for the differentiation of subgroups of breast cancer patients with the decision tree algorithms and Cox regression analysis using Kaplan–Meier analysis. Random Survival Forests (RSF) was used to choice the best method and prognostic index.

* Corresponding author. Tel.: +90 284 2357641/1631; fax: +90 284 2357652.
E-mail address: ture@trakya.edu.tr (M. Ture).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.10.014

2. Patients and methods

A retrospective analysis was performed in 381 breast cancer patients diagnosed between 1997 and 2007. In all patients' age, menopausal status, age of menarche, family history of cancer, histologic tumor type, quadrant of tumor, tumor size, estrogen and progesterone receptor status, histologic and nuclear grading according to Scarf–Bloom–Richardson criteria (Bloom & Richardson, 1957), axillary nodal status, pericapsular involvement of lymph nodes, lymphovascular and perineural invasion, adjuvant radiotherapy, chemotherapy and hormonal therapy were assessed and documented.

Descriptive statistics of clinical and pathologic data for the entire patient population was listed in Table 1. We performed the classical statistical analysis to examine the differences in the distribution of variables between patients who had recurrence or not. The Kolmogorov Smirnov test was used to assess the normality of numeric variables. For all the numeric variables that were non-normally distributed, comparison between two groups was made by the Mann–Whitney U-test and results were expressed as median and interquartile range. Association of recurrence with nominal variables was assessed using the chi-square test.

New prognostic indexes which were solely based on standard factors were developed using C&RT, CHAID, QUEST, ID3, C4.5, C5.0 and Cox regression analysis. A 10-fold cross-validation analysis was performed as an initial evaluation of the test error of the decision tree algorithms. Briefly, this process involves splitting up the dataset into 10 random segments and using 9 of them for training and the 10th as a test set for the algorithm.

For the terminal nodes of the C&RT, CHAID, QUEST, ID3, C4.5, C5.0 and the Cox model, survival curves of DFS were estimated by the Kaplan–Meier method and the difference between the curves was evaluated by Log-Rank test. Follow-up time for each patient was calculated in months from the last day of the initial treatment to the date of death or the date of last visit. RSF with Log-Rank splitting rule was used to choice the best method and prognostic index.

2.1. Decision tree algorithms

2.1.1. Classification and regression tree (CART)

CART is a recursive partitioning method to be used both for regression and classification. CART is constructed by splitting subsets of the data set using all predictor variables to create two child nodes repeatedly, beginning with the entire data set. The best predictor is chosen using a variety of impurity or diversity measures. The goal is to produce subsets of the data which are as homogeneous as possible with respect to the target variable (Breiman, Friedman, Olshen, & Stone, 1984).

2.1.2. Chi-squared Automatic Interaction Detector (CHAID)

CHAID method is based on the chi-square test of association. A CHAID tree is a decision tree that is constructed by repeatedly splitting subsets of the space into two or more child nodes, beginning with the entire data set (Michael & Gordon, 1997). To determine the best split at any node, any allowable pair of categories of the predictor variables is merged until there is no statistically significant difference within the pair with respect to the target variable.

Table 1
Clinical and laboratory characteristics of the study groups.

Independent variables		Recurrence		p
		Absent (n = 264)	Present (n = 117)	
Age (years-old) median (IQR)		48 (14)	48 (19)	0.603
Age of Menarche (years-old) median (IQR)		13 (1)	13 (2)	0.701
Tumor Size (cm) median (IQR)		2.9 (2)	3.5 (3.5)	<0.001
		n (%)	n (%)	
Menopausal status	Post	114 (43.2)	57 (48.7)	0.316
	Pre	150 (56.8)	60 (51.3)	
Nuclear grade	I–II	191 (72.3)	73 (62.4)	0.052
	III	73 (27.7)	44 (37.6)	
Estrogen receptor status	Negative	64 (24.2)	40 (34.2)	0.044
	Positive	200 (75.8)	77 (65.8)	
Progesterone receptor status	Negative	63 (23.9)	46 (39.3)	0.002
	Positive	201 (76.1)	71 (60.7)	
Adjuvant Radiotherapy	Absent	56 (21.2)	43 (36.8)	0.001
	Present	208 (78.8)	74 (63.2)	
Chemotherapy	Absent	35 (13.3)	9 (7.7)	0.163
	Present	229 (86.7)	108 (92.3)	
Hormonal therapy	Absent	52 (19.7)	44 (37.6)	<0.001
	Present	212 (80.3)	73 (62.4)	
Family history of cancer	Absent	183 (69.3)	77 (65.8)	0.063
	Breast cancer	64 (24.2)	24 (20.5)	
	Other cancers	17 (6.4)	16 (13.7)	
Perineural invasion	Absent	187 (70.8)	64 (54.7)	0.002
	Present	77 (29.2)	53 (45.3)	
Lymphovascular Invasion	Absent	127 (48.1)	34 (29.1)	0.001
	Present	137 (51.9)	83 (70.9)	
Axillary nodal status	Negative	118 (44.7)	22 (18.8)	<0.001
	1–3 Lymph nodes positive	78 (29.5)	30 (25.6)	
	$\geqslant$4 Lymph nodes positive	68 (25.8)	65 (55.6)	
Histologic grade	I–II	169 (64.0)	58 (49.6)	0.008
	III	95 (36.0)	59 (50.4)	
Histologic tumor type	Ductal	220 (83.3)	104 (88.9)	0.212
	Non-ductal	44 (16.7)	13 (11.1)	
Quadrant of tumor	Unicentric	246 (93.2)	91 (77.8)	<0.001
	Multicentric	18 (6.8)	26 (22.2)	
Pericapsular involvement of lymph nodes	Negative	187 (70.8)	52 (44.4)	<0.001
	Positive	77 (29.2)	65 (55.6)	