Artificial Intelligence in Medicine (2005) **34**, 113–127

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

Predicting breast cancer survivability: a comparison of three data mining methods

Dursun Delen*, **Glenn Walker**, **Amit Kadam**

Department of Management Science and Information Systems, Oklahoma State University, 700 North Greenwood Venue, Tulsa, OK 74106, USA

Received 13 January 2004; received in revised form 30 June 2004; accepted 15 July 2004

KEYWORDS
Breast cancer survivability;
Data mining;
k-Fold cross-validation;
SEER

Summary

Objective: The prediction of breast cancer survivability has been a challenging research problem for many researchers. Since the early dates of the related research, much advancement has been recorded in several related fields. For instance, thanks to innovative biomedical technologies, better explanatory prognostic factors are being measured and recorded; thanks to low cost computer hardware and software technologies, high volume better quality data is being collected and stored automatically; and finally thanks to better analytical methods, those voluminous data is being processed effectively and efficiently. Therefore, the main objective of this manuscript is to report on a research project where we took advantage of those available technological advancements to develop prediction models for breast cancer survivability.

Methods and material: We used two popular data mining algorithms (artificial neural networks and decision trees) along with a most commonly used statistical method (logistic regression) to develop the prediction models using a large dataset (more than 200,000 cases). We also used 10-fold cross-validation methods to measure the unbiased estimate of the three prediction models for performance comparison purposes.

Results: The results indicated that the decision tree (C5) is the best predictor with 93.6% accuracy on the holdout sample (this prediction accuracy is better than any reported in the literature), artificial neural networks came out to be the second with 91.2% accuracy and the logistic regression models came out to be the worst of the three with 89.2% accuracy.

Conclusion: The comparative study of multiple prediction models for breast cancer survivability using a large dataset along with a 10-fold cross-validation provided us with an insight into the relative prediction ability of different data mining methods. Using sensitivity analysis on neural network models provided us with the prioritized importance of the prognostic factors used in the study.

* Corresponding author. Tel.: +1 918 594 8283; fax: +1 918 594 8281.
E-mail address: delen@okstate.edu (D. Delen).

0933-3657/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2004.07.002

1. Introduction

Breast cancer is a major cause of concern in the United States today. At a rate of nearly one in three cancers diagnosed, breast cancer is the most frequently diagnosed cancer in women in the United States. The American Cancer Society projected that 211,300 invasive and 55,700 in situ cases would be diagnosed in 2003 [1]. Furthermore, breast cancer is the second leading cause of death for women in the United States, and is the leading cause of cancer deaths among women ages 40–59 [1,2]. According to The American Cancer Society 39,800 breast cancer related deaths are expected in 2003 [2]. Though predominantly in women, breast cancer can also occur in men. In the United States, of the 40,600 deaths from breast cancer in 2001, 400 were men [3]. Even though in the last couple of decades, with increased emphasis towards cancer related research, new and innovative methods for early detection and treatment have been developed, which helped decrease the cancer related death rates [4–6], cancer in general and breast cancer in specific is still a major cause of concern in the United States.

Although cancer research is generally clinical and/or biological in nature, data driven statistical research is becoming a common complement. In medical domains where data and statistics driven research is successfully applied, new and novel research directions are identified for further clinical and biological research. For instance, Dr. John Kelsoe of the University of California, San Diego, demonstrated through his research study that a flawed gene appeared to promote manic-depression [7]. His data driven study found statistical evidence to tie the gene to the disease, and now researchers are looking for biological and clinical evidence to support his theory.

Predicting the outcome of a disease is one of the most interesting and challenging tasks in which to develop data mining applications. Survival analyses is a field in medical prognosis that deals with application of various methods to historic data in order to predict the survival of a particular patient suffering from a disease over a particular time period. With the increased use of computers powered with automated tools, storage and retrieval of large volumes of medical data are being collected and are being made available to the medical research community who has been interested in developing prediction models for survivability. As a result, new research avenues such as knowledge discovery in databases (KDD), which includes data mining techniques, has become a popular research tool for medical researchers who seek to identify and exploit patterns and relationships among large number of variables, and be able to predict the outcome of a disease using the historical cases stored within datasets [8,9].

It is the combination of the serious effects of breast cancer, the promising results of prior related research, the potential benefits of the research outcomes and the desire to further understand the nature of breast cancer that provided the motivation for this research effort. In this paper, we report on our research project where we developed models that predict the survivability of diagnosed cases for breast cancer. One of the salient features of this research effort is the authenticity and the large volume of data processed in developing these survivability prediction models. We used the SEER cancer incidence database, which is the most comprehensive source of information on cancer incidence and survival in the United States [2]. We used three different types of classification models: artificial neural network (ANN), decision tree, and logistic regression along with a 10-fold cross-validation technique to compare the accuracy of these classification models.

1.1. Breast cancer

Breast cancer is a malignant tumor that develops when cells in the breast tissue divide and grow without the normal controls on cell death and cell division [2,3]. It is the most common cancer among women [1]. Although scientists do not know the exact causes of most breast cancer, they do know some of the risk factors that increase the likelihood of a woman developing breast cancer. These factors include such attributes as age, genetic risk and family history [3].

Treatments for breast cancer are separated into two main types, local and systematic. Surgery and radiation are examples of local treatments whereas chemotherapy and hormone therapy are examples of systematic therapies. Usually for the best results, the two types of treatment are used together [2]. Although breast cancer is the second leading cause of cancer death in women, the survival rate is high. With early diagnosis, 97% of women survive for 5 years or more [3,10].

1.2. Knowledge discovery in databases and data mining

The amount of data being collected and stored in databases (both in medical and in other fields) has increased dramatically due to the advancements in software capabilities and hardware tools that enabled the automated data collection (along with