Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 34 (2008) 108–118
Expert Systems with Applications
www.elsevier.com/locate/eswa

The feasibility of constructing a Predictive Outcome Model for breast cancer using the tools of data mining

Thora Jonsdottir [a,*], Ebba Thora Hvannberg [b], Helgi Sigurdsson [a], Sven Sigurdsson [b]

[a] Cancer Center for Research and Development, Landspitali-University Hospital, Kopavogsbraut 5-7, 105 Kopavogur, Iceland
[b] University of Iceland, VR-II, Hjardarhaga 2-6, 107 Reykjavik, Iceland

Abstract

A *Predictive Outcome Model* (POM) for breast cancer was built, and its ability to accurately predict the (5 year) outcome of an incidence of cancer was assessed. A wide range of different feature selection and classification methods were applied in order to find the best performing algorithms on a given dataset. A special *Model Selection Tool*, MST, was developed to facilitate the search for the most efficient classifier model. The MST includes programs for *choosing different classification algorithms*, *selecting subsets of features*, *dealing with imbalance in the data* and *evaluating the predictive performance by various measures*. These steps are important in most data mining tasks and it would be time consuming to conduct them manually. The dataset, *Rose*, was assembled retroactively for this study and contains data records from 257 women diagnosed with primary breast cancer in Iceland during the years 1996–1998. An extra feature, containing the risk assessment of a doctor was added to the dataset which initially contained 400 features, both to see how much that could enhance the performance of the model and to investigate to what extent such a subjective assessment can be predicted from the remaining features. The main result is that similar performance is achieved regardless of which algorithm is used. Furthermore, the inclusion of the doctor's assessment does not appear to significantly enhance the performance. That is also reflected in the fact that the models are in general more successful in predicting the doctors risk assessment than the actual outcome if resulting Kappa values are compared.

Keywords: Data mining; Feature selection; Breast cancer; Classification; Accuracy

1. Introduction

The process of data mining is to find patterns and relationships in the data. A relatively large amount of data on cancer patients has been collected over the last few years, but the results of data mining are directly affected by the quantity and quality of the data. The Nobel price winner Joshua Lederberg stated: "*Data are the building blocks of knowledge and the seeds of discovery. They challenge us to develop new concepts, theories, and models to make sense of the patterns we see in them*" (Lederberg, 1999). Successful application of data mining to cancer patient data can result in new knowledge which can assist in cancer diagnosis and in the choice of treatment.

More than one million people were diagnosed worldwide with breast cancer in the year 2000, according to the International Agency for Research on Cancer's (IARC) extensive databases (Ferlay, Bray, Pisani, & Parkin, 2004). The number of new cases has been increasing for the last few decades, especially in the western part of the world. In Iceland, the number of newly diagnosed patients has been steadily increasing since 1958, whilst the number of patients dying of breast cancer has remained nearly the same (Jonason & Tryggvadottir, 2004). Survival at five years after the initial diagnosis has changed from being less than 50% during the years 1959–1963 to about 85% during the years 1994–1998, making the prognosis of patients with breast cancer one of the best among all cancers.

The purpose of this study was to build a Predictive Outcome Model (POM) that could accurately classify newly diagnosed patients into either of the following two

* Corresponding author. Tel.: +354 543 6901.
E-mail addresses: thora@landspitali.is, thora@lsh.is (T. Jonsdottir).

0957-4174/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2006.08.029

classes: no-event and recurrence-event of cancer, five years after diagnosis. The research is based on a data set *Rose*, which was assembled in cooperation with the Cancer Centre of Research and Development at the University Hospital in Iceland, during the years 1996–1998. The Rose database includes a relative small number of instances (257) but a large number of features (400). The features that could be used for this study had to be facts collected from the time when the patient was first diagnosed and treated.

In clinical practice, patients are classified into risk groups when diagnosed with breast cancer. It was therefore of interest to be able to conduct an experiment where the result of the classifier model would be compared to the results obtained from the clinical practice. This resulted in the introduction of a new three-valued feature named Risk to the Rose dataset. The Risk feature was the estimate, evaluated by a doctor, of the risk for a newly diagnosed patient to show marks of the disease within five years of diagnosis. The medical doctor grouped the patients into three risk groups: high, intermediate or low risk of recurrence. It was expected that the performance of the POM could be improved by adding this feature to the data set. A secondary objective was to use the specialist's risk estimate for each patient as a class attribute to see whether the POM could simulate the pattern implicitly used by the medical doctor for estimating this risk.

For the prediction, a POM was built from a training set of instances, whereas each instance was characterized by some set of given features. Building a valid and reliable POM can both be difficult and time consuming. Firstly, the modeller needs to clean the data, and select the most appropriate features and class attribute. Secondly, the data instances which the learning process will be based upon have to be chosen, and a learning method has to be selected from a range of algorithms currently available. Finally, the modeller needs to assess the reliability of the obtained results. Herein, a Model Selection Tool (MST) was constructed in order to ease the process of building an effective POM. As already mentioned, the resulting POM was to be used to predict the five years outcome for a newly diagnosed breast cancer patient using appropriate information about the patient. This tool was implemented on top of the data mining package WEKA (Witten & Frank, 2000).

An important aspect of the study was to gain better understanding of the relative importance of the features included and to prepare the data for the data mining task. Specific questions that were addressed are how many and what type of features have to be selected to reach a satisfactory prediction, whether there is a learning algorithm that is significantly better than others for this type of data, and whether a subjective evaluation from a doctor has a marked influence on the results.

2. Brief review of related work

Lee and co-workers (1999) used a linear support vector machine (SVM) to extract 6 out of 31 features from a data set including 253 breast cancer patients. The data set, WPBCC (Wisconsin Prognostic Breast Cancer Chemotherapy database), is publicly available (Wolberg, Lee, & Mangasarian, 1999) and contains features which were obtained before and during surgery. Their classification was based on dividing the patients into those with node-negative disease (no lymph nodes metastases) and node-positive disease (with lymph node metastases) patients. The patients were clustered into three prognostic groups: good (node-negative), intermediate (1–4 lymph node metastases) and poor (more than 4 lymph node metastases), whereas each group had a distinct survival curve. Based on the 6 selected features, the model could be used to assign new patients to one of the three prognostic groups with its associated survival curves. In 2003, Lee, Mangasarian, and Wolberg (2003) improved their cluster selection method even further, resulting in a classifier that could classify the breast cancer instances into the same three survival categories with 82.7% accuracy.

Fung, Mangasarian, and Shavlik (2001) carried out numerical tests on the WPBCC data set with a 60-month cut off for predicting recurrence or no recurrence of breast cancer. They used a support vector machine and reported a test accuracy of 66.2% using 10-fold cross validation.

Another publicly available data set containing information about breast cancer is the "Breast Cancer Data Set" (Zwitter & Soklic, 1988) originally collected at the University Medical Center, Institute of Oncology, Ljubljana, Slovenia. The outcome classes were no-event and recurrence-event. The accuracy obtained was between 62% and 78% using different classification methods (Holte, 1993). Tsai, King, and Higgins (1997) used an expert-guided decision tree to classify this data set. Their result (maximum 71.4% accuracy) indicated that an expert-guided approach was comparable to the optimal inductive learning approach. Both the WBPCC and the "Breast Cancer Data Set" have been widely used for comparing different classifiers.

Pendharkar, Khosrowpour, and Rodger (2000) applied a Bayesian network classifier and a data envelopment analysis (DEA) to a dataset collected from breast cancer patients in a large hospital in Pennsylvania, in order to discover patterns in the data. The result of their experiments indicated that DEA and Bayesian network classifier outperform statistical linear discriminates analysis. Pendharkar, Rodger, and Yaverbaum (1999) had previously shown that data mining can be used for breast cancer diagnosis.

Delen, Walker, and Kadam (2005) reported a 91.2–93.6% accuracy using artificial neural networks and decision tree respectively on a large dataset containing more than 200,000 cases collected from the years 1973–2000. This prediction accuracy is the best reported in the literature so far, and the size of this dataset is about 1000 times bigger than the size of other datasets reported. The outcome classes were defined as "any incidence of breast cancer where the person is still living after 60 months (5 years) from the date of diagnosis."