Application of Data Mining Techniques for Breast Cancer Prognosis

M. Mehdi Owrang O.
American University, USA

INTRODUCTION

Breast cancer is a malignant cancer causing tumor that begins when cells in the breast tissue grow abnormally, without managing cell division and cell death rates. Breast cancer is the most common female cancer in the US, the second most common cause of cancer death in women, and the main cause of death in women ages 40 to 59 (Costanza & Chen, 2012). Approximately 232,340 new cases of invasive breast cancer are expected to be diagnosed in the United States in 2013, and almost 40,000 will die from the disease ("Breast Cancer," 2012; Siegel, Naishadham, & Jamal, 2012). The lifetime probability of developing breast cancer is one in six overall (one in eight for invasive disease) (American Cancer Society (ACS), 2011-2012; "Breast Cancer," 2012; "Surveillance & Epidemiology," n.d.). Breast cancer death rates are declining. This is probably the result of detecting the cancer earlier and doing enhanced treatment.

Breast cancer treatments can be classified as local or systematic. Surgery and radiation fall under local while chemotherapy and hormone therapy are examples of systematic therapies. Usually for the best results, the two types of treatments are used collectively ("Breast Cancer," 2012). 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 (Costanza & Chen, 2012; "Surveillance & Epidemiology," n.d.; Siegel et al., 2012).

Although cancer research is generally clinical and/or biological in nature, data mining research is becoming a common match. In medical domains where data and statistics driven research is successfully applied, new and fresh research directions are recognized to promote clinical and biological research.

Forecasting the result of a disease or discovering information previously unknown is one of the most inspiring and challenging tasks in which to develop data mining applications. Survival analysis is a field in medical prognosis that deals with application of various techniques to historical data in order to predict the survival of a particular patient suffering from a disease over a time period (Bellaachia & Guven, 2006; Delen, Walker, & Kadam, 2005; Delen, 2009; Gupta, Kumar, & Sharma, 2011; Jonsdottir, Hvannberg, Sigurdsson, & Sigurdsson, 2008; Kharya, 2012; Thongkam, Zhang, & Huang, 2008).

Breast cancer survivability prediction has mostly been studied through clinical approaches. In this study, we present our analysis of the prediction of survivability rate of breast cancer patients using data mining techniques such as Naïve Bayes and association rules and data mining tool of XLMiner ("XLMiner On Line," n.d.). Through data mining experiments, we are studying the significance of the prognostic factors in the survivability rate of the breast cancer patients.

We used Naïve Bayes classification model for its performance and most adaptability in using diverse types of data without limitations on variables being categorical or continuous. Naïve Bayes is used to predict one on one relation with the dependent variable. Association rules technique, on the other hand, could show the combination of the prediction factors with the dependent variable. It would be beneficial in running the two mining techniques in order to get much clearer picture. Our goal was to discover rules that can easily be interpreted by medical people.

It was our intension, through experiments, to show the significance of the established prognostic factors and their combinations on the prediction of the survivability rates of the breast cancer patients. Such knowledge would enable us to further expand the current prognostic tools in order to produce more accurate breast cancer survivability rate.

DOI: 10.4018/978-1-4666-5888-2.ch158

BREAST CANCER PROGNOSIS

Although scientists do not know the exact cause of most breast cancers, 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 among others.

Medical prognosis is a field in medicine that encompasses the science of estimating the complication and recurrence of disease and to predict survival of patient (ACS, 2011-2012; Bellaachia & Guven, 2006; Delen et al., 2005; Gupta, 2011; Jonsdottir et al., 2008; Thongkam et al., 2008). Survival analysis is a field in medical prognosis that deals with the application of various methods to estimate the survival of a particular patient suffering from a disease.

A prognostic factor may be defined as a measurable variable that correlates with the natural history of the disease. Some of the key factors are the size of the tumor, lymph node involvement, number of positive nodes, grade, and the stage of cancer. In general, but not always, the smaller the size of the tumor, the better the chance the patient has for successful treatment. Doctors define small as less than 2 centimeters or about three fourths of an inch. The presence of cancer cells is known as lymph node involvement. If lymph nodes have some cancer cells in them, they are called positive nodes. The grade of a cancer is a measure of how much the tumor looks like the normal tissue from where it originated. A grade 1, or well-differentiated carcinoma, looks very much like the normal, nearby breast tissue. Grade 2, or moderately differentiated carcinoma, looks less like the normal tissue. Grade 3 show very little similarities to the normal breast ducts or lobules. Stage of cancer is usually expressed as a number on a scale of 0 through IV — with stage 0 describing non-invasive cancers that remain within their original location and stage IV describing invasive cancers that have spread outside the breast to other parts of the body.

The prognostic factors used in the prediction of survival of breast cancer can be separated into two categories: chronological (based on the amount of time present, i.e., Stage of Cancer), or biological (based on the potential behavior of the tumor, i.e., Histological Grade) (Bradley, 2007; Costanza & Chen, 2012; Maskarinec et al., 2011; Soergomataram, Louwman, Ribot, Roukema, & Coebergn, 2008).

Lymph node status, tumor size, histological grade are among the prognostic factors in use today (Bradley, 2007; Costanza & Chen, 2012; Maskarinec et al., 2011; Soergomataram et al., 2008). Lymph node status is time-dependent factor and is directly related to survival. One of the most significant prognostic factors in breast cancer is the presence or absence of axillary lymph node involvement, which is usually assessed at the time of surgery using sentinel lymph node biopsy or axillary dissection (Bradley, 2007). Macrometastases (>0.2 cm in size) have clearly been shown to have prognostic significance.

Tumor size has long been recognized as an independent prognostic factor and as a predictor of axillary node status, with larger tumors being associated with a worse prognosis and an increased likelihood of nodal metastasis. In Bellaachia and Guven (2006), authors have used Weka mining tool and ranked the survivability attributes. The result indicates that Extension of Tumor has a higher rank than the tumor size.

Histological grade is being identified as being highly correlated with long term survival. Patients with a grade 1 tumor have a much better chance of surviving than patients with grade 3 tumor ("Breast Cancer," 2012). In Delen et al. (2005), authors also conducted sensitivity analysis on artificial neural networks model in order to gain insight into the relative contribution of the independent variables to predict survivability. The sensitivity results indicated that the prognosis factor "Grade" is by far the most important predictor, followed by "Stage of Cancer," "Radiation," and "Number of Primaries."

Other factors include the patient's age, general health, estrogen-receptor and progesterone-receptor levels in the tumor tissue.

DATA MINING TECHNIQUES

Naïve Bayes

Bayesian classifiers (Endo, Shibata, & Tanaka, 2008; Han, Kamber, & Pei, 2011; Witten, Ian, & Frank, 2011; "XLMiner On Line," n.d.) operate by saying "If you see a fruit that is red and round, which type of fruit is it most likely to be, based on the training data set? In future, classify red and round fruit as that type of fruit." The variables are assumed to be independent