Knowledge Discovery from a Breast Cancer Database

Subramani Mani\textsuperscript{1}, Michael J. Pazzani\textsuperscript{1} and John West MD\textsuperscript{2}

\textsuperscript{1} Department of Information and Computer Science, 444 Computer Science Building, University of California at Irvine, Irvine, CA 92697-3425 (mani,pazzani)@ics.uci.edu
\textsuperscript{2} Breast Care Center, 1140 W. La Veta, Orange CA 92668

Abstract. We report on the use of various Machine Learning algorithms on an electronic database of breast cancer patients. The task was to predict breast cancer recurrence using a short subset of clinical attributes such as tumor presence, tumor size, invasive nature of tumor, number of lymph nodes involved, severity of lymphedema and stage of tumor. The predictive accuracy over fifty runs employing test sets not used to build the model were 63.4\%(Cart), 63.9\%(C45), 62.5\%(C45rules), 66.4\%(FOCL) and 68.3\%(Naive Bayes). An extension of the model using additional features and larger datasets is contemplated.

1 Introduction

Breast cancer affects approximately ten per cent of women in US. Much of the research effort has been focussed on early detection for better treatment of the condition. Breast cancer staging and TNM(Tumor, Node, Metastasis) classification were applied for devising optimal treatment plans. In spite of the advances made, relapse or recurrence of the malignancy continues to be a significant problem. Tumor characteristics, extent of lymph node involvement and patient parameters have been postulated to be predictors of prognosis. The primary treatment selected for the condition also plays a modifying role.[1] Machine Learning(ML) methods have been employed in many domains including some medical domains for induction of useful rules from data. ML algorithms have been used successfully to classify benign and malignant tumors from histological data.[2] There has also been earlier attempts at application of these techniques to the problem of relapse but they emphasized development of ML methods to noisy data.[3]

2 Aim of the Study

In this work we explore the use of various machine learning techniques to identify with a high degree of certainty the tumor characteristics responsible for recurrence of carcinoma of the breast.
3 Methods

3.1 Database

For this study we used the electronic medical record database of Breast Care Center, Orange, CA. Data consisted of demographic factors, clinical features, tumor parameters (macroscopic and microscopic), investigative procedures performed, diagnostic variables, treatment given and follow up measures. There were 887 patients.

3.2 Attribute Selection

Six attributes were identified by the surgeon Dr. John West for usefulness in predicting recurrence. The features considered for evaluation were presence of tumor, size of tumor, whether the tumor was of invasive nature, number of lymph nodes involved, severity of lymphedema and stage of tumor. The class variable was recurrence (coded yes and no).

3.3 Datasets for ML Runs

Out of the total sample of 887, 802 (90%) did not develop recurrence during follow up and 85 (10%) had relapse. To eliminate the bias of class distribution (90% versus 10%), the major class was randomly partitioned and six datasets created with class proportions of approximately 60% and 40% (total n = 233). Each algorithm was run fifty times on all the datasets. The datasets were randomly split into a two-third training set and a one-third testing set. The accuracies reported are of the test set.

3.4 Machine Learning Algorithms

Machine Learning algorithms which generate classification trees and rules are of particular importance in our context. The rules and trees provide necessary explanation for their predictions. The tree classifiers (C4.5[4] and CART[5]) and rule inducers (C4.5rules and FOCL[6]) were tried for this work. The results were compared in terms of prediction accuracy with a robust classifier Naive Bayes which assumes that the attributes are conditionally independent given the class. Nevertheless, it gives a good baseline performance.

4 Results

The comparative results of the different ML algorithms on these six datasets are shown in Table 1. An example of a rule set from C45rules run and a Cart induced tree are given in subsection 4.1 and subsection 4.2.