Comparison of Eight Algorithms To Predict Breast Cancer 5-year Survival

Arihito Endo\textsuperscript{a,1}, Takeo Shibata\textsuperscript{b}, Hideaki Takata\textsuperscript{a}, and Hiroshi Tanaka\textsuperscript{a}

\textsuperscript{a} Tokyo Medical and Dental University, Tokyo, Japan
\textsuperscript{b} School of Medicine, Tokai University, Isehara, Japan

Keywords: Artificial Neural Network, Logistic Regression Model, Decision Trees, Bayesian Model, Cox Proportional Hazard Model, SEER Program

In this paper, we present optimal models to predict the survival rate of breast cancer patients in five years.

This study used the 37,256 follow-up patients by 2002 that were diagnosed as breast cancer and registered in the SEER program from 1992 to 1997. We adopted eight common algorithms (Logistic Regression model, Decision Trees (J48), Artificial Neural Network, Cox Proportional Hazard, Decision Trees with naive Bayes, Naive Bayes, Bayes Net and Decision Trees (ID3)) to develop the prediction models. SPSS software for the statistical analysis, the open source software R and Weka were employed to establish the models.

To estimate performances in the models, accuracy, sensitivity and specificity were used as criteria, and were compared. The 10-fold cross-validation method was adopted in our estimation to reduce possible bias and calculate 95\% confidence interval.

Logistic Regression model showed the highest accuracy. Cox Proportional Hazard model had the highest sensitivity and Artificial Neural Network showed the highest specificity. The Decision Trees model tended to be more sensitive to survival data and the Bayesian model tended to be more sensitive to death data. Decision Trees with naive Bayes model showed stable performance for sensitivity and specificity in our dataset.

Our results show that relatively high sensitivity and accuracy were achieved. Specificity was relatively low. As the 5-year survival rate was 81.6\% in our dataset, Cox Proportional Hazard model might show the highest accuracy in the employed models.

We considered sensitivity as a more reliable indicator than specificity, since the probability of breast cancer patient survival was high. Use of higher-sensitivity algorithms allows better prediction of breast cancer survivability.

For patients who need survival information for 5 years, sensitivity may be more important than the other measures. High specificity means good prediction of risk of death within 5 years. Using sensitivity as an indicator to predict breast cancer survival would be better, as high sensitivity can predict survival and the mental damage from

\textsuperscript{1} Corresponding Author: 1-5-45 Yushima, Bunkyo, Tokyo 113-8510, Japan; E-mail: endo@arihito.net
these figures would be decreased. The optimal algorithms for predicting survival might be Cox Proportional Hazard model.