Data Mining Methods in Health Care Management

5-year Breast Cancer Survivability

Kung-Jeng Wang*[a], Bunjira Makond[b] and Yu-Siang Lin[c]
Department of Industrial Management, School of Managemen
National Taiwan University of Science and Technology, Taipei, Taiwan, R.O.C.
[a]kjwang@mail.ntust.edu.tw
[b]D9801802@mail.ntust.edu.tw [c]m9601106@mail.ntust.edu.tw

Abstract

The purpose of this paper is to study the performances of two data mining techniques for predicting 5-year survivability of breast cancer patients. The data set consisted of information about patients who have cancer diagnosis collected by SEER. In this study, data set is pre-classified into survival and non-survival with 90.66% and 9.34%, respectively. The selected variables used to predict 5-year survivability of breast cancer patients are race, grade, extension of disease, site-specific surgery code, stage of cancer, and SEER modified AJCC stage 3rd. The performances of two methods are considered from the perspective of three criteria (i.e. accuracy, g – mean and ROC); the results show that logistic regression is better than decision tree.

Keywords: Data mining, logistic regression, decision tree, SEER database, breast cancer, survivability

I. Introduction

Breast cancer is one of leading causes of death in the world [1] and the second cause of death of women in United States [2]. The diagnosis of breast cancer has impacts not only to mental effects but also family economics. According to the report of [3], the incidence of breast cancer decreased 2.0% per year from 1999 to 2006 and the death decreased 1.2% per year from 1999 to 2006. Since recently many new cancer detection and treatment approaches were developed, the cancer incidences and death of breast cancer decreased constantly. Though recently, the advance of cancer diagnosis and treatments have been able to reduce the number of mortalities and increase the survival time for patients. Meanwhile, patients and their family and friends are concerned about survival time after diagnosis in order to plan regarding their treatments and finances. Thus, the accurate prognosis is necessary; however, it is difficult for a physician to have accurate answers because the survivability is related to many factors. Presently, data mining techniques are used to obtain useful information from the large amounts of data which can help the physician for decision making regarding the prognosis. Data mining techniques are widely applied to medical data to predict the survivability of breast cancer patient; in [4]'s study, decision tree can predict the survivability more accurate than logistic regression, whereas [5]'s study concluded that logistic regression has better performance. Because the

performances of decision tree and logistic regression in [4] and [5] are controversial, the purpose of this research is to study the performance of the decision tree and logistic regression for predicting 5-year survivability of breast cancer patients based on Surveillance, Epidemiology and End Results (SEER) database [6] which covers most current records.

Our study finds that cancer survivability class distribution (90.66% for survival and 9.34% for non-survival) differs from [4]'s study (46% for survival and 54% for non-survival). The accuracy of decision tree (90.94%, 91.19%) and logistic regression (91.11%, 91.34%) are higher than the results of decision tree (85.4%, 85.8%) and logistic regression (85.6%, 86.0%) in the study of [5]. In addition, logistic regression has a better performance than the decision tree which is inconsistent with the results of [4], when considering from the perspective of g-mean and AUC.

II. Data and data preparation

This study uses data from SEER. To study breast cancer, we gained data from the SEER_1973_2007_TEXTDATA directory which were requested through website **(**www.seer.cancer.gov). The dataset is pre-classified into two classes (i.e. survival and non-survival). Herewith, the instances that did not survive for 5 years from the diagnosis date and that recorded the cause of death other than breast cancer are removed. The instances that had the follow-up cutoff date not completed in sixty months (who had diagnosis after December 31**,** 2002) are removed. Afterward, the remained instances are indicated as survival if they survived for 5-years after the diagnosis date; otherwise non-survival. Herein, irrelevant variables are deleted (such as Patient ID number, Registry ID, Birthplace, Sequence number and so on). Redundant information variables such as "year of birth" and "Age recode with < 1 year olds" are excluded from the data set because these types of information are redundant with "Age at diagnosis". Furthermore, predictor variables are selected by considering from the previous papers **(**[4];[5]; [7]; [8]; [9]; **[**10**]**). The variables with high missing value are deleted and the variables which had a single value more 90% of all values are also deleted. The values of important variables such as "Extent of Disease" and "AJCC stage of cancer" are missing before 1988, so we delete the records of missing values (also refer to **[**4]; **[**7**]**; [8]; **[**9**]** using the same procedure). Site-specific surgery variable is recorded separately in two variables in different time periods and codes, thus mapping data and recoding data procedures are applied to deal with this variable. Outlier is considered; for instance, the values of "Tumor size" for those are greater than 200 mm. are removed from the data set. The results of classification are presented in Tables 1.

Table 1: Cancer survivability class distribution

Class	Number of records	Percentage
Survival =1	195,252	90.66%
Non-survival = 0	20,123	9.34%
Total	215,375	100%

III. Methods

In this study, the proposed prediction models of survival were developed by using two data mining methods: logistic regression and decision tree.

Logistic regression

Logistic regression is a statistical method to describe the relation between predictor variables and response variable which is categorical variable with two values ("survival" or "non-survival"). The relation of response variable and predictor variables is found to be non-linear. To interpret logistic regression, odds ratio is used. Odds ratio is defined as the probability that an event occurs divided by the probability that the event does not occur. Odds ratio is greater than 1 when the event is more likely than not to occur [11].

Logistic regression model for *p* predictor variables can be written as

$$P(Y=1) = \frac{e^{(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}}{1+e^{(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}} \quad (1)$$

where $P(Y=1)$ is the probability of the patients who are survival, and $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are regression coefficients.

A useful transformation of logistic regression is logit transformation defined as:

$$g(x) = ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_px_p \quad (2)$$

Odds ratio (*OR*) is thus as following formula:

$$OR = e^{(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p)}$$

Decision tree

This study adopts J48 algorithm to predict breast cancer survivability. J48 is a release of C4.5 which has high accuracy, comprehensibility and stability. In addition, C4.5 which was developed from ID3 algorithm also deals with problems of missing data, continuous data, pruning rules and splitting criterion [12].

Model evaluation

Classifier methods are commonly evaluated by their accuracy, sensitivity, specificity, g-mean, receiver operating characteristic (ROC) curve and area under the curve (AUC).The accuracy, sensitivity, specificity, and g-mean are calculated as the following formulas.

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$sensitivity = \frac{TP}{TP+FN} \quad (4)$$

$$specificity = \frac{TN}{TN+FP} \quad (5)$$

$$\text{g}-\text{mean} = \sqrt{\frac{TN}{(TN+FP)} \times \frac{TP}{(TP+FN)}} \quad (6)$$

where TP denotes true positives, TN denotes true negatives, FP denotes false positives, and

FN denotes false negatives. These values are often displayed in a confusion matrix as be presented in Table 2.

Table 2: Confusion matrix

	Predicted class		
Actual class		Survive	Non-survive
	Survive	TP	FN
	Non-survive	FP	TN

Receiver operating characteristic (ROC) curve and area under the curve (AUC)

The ROC curve is widely used to characterize the performance of models obtained by using data mining methods. It shows the relationship between the number of positive included in the sample (i.e. sensitivity) and the number of negatives included in the sample (i.e. 1-specificity). Since the values of sensitivity and 1- specificity vary, therefore ROC curve is generally summarized in a single quantity by the area under the curve (AUC). It can be obtained using the Wilcoxon non-parametric approach, whereas the larger the area under the curve, the better the model [13], [14].

10–fold cross-validation

In this study, data is divided into 2 types: training and testing sets. A 10–fold cross-validation is employed so that the bias caused by random sampling for training and testing sets can be reduced [14].

IV. Results

The confusion matrix of each method is presented in Table 3; the values to measure the performance of the methods (i.e. accuracy, sensitivity, specificity, and g-mean) are derived from the confusion matrix and showed in Table 4.

Table 3: Confusion matrix

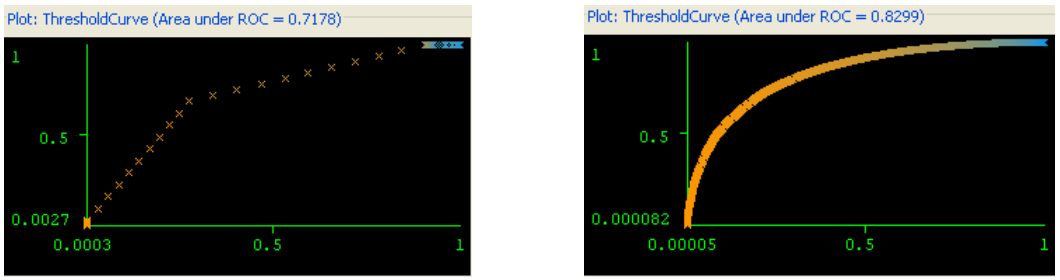
	Decision Tree Predicted class			**Logistic Regression** Predicted class		
Actual class		Survive	Non-survive		Survive	Non-survive
	Survive	194,495	757	Survive	193,182	2,070
	Non-survive	18,475	1,648	Non-survive	16,809	3,314

Table 4: Accuracy, sensitivity, specificity and g-mean

Methods	**Accuracy**	**Sensitivity**	**Specificity**	**g-mean**
Decision tree (J48)	91.07%	99.61%	8.18%	0.285
(Confidence interval)	(90.94%, 91.19%)	(99.58%, 99.63%)	(8.06%, 8.29%)	
Logistic regression	91.23%	98.93%	16.46%	0.403
(Confidence interval)	(91.11%, 91.34%)	(98.89%, 98.97%)	(16.30%, 16.61%)	

The accuracy value shows that there is no significant difference in accuracy between logistic regression and decision tree models while g-mean indicates that logistic regression

outperforms decision tree (0.403 > 0.285). ROC curve are plotted to measure the performance of two methods, as showed in Figure 1. The area under ROC curve indicates that logistic regression has a better performance than decision tree to predict 5-year breast cancer survivability (0.829>0.717).



(a) Decision tree model (b) Logistic regression model

Figure 1: ROC curve and AUC

V. Conclusions

In this study, we utilize two methods of data mining to investigate the 5-year survivability of breast cancer patients. The data preparation method is conformed to the literature of [4]; however, our outcomes are not similar to theirs (90.66% vs. 46% for survival and 9.34% vs. 54% for non-survival). This study employs data set which is twice the size of theirs and covers more current records. Since the death of breast cancer decreased constantly from 1999 to 2006; such change would contribute to the significant differences between our research conclusion and [4].

In addition, this study finds that while the accuracy index shows the performance of both methods seems to be indifferent, notably logistic regression outperforms decision tree according to AUC and g-mean indices. Some measures are not perfectly suited to measure survivability when data is imbalanced (as in the case in this study) however, AUC and g-mean are better at measuring, due to their robustness towards imbalanced class distributions (also refer to the claim of [15]).

Likewise, the accuracy of decision tree and logistic regression in this study is higher than [5]. Note that this study reaches a smaller variation than [5] (i.e. the confidence interval of the accuracy of decision tree and logistic regression in this study are (90.94%, 91.19%) and (91.11%, 91.34%) respectively, whereas (85.4%, 85.8%) and (85.6%, 86.0%) in [4]'s study, because this study used larger data set than theirs and the models are refined and tuned to best fit the data. This confirms to the assertion that the accuracy of decision tree and logistic regression are improved when the data set is increased, even the data set is decreased the accuracy is increasing [16], [17], [18].

On the other hand, the inconsistent results between this study and [4] can be explained by considering the property of variables. Some variables are found to be non-linear and two predictor variables have particularly high correlation. Such non-linearity problem in data harms the correct classification of decision tree due to its linear-separable property [19]. In contrast, logistic regression suits when the relationships of predictor variables and response variable are non-linear. Multicollinearity is also a problem for decision tree when two variables have high correlation as only the best variable will be chosen, whereas logistic regression will be prone to use both and can solve the problem of multicollinearity by using Ridge estimation to estimate parameters.

In term of health care management, this study purposes the optimal data mining method for predicting 5-year breast cancer survivability. The accurate prediction has benefits for physicians, patients and their families that is physicians are able to provide the appropriate treatments for patients, while patients and their families can do decision making and planning about their quality of life according to their finances. It also has benefits for health policy and planning makers to provide health insurance system for breast cancer patients.

References

World Health Organization (2010). Quick Cancer Facts. Retrieved September 22, 2010. from http://www.who.int/cancer/en

National Cancer Institute (2010). A Snapshot of Breast Cancer. Retrieved September 22, 2010. from http://www.cancer.gov/aboutnci/servingpeople/snapshots/breast.pdf

Edwards et al. (2010). Annual Report to the Nation on the Status of Cancer, 1975-2006, Featuring Colorectal Cancer Trends and Impact of Interventions (Risk Factors, Screening, and Treatment) to Reduce Future Rates. Cancer, 116(3), 544-573.

Delen, D., Walker, G., & Kadam, A. (2005). Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine, 34, 113-127.

Endo, A., Shibata, T., & Tanaka, H. (2008). Comparison of Seven Algorithms to Predict Breast Cancer Survival. Biomedical Soft Computing and Human Sciences, 13, 11-16.

SEER (2010) Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973-2007), National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2010, based on the November 2009 submission.

Bellaachia, A., & Guven, E. (2006). Predicting Breast Cancer Survivability Using Data Mining Techniques. Department of Computer Science, George Washington University, 1-4. Technical report.

Liu, Y., Cheng, W., & Lu, Z. (2009). Decision tree based predictive models for breast cancer survivability on imbalance data. Bioinformatics and Biomedical Engineering, 2009. ICBBE 2009, 1-4. doi: 10.1109/ICBBE.2009.5162571

Khan, M.U., Choi, J.P., Shin, H., & Kim, M. (2008). Predicting Breast Cancer Survivability Using Fuzzy Decision Trees for Personalized Healthcare. Proceedings of 30 th Annual International IEEE EMBS Conference. (pp 5148-5151).

Palaniappan, S., & Hong, T.K. (2008). Discretization of Continuous Valued Dimensions in OLAP Data Cubes. International Journal of Computer Science and Network Security, 8, 116-126.

Larose, D.T. (2006). Data mining methods and models. Hoboken, NJ: John Wiley & Sons.

Fountoulaki, A., Karacapilidis, M., & Manatakis, N. (2010). Using Decision Trees for the Semi-automatic Development of Medical Data Patterns: A Computer-Supported Framework. Web-Based Applications in Healthcare and Biomedicine, 229-242.

Liu, H., & Wu, T. (2003). Estimating the Area under a Receiver Operating Characteristic Curve For Repeated Measures Design. Journal of Statistical Software, 8(12), 1-18.

Witten, I.H. &Frank, E. (2005). Data mining: practical machine learning tools and techniques. San Francisco, CA: Morgan Kaufmann.

Gu, Q., Cai., Z., Zhu., L. & Huang., B. (2008). Data mining on imbalanced data sets. International Conference on Advanced Computer Theory and Engineering, 1020-1024.

Morgan, J., Dougherty, R., Hilchie, A., & Carey, B. (2003). Sample Size and Modeling Accuracy with Decision Tree Based Data Mining Tools. Academy of Information and Management Sciences Journal, 1-19.

Stockwell, D.R.B., & Peterson, A.T. (2002). Effects of sample size on accuracy of species distribution. Ecological modelling, 148, 1-13.

Komarek, P.R., & Moore, A.W. (2003). Fast Robust Logistic Regression for Large Sparse Datasets with Binary Outputs. Department of Mathematical Sciences and School of Computer Science, Carnegie Mellon University,1-8.

Piramuthu, S. (2008). Input data for decision trees. Expert Systems with Applications, 34, 1220-1226.