Journal of Theoretical and Applied Information Technology
10th February 2014. Vol. 60 No.1

ISSN: **1992-8645** www.jatit.org E-ISSN: **1817-3195**

PREDICTION OF SURVIVAL IN PATIENTS WITH BREAST CANCER USING THREE ARTIFICIAL INTELLIGENCE TECHNIQUES

[1]CHENG-TAO YU, [2] CHENG-MIN CHAO, [3]BOR-WEN CHENG
[1]Lecture, Department of Industrial Engineering and Management, National Yunlin University of Science and Technology, Taiwan
[2]Ph.D., Department of Industrial Engineering and Management, National Yunlin University of Science and Technology, Taiwan
[3]Prof., Department of Industrial Engineering and Management, National Yunlin University of Science and Technology, Taiwan
E-mail: [1]g9421802@ yuntech.edu.tw, [2]g9521807@yuntech.edu.tw, [3]chengbw@yuntech.edu.tw

ABSTRACT

As medical technology advances, has accumulated a large number of health-related data. Faced with increasingly complex analytical requirements, predictive data mining has become an essential instrument for hospital management and medical research. In this study, the breast cancer dataset is collected from a regional teaching hospital in central Taiwan between 2002 and 2009. The prognostic factors composed of 8 attributes including 967 subjects, of which 861 are survival after treatment. The three techniques, artificial neural networks (ANNs), support vector machine (SVM) and Bayesian classifier, have been discussed which is used to investigated and evaluated for predicting breast cancer survival. As can be seen from the results, the prediction accuracy of a 10-fold cross validation is 90.31%, 89.79% and 88.64%, respectively. Classification results of SVM are slightly better as compared to ANN and Bayesian classifier, however, from a relatively low variance, the results show that the SVM will be the best prognosis in clinical practice.

Keywords: *Breast Cancer, Bayesian classifier, Support vector machines (SVM), Artificial Neural Networks (ANNs)*

1. INTRODUCTION

As a malignant disease, cancer has caused millions of human deaths. In Taiwan, malignant neoplasm is the top 1 of the 10 death causes since 1982, and the death rate is still increasing every year. Malignant neoplasm not only causes high mid-to-long term medical care expenses, it is also a heavy burden of a patient's family and community. For this reason lots of advanced countries in the world has reinforced health medical care and invested considerable budget and human resources into malignant neoplasm research and education, trying to lower the fatality rate, morbidity rate and sequel, as well as the burden on individuals, families, communities and countries.

Malignant neoplasm includes liver cancer, lung cancer, colorectal cancer, breast cancer, gastric cancer, etc. Among them, breast cancer is the commonest cancer in women and is currently one of the most pressing medical issues. According to the extensive database of International Agency for Research on Cancer (IARC), in 2000, more than one million people around the world were diagnosed with breast cancer and about one-third of women died from the disease, despite that it can be cured at early stages [1]. In the United States, breast cancer is the most frequently diagnosed malignancy and the main cause of cancer death in women [2]. In Taiwan, the number of new cases has been increasing during the last few decades. Breast cancer is a major cause of death, and the fourth or fifth of top 10 death causes for women since 1996 [3], with the death rate increasing every year. Recent research showed that in a woman's lifetime, the chance of being affected by invasive breast cancer is approximately one in eight, and the chance of death is one in thirty five. Because the causes of breast cancer are still uncertain, accurate early detection is very important in order to lower the mortality rate [4]. Therefore, early breast cancer detection is a difficult and important issue from clinical perspectives.

Many data has been gathered with the advancement of technology, and computers are used to handle such large amount of data. The technology of detecting relation and knowledge from data is called data mining techniques (DMT). DMT forms a branch of applied artificial intelligence (AI), and is the process of choosing, finding and modeling huge amounts of data so as to discover unknown patterns or relationships which provide a clear and practical result [5]. There are two phases in data mining classification technique: classification model construction and model classification efficiency evaluation. In classification model construction, the algorithm is trained through a dataset to build the predictive classification model. In classification efficiency evaluation, a testing dataset is applied. Every data in a training dataset or a testing dataset includes distinct number of attributes and a target class [6-7].

The DMT has largely advanced in recent years, and have been commonly adopted in several professional fields, including medical issues, health sciences, commercial purposes, and manufacturing industry environments, etc. During the last few years, the medical field has been paying more attention on data mining techniques to build and compare predictive models [4, 8]. DMT has been widely applied in various medical tasks, such as breast cancer diagnosis and prediction [9], cerebrovascular disease prediction [7], and forecasting the outcomes of the cognitive rehabilitation of patients with acquired brain injury (ABI) [6].

The artificial neural networks (ANNs) algorithm has been applied in many medical issues, for example, assessing and building predictive models for diabetes [8], predicting the survival rates for diagnosed cases of breast cancer [9]. The support vector machine (SVM) applications in medical issues could be found in literatures, for instance, Karabatak and Ince[10] conducted differential diagnosis of breast cancer, Luo and Cheng[11] assisted diagnosis of breast cancer. The Bayesian classifier has been applied in many medical issues including cerebrovascular disease forecasting [7].

As the pathogenesis of breast cancer is variable, it is difficult to accurately diagnose beforehand. However, in the perspective of preventive medicine, it is necessary to develop successful identification method and a predictive model to recognize the breast cancer and used to improve the diagnosis and prediction of breast cancer. Statistical techniques and artificial intelligence techniques are important in unearthing previously unknown knowledge; several researchers have applied these methods to forecast breast cancer [12-13]. Motivated by the need of a strong diagnostic tool for breast cancer, this research was to investigate the use of artificial intelligence methods and data mining techniques for predicting breast cancer accuracy. The data were collected from the cancer registry of a regional teaching hospital in central Taiwan between 2002 and 2009. This research adopted several data mining techniques, artificial neural networks (ANNs), support vector machine (SVM) and Bayesian classifier, to construct an optimum breast cancer prediction model. Accuracy was used for evaluation.

The remainders of this research are organized as follows. Section 2 provides materials and methods including participants, data collection, data processing and prediction methods. Section 3 provides prediction results of all three algorithms. The research is discussed in Section 4. Finally, Section 5 concludes our findings and suggestions for future research.

2. MATERIALS AND METHODS

2.1. Participants and Data collection

To conduct the research reported here, the data is gathered from a regional teaching hospital in Central Taiwan between 2002 and 2009. The prognostic factors composed of 8 attributes including 967 subjects, of which 861 are survival after treatment. The subjects contained a population representing different ethnic or racial groups residing in Taiwan. The main purpose is to distinguish between the people survival after treatment with breast cancer from the data, according to “status” variable which is assign to 1 for survival and 2 for death. The attributes of the dataset comprise seven variables, age, chemotherapy, radiotherapy, tumor size, the number of examined lymph nodes, number of attacked lymph nodes, and pathological staging. These variables contain critical information for the cancer staging system TNM and NPI (an indicator used for predicting cancer patients’ survival). Several studies [14-15] on histopathological factors found higher lymph node involvement, histologic grading, stage at presentation, greater tumor size, and more triple-negative tumors in younger patients. D’Eredita et al. [16] reported the lymph status, tumor size, and histological grade are usually the most critical factors for breast cancer survivability, and these three features also emerged as key variables in our study. Delen et al. [9] selected three key variables: pathological staging, radiotherapy, and tumor size. These findings are partly similar as in this research. All of the attributes are described in Table 1 and Table 2.

Table 1: Descriptive Statistics of dependent variable

Status	Frequency	Percentage
1(survived)	861	89.0%
2(death)	106	11.0%
Total	967	

Table 2: Predictor variables for survival modeling

Categorical variable name	Number of unique values		
Chemotherapy	5		
Radiotherapy	4		
Pathological staging	5		
Continuous variable name	Mean	S.D	Range
Age	573.84	11.47	26-90
Tumor size	2.95	1.91	0-15
Number of lymph nodes examined	22.02	7.99	3-76
Number of lymph nodes attacked	3.62	7.45	0-66

To ensure that the results are objective and would be valid for making predictions regarding new data, this research randomly partitioned the dataset into independent training and testing sets with stratified 10-fold cross-validation by the partition node of SPSS statistical analysis tool. The 10-fold cross-validation method is used to measure the unbiased estimate of the three prediction models for the purposes of comparing their performances. Each fold has approximately an equal size. The dataset is randomly divided into two splits, 90 % training and 10% testing.

2.2. Data Processing

In this research, the breast cancer model was constructed in three stages. First stage: data pre-processing and screening. Second stage: Classification model construction by ANNs, SVM and Bayesian classifier. Third stage: Classification efficiency comparison, optimum predictive model determination, and diagnosis classification rules extraction.

Data analyses were performed using SPSS statistical analysis tool and data mining prediction models were constructed using Clementine 7.2, MySVM toolkit, and Weka. These software packages were chosen to explore and manipulate the data. The next section describes the surface complexities and structure of the data. We executed experiments using the published template datasets and compared the results of ANNs, SVM, and Bayesian classifier. Because we did not know which parameter or kernels were appropriate for the samples, we tested several parameters and randomly selected kernels for the ANNs, SVM, and Bayesian classifier to find the best predictors.

3. RESULTS

Classification is an important data mining method. Different classification algorithms, artificial neural networks (ANNs), support vector machine (SVM), Bayesian classifier, logistic regression (LR), and decision tree (DT), have been proposed in the medical fields [4, 7-8, 9-11]. On the basis of popularity and efficiency, this research adopted three well-known and widely used data mining classification models, including ANNs, Bayesian classifier and SVM. In this research, a large dataset (967 cases and seven risk factors) from the seven programs were analyzed, and after a long process of data selection and transformation, we constructed the prediction models.

All three predictive models applied in this research were trained and tested with the same data and validated using 10-fold cross-validation. Finally, use average of results of 10 tests to evaluate predictive model classification efficiency, and selected the best network structure for each model as explained in the following section.

The dependent variable was a binary categorical variable, which was calculated from the variables in the raw dataset to represent breast cancer survival (where survival was represented with a value of ‘‘1’’ and death was represented with ‘‘2’’). The independent variables were seven risk factors of breast cancer. There were patients' age, chemotherapy, radiotherapy, pathological staging, tumor size, number of lymph nodes examined, and number of lymph nodes attacked. To examine whether these variables affected a breast cancer diagnosis, we adopted three data mining methods: ANNs, Bayesian classifier and SVM. The accuracy of classification was measured by these three data mining methods. The comparisons were based on 10-fold cross-validation. In table 3, the test performance of the data mining techniques are shown. This was derived through measuring the performance of the training dataset and testing dataset on the test data subsets.

The training dataset results showed that the SVM techniques performed the best among the three data mining techniques, with a classification accuracy of 90.31%. The ANNs technique was the second best, with a classification accuracy of 89.79%. The Bayesian classifier techniques performed worst, with a classification accuracy of 88.64%. The SVM techniques had the best results among the three techniques. In the testing dataset, the ANNs techniques achieved a classification accuracy of 90.23%. The SVM techniques gave a classification accuracy of 91.24% and the Bayesian classifier techniques presented a classification accuracy of 88.61%. The SVM techniques had the best results among the three techniques.

In SVM model, most accuracy results were better than in ANNs model. But Bayesian classifier model accuracy was lower than SVM and ANNs, and was rather unstable in 10-fold cross-validation. In training dataset, the standard deviation of Bayesian classifier model was 5.36 and testing dataset scored 2.57, both were the highest scores of all three models. The results showed that, in the experiment, we failed to find the best solution of Bayesian classifier.

In addition, this research chose input factors based on seven impact factors and survived and death as output factors, and then found important impact factors via the three data mining techniques. The results indicated that the key impact factors are age, radiotherapy, tumor size, the number of examined lymph nodes, number of attacked lymph nodes, and pathological staging.

Table 3: Results for 10-fold cross-validation for all folds and all model types

Fold No.	Training dataset			Testing dataset		
	Accuracy			Accuracy		
	ANN	SVM	Bayes	ANN	SVM	Bayes
1	90.62	89.58	76.04	88.75	90.50	85.70
2	90.62	91.67	86.46	89.62	92.71	86.46
3	91.67	91.67	94.79	91.75	87.50	92.76
4	87.50	87.50	91.67	86.50	91.67	91.67
5	85.42	92.71	90.62	87.50	92.71	90.62
6	92.71	89.58	92.71	91.67	92.71	86.46
7	91.67	91.67	90.62	92.71	87.50	91.67
8	92.71	91.67	91.67	91.50	91.67	90.62
9	87.50	87.50	86.46	89.67	89.67	86.46
10	87.50	89.58	85.42	89.67	92.71	89.67
Mean	89.79	90.31	88.64	90.23	91.24	88.61
S.D.	2.58	1.84	5.36	1.63	1.73	2.57

4. DISCUSSION

In this research, the accuracy of SVM (90.31%) and ANNs (89.79%) are better than Bayesian classifier (88.64%). Moreover, Bayesian classifier model has the highest accuracy standard deviation in 10-fold cross-validation (5.36 of training dataset and 2.57 of testing dataset), indicating worse forecasting ability comparing to SVM model and ANNs model.

Compared with the present study, a number of researchers have derived different results from the classification of their own choosing. Delen et al.[9] uses logistic regression model, artificial neural networks and decision tree on a large dataset, three key variables, pathological staging, radiotherapy, and tumor size, have been selects. Accuracy is reported in the range of 89.2 -93.6%.In the examinations of Yeh et al.[7], a 91.30%- 98.01% accuracy was shown using decision tree, Bayesian classifier and back propagation neural network respectively on a 493 valid sample patients. Back propagation neural network had the highest accuracy than decision tree and Bayesian classifier Meng et al.[8] compares the performance of logistic regression, artificial neural networks (ANNs) and decision tree models for predicting diabetes or pre-diabetes, presented 77.87%-82.18% accuracy. From the above results, the accuracy of this study is reasonably acceptable.

5. CONCLUSION

This research adopted several data mining techniques, artificial neural networks (ANNs), support vector machine (SVM) and Bayesian classifier, to construct an optimum breast cancer prediction model. The data of the breast cancer patients were accessed from a regional teaching hospital in central Taiwan. The dataset composed of 8 attributes including 967 subjects, of which 861 are survival after treatment. In comparison of data mining techniques, this research used accuracy indicator to evaluate classification efficiency of different algorithms. Overall, SVM classification accuracy is better than artificial neural networks and Bayesian classifier, both from the observation of mean and standard deviation. However, from a relatively low variance, the results show that the SVM will be the best prognosis in clinical practice.

The optimum breast cancer disease predictive model obtained in this study adopts SVM classification algorithm, this research may provide references for future research on selecting the optimal predictive models to lower the incidence of breast cancer. However, we are hopeful that future research will provide more details about data mining approaches in order to reach higher prediction rates.

REFRENCES:

[1] Ferlay J, Bray F, Pisani P and Parkin DM. GLOBOCAN, “Cancer incidence, mortality and prevalence worldwide. IARC Cancer Base

No. 5. version 2.0. Lyon (France)", IARC Press 2004.

[2] Fabregue, M., Bringay, S., Poncelet, P., Teisseire, M., and Orsetti, B., "Mining microarray data to predict the histological grade of a breast cancer", Journal of Biomedical Informatics, 44 Suppl. 1:S12-6, 2011. doi: 10.1016/j.jbi.2011.03.002.

[3] Ministry of Health and Welfare, "Ministry of Health and Welfare", Executive Yuan, Republic of China (Taiwan), 2013. http://www.mohw.gov.tw/cht/DOS/Statistic.aspx?f_list_no=312&fod_list_no=1601.

[4]. Nahar, J., Imam, T., Tickle, K. S., Ali, A. B. M. S., and Chen, Y. P. P., "Computational intelligence for microarray data and biomedical image analysis for the early diagnosis of breast cancer", Expert Systems with Applications, Vol. 39 No. 16, 2013, pp. 12371-12377. doi:10.1016/j.eswa.2012.04.045.

[5] Liao, S.-H., Chu, P. -H., and Hsiao, P. -Y., "Data mining techniques and applications - A decade review from 2000 to 2011', Expert Systems with Applications, Vol. 39 no. 12, 2012, pp. 11303-11311. doi: 10.1016/j.eswa.2012.02.063.

[6] Marcano-Cedeño, A., Chausa, P., García, A., Cáceres, C., Tormos, J. M., and Gómez, E. J., "Data mining applied to the cognitive rehabilitation of patients with acquired brain injury", Expert Systems with Applications, Vol.40, No. 4, 2013, pp. 1054-1060. doi: .1016/j.eswa.2012.08.034.

[7] Yeh, D. -Y., Cheng, C. -H., and Chen, Y. W., "A predictive model for cerebrovascular disease using data mining", Expert Systems with Applications, Vol. 3, No. 7, 2011, pp. 8970-8977. doi: 10.1016/j.eswa.2011.01.114.

[8] Meng, X. -H., Huang, Y. -X., Rao, D. -P., and Liu, Q. Z., "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors", Kaohsiung Journal of Medical Sciences, Vol. 29, 2013, pp. 93-99. doi: 10.1016/j.kjms.2012.08.016.

[9] Delen, D., Walker, G., and Kadam, A., "Predicting breast cancer survivability: A comparison of three data mining methods", Artificial Intelligence in Medicine, Vol. 34, No 2, 2005, pp. 113-127. Doi: 10.1016/j.artmed.2004.07.002.

[10] Karabatak, M., and Ince, M. C., "An expert system for detection of breast cancer based on association rules and neural network", Expert Systems with Applications, Vol. 36, No. 2, 2009, pp. 3465-3469. doi: 10.1016/j.eswa.2008.02.064.

[11] Luo, S. -T., and Cheng B. -W., "Diagnosing breast masses in digital mammography using feature selection and ensemble methods", Journal of Medical Systems, Vol. 36, No. 2, 2012, pp. 569-577. doi: 10.1007/s10916-010-9518-8.

[12] Kovalerchuck, B., Triantaphyllou, E., Ruiz, J. F., & Clayton, J., "Fuzzy logic in computer-aided breast-cancer diagnosis: Analysis of lobulation", Artificial Intelligence in Medicine, Vol. 11, 1997, pp. 75-85. Doi: 10.1016/S0933-3657(97)00021-3.

[13] Pendharkar, P. C., Rodger, J. A., Yaverbaum, G. J., Herman, N., & Benner, M., "Associations statistical, mathematical and neural approaches for mining breast cancer patterns", Expert Systems with Applications, Vol.17, 1999, pp. 223-232. doi: 10.1016/S0957-4174(99)00036-6.

[14] Anders, C. K., Hsu, D. S., Broadwater, G., et al., "Young age at diagnosis correlates with worse prognosis and defines a subset of breast cancers with shared patterns of gene expression", Journal of Clinical Oncology, Vol. 26, No. 20,2008, pp. 3324-3330. doi: 10.1200/JCO.2007.14.2471.

[15] Hartmann, S., Reimer, T., and Gerber, B., "Management of Early Invasive Breast Cancer in Very Young Women (<35 years)", Clinical Breast Cancer, Vol. 11, No. 4, 2011, pp. 196-203. doi: 10.1016/j.clbc.2011.06.001.

[16] D'Eredita' G, Giardina C, Martellotta M, Natale T, Ferrarese F, "Prognostic factors in breast cancer: the predictive value of the Nottingham Prognostic Index in patients with a long-term follow-up that were treated in a single institution", European Journal of Cancer, Vol. 37, No. 5, 2001, pp. 591-6. doi: 10.1016/S0959-8049(00)00435-4.