Volume 4 ,Issue1

Available online at **http://UCTjournals.com**

UCT Journal of Research in Science, Engineering and Technology
UCT. J. Resea. Scien. Engineer. Techno. (UJRSET)

37-42 (2016)

ISSN 2382-977X

A survey on predicting breast cancer survivability and its challenges

Samaneh Miri Rostami [1], ***Mohammad Reza Parsaei*** [1,*], ***Marzieh Ahmadzadeh*** [1]

[1] *Department of Computer Engineering and IT, Shiraz University of Technology, Shiraz, Iran*
Email: S.Miri@sutech.ac.ir, mr.parsaei@sutech.ac.ir, Ahmadzadeh@sutech.ac.ir

Original Article:
Received 22 March. 2016, Accepted 28 April. 2016, Published 3 May. 2016

ABSTRACT

Data mining is a powerful technology that can be used in all domains in order to detect hidden patterns from a large volume of data. A huge amount of medical data gives opportunities to health research community to extract new knowledge in different parts of medicine such as diagnosis, prognosis, and treatment by using data mining applications in order to improve the quality of patient care and reduce healthcare costs. Breast cancer is the most common cancer in women worldwide and it is the leading cause of death among women. Data mining can be used as a decision support system to predict survival of new patients. In this study, related works in the field of breast cancer survival prediction are reviewed and by compromising these works challenging issues are presented.

Keyword:

Breast Cancer,
Prognosis,
Survival Analysis,
Medical Data Mining

*** Corresponding author: Mohammad Reza Parsaei**
Email: mr.parsaei@sutech.ac.ir

I. INTRODUCTION

Today, with the rapid growth of information and the ability to store massive data, knowledge acquisition from these resources has become a new science. The volume of data collected and stored in databases in the medical field and in other areas has increased significantly. As a result, traditional methods of data analysis for processing of such a volume of data is inefficient. For this purpose, new techniques are developed. Much of this progress is named Knowledge Discovery in Databases (KDD). KDD includes a variety of statistical analysis, pattern recognition, and machine learning techniques. In fact, KDD process by using the domain understanding, data understanding, data preparation, data collection and compiling knowledge from extracted patterns and post-processing want to extract knowledge from large volumes of registered data [1]. Data collection and compiling knowledge from extracted patterns are usually referred as data mining [2-6].

The enormous collection of health data has the opportunity to health research community to extract new patterns in different parts of medicine, including diagnosis, prognosis, and treatment by using data mining applications [7]. In recent years, data mining in the field of health care management has used to classify patients and medical diagnosis aimed to find out the pattern of survival of patients based on demographic data and clinical data [8]. Predict survival of breast cancer patients by considering its risk factors is difficult. The role of experts in predicting the survival of breast cancer is undeniable. But the experienced oncologist is limited. Given these circumstances, using hidden knowledge stored in electronic or paper records can be an effective way to support the less experienced physicians to use it in their daily decisions [9].

The rest of the paper is organized as follows: In section II medical data mining and the main topic of this study are described. Related works are reviewed in section III. In section IV, discussion is presented and concludes the paper in the end.

II. MEDICAL DATA MINING

Data mining is a three-step process: data preprocessing, data modeling and data post-processing [10]. Today, data mining can be declared as a multidisciplinary science. The aim of data pre-processing is to prepare raw data for extracting knowledge. The data modeling searches relationships among data to extract pattern. The extracted pattern should be assessed and verified to be considered as knowledge in the post-processing step. Also, background knowledge can be used to clarify the issues and verify the extracted knowledge.

The definition of Medical data mining is different based on the author's view. Definition presented here is based on [11] as follows:
"Extraction of implicit, potentially useful and novel information from medical data to improve accuracy, decrease time and cost, construct decision support system with the aim of health promotion".
Based on this definition, four objectives for medical data mining delivers:

a. Improving efficiency and reducing human error.
b. Reduce the time and cost.
c. Medical decision support system: inexperienced physician can benefit from such a system.
d. Knowledge extraction: extraction of relations between variables, identify risk factors and explore new knowledge.

A. Breast cancer

Recently cancer has been a major health problem for humans. According to the statistics, cancer caused 13% of all human deaths [12]. Different types of cancer, such as a bladder, breast, colon, kidney, and lung are there. After lung cancer, breast cancer ranked second in terms of incidence. Breast cancer is a malignant tumor that occurs when cells in the breast tissue divide and grow without normal controls on cell death and cell division [13]. Although scientists do not know the main cause of breast cancer, but some risk factors that increase the likelihood of breast cancer have identified. These factors include age, genetic factors and family history [14]. Treatment of breast cancer can be divided into two types: local and systematic. Surgery and radiation therapy are examples of local treatment, and chemotherapy and hormone therapy are examples of systematic one. Usually for a better result, two types of treatments are used together. Breast cancer is the most common cancer in women worldwide and it is the leading cause of death among women, but the chance of survival for breast cancer patients is high. With early diagnosis, about 97% of women survive for 5 years or longer [14]. Women's risk of developing invasive breast cancer during her lifetime is about 1 in 8, and 1 in 35 the likelihood of death [15]. In 2004, breast cancer caused 519,000 deaths worldwide (7% of cancer deaths; almost 1% of all deaths) [16]. In 2011, about 230,480 US women were diagnosed with breast cancer and approximately 39,520 of them died due to this disease [17]. In Iran, according to the National Center for Cancer Registration, breast cancer incidence has increased dramatically from 2001; in 2010, 23% of all cancers diagnosed in women were breast cancer cases [18].

B. Prognosis of breast cancer

When breast cancer is diagnosed, malignant neoplasms should be removed. During this stage, physicians need to make a prognosis of a disease. The estimate of how the disease will go for you and chance of survival is called prognosis. It is important because the type and amount of drugs is determined based on it [19]. Cancer prognosis has three aspects, predict susceptibility to cancer (e.g. risk assessment), predict cancer recurrence, and predict cancer survival [19].

Survival Analysis is a topic in medical prognosis that predicts patient survival for a specific time period [20]. In many types of research in the field of breast cancer, the term survival is considered for patients who are still alive after 5 years from the date of diagnosis. Prognostic factors are necessary for accurate identification of patients who may, or may not, benefit from treatment [15]. Knowledge of prognostic factors plays an important role in the treatment and care of patients. In fact, physicians become more informed about the outcome of a disease. This knowledge helps them to make more accurate treatment decisions. Also, patients won't suffer from the exorbitant costs in cases where treatment is not effective. In several studies on survivability of breast cancer patients, prognostic factors such as tumor grade, estrogen receptor (ER), progesterone receptor (PR), and tumor size were investigated. Since in many cases the disease is diagnosed at an advanced stage, the studying of prognostic factors for earlier detection of the disease can be effective. Data mining as a decision support system for predicting survival of new patients is a great advantage. It is a new topic for health researchers that are trying to find the relationship between risk factors and cancer survival [21].

III. RELATED WORKS

Today, treatment decisions for cancer patients are done specifically for each patient. Since performing specific tests is costly, focus on prognostic factors may be effective for next treatment decisions. Finding a predictive model to determine the survival of cancer patients is one of the ways to improve health care, programs related to cancer and help to evaluate the effectiveness of new therapies. In this section, some related works that have utilized data mining techniques for predicting breast cancer survivability are reviewed.

Delen et al. [11] applied two data mining techniques such as Artificial neural network (ANN) and decision tree (C5) with the most common statistical Logistic Regression method to SEER data set to predict 5-year survivability of breast cancer patients. 16 predictor variables were used for predicting dependent variable. The dependent variable was classified into two classes "survived" and "not survived" according to Survival Time Record (STR). Decision tree (C5) showed the best performance based on 3 metrics of accuracy, sensitivity, and specificity. Bellaachia and Guven. [22] used SEER data set to predict breast cancer survival. This paper unlike [15] used 2 other variables (Vital status record and Cause of Death) with STR for classifying dependent variable. After data preprocessing, three data mining techniques such as Naïve Base, Back-propagated Neural Network and decision tree (C4.5) were used for predictive modeling. Among three techniques C4.5 showed better performance.

Chao et al. [23] Utilized support vector machine, logistic regression, and C5 decision tree to make a predictive model of breast cancer survival with the aim of offering a treatment decision-making reference for women with breast cancer in Taiwan. Among these three techniques, SVM showed the best result based on average accuracy. Endo et al. in [24] studied another version of SEER data set to predict survival for breast cancer patients. Depended variable classifying into two classes: "survived" 81.5% and "not survived" 18.5%. Then 5 data mining algorithm such as ANN, NB, decision tree (ID3), decision trees (J48) and Logistic Regression were applied to the data set to build a predictive model. Logistic Regression provided best result in term of accuracy with 85.8%. Thongkam et al. [25] proposed a hybrid approach to improve breast cancer survival prediction by qualifying data set. First identified and eliminated outliers based on C-SVC technique and then solved skew data problem by oversampling with replacement. Four data mining techniques like AdaBoost, Bagging, C4.5 and SVM applied to data set to evaluate the hybrid approach. SMV showed the best performance based on accuracy, sensitivity, and specificity.

Liu et al. [26] used under-sampling method to deal with skew problem to build more accurate prediction model for breast Cancer survival. C5 decision tree was applied to balanced data set to evaluate the accuracy of proposed approach. Despite the 85.5% accuracy, the ability to identify the dead samples was weak with 0.2325 in term of specificity. Afshar et al. in [9] utilized data mining techniques to build a predictive model for breast cancer survival and also to explore the relationship between "specific predictor variables" and "Survival". In this study imbalanced

Problem was solved by an over-sampling method. Wang, et al. [27] used SMOTE to balance SEER data set for prediction breast cancer survivability. After data balancing, PSO [28] and classifiers combined for obtaining most relevant feature for classification. Results showed that adding SMOTE to PSO + classifiers can sufficiently improves the effectiveness of classification for massive imbalanced data sets. Most standard prediction methods can't handle incomplete records with missing values. Missing data imputation approach is widely used to solve this problem. Garcia-Laencina et al. in [29] used three methods such as Mode imputation, K-Nearest Neighbors Imputation, and Expectation-Maximization Imputation to deal with missing values. The dataset used in this study is based on data gathered from the Institute Portuguese of Oncology of Porto with 399 records. More than 18 percent of the data has been missing values. For building a predictive model for the survival of breast cancer four data mining approaches such as K-Nearest Neighbors, Classification Trees, Logistic Regression and Support Vector Machines were used. In the best case SVM offered 81.73% of accuracy and 0.78 of area under the Receiver Operator Characteristic curve.

TABLE I. Comparison of related works

Work	***Data set***	***Number of records after preprocessing***	***Best classifier + accuracy***	***Evaluation metric***	***Solution for Missing value***	***Solution for Imbalanced problem***
Delen et al. (2005)	SEER	202,932	C5=93.6%	Accuracy, Sensitivity Specificity	Deletion	–
Bellaachia and Guven (2006)	SEER	151,86	C4.5=86.7%	Accuracy, Precision, Recall	Deletion	–
Endo et al. (2007)	SEER	37,256	LR=85.8%	Accuracy	Deletion	–
Thongkam et al. (2009)	Data obtained from Srinagarind Hospital in Thailand	1479	C-SVC+SVM= 98.13%	accuracy, sensitivity, specificity, AUC[1], F-measure	–	Oversampling with replacement
Liu et al. (2009)	SEER	182,517	C5=88.05%	accuracy, sensitivity, specificity, AUC	Not exactly mentioned	Under-sampling
Chao et al. (2014)	Data obtained from specific hospital in Central Taiwan	1,340	SVM=98.34%	Accuracy, type I and type II error	–	–
Afshar et al. (2014)	SEER	22,763	SVM=96.7%	Sensitivity, Specificity, Accuracy, Adjusted Propensity scores	Multiple imputation	Oversampling
Wang et al. (2014)	SEER	215,221	PSO+C5=94.25%	G-mean, Accuracy, Sensitivity, Specificity	Deletion	SMOTE
Garcia et al. (2015)	Data gathered from Institute Portuguese of Oncology of Porto	399	KNN+ KNNimp = 81.73%	accuracy, sensitivity, specificity, AUC	Mode imputation, K-Nearest Neighbors Imputation, and Expectation-Maximization Imputation	–

[1] Area Under the receiver operating characteristic Curve

IV. CONCLUSION

According to related works and comparison in Table I, we can see that there are two basic problems for building prediction model of breast cancer survivability. One is missing value problem and another is the imbalance problem that causes quite different results for preprocessing step and makes predictive model not reliable.

Data quality is an important issue for building accurate predictive model, especially for medical data set. Since the medical data often are collected without any particular research goal, so usually has quality problem, such as missing data and outliers. As mentioned earlier most standard prediction methods can't handle incomplete records with missing values. Also, recently with advances in diagnosis and treatment of cancer, death rates have fallen and survival time has improved as well, so the breast cancer data sets have been imbalanced. Classification of an imbalanced data set is a challenging issue for researchers. Most standard data mining techniques consider balanced data set and when they work with imbalanced data set, results are biased toward numerous majority class samples. So the accuracy of classification for majority class is high and is low for minority class. This can be considered as a good performance in terms of simple accuracy, but in some cases, simple accuracy isn't a good criterion because it is important for the classifier to detect minority class. Especially In medical, the ramifications of such a result can be critical.

There are some problems with solutions that were presented for imbalanced problem in the previous papers. For example random undersampling method might eliminate valuable information that will be useful for building a predictive model. On the other hand random oversampling by generating too many repeated samples for minority class increases the likelihood of over-fitting. In this situation data mining techniques give good performance based on the accuracy of training set but for unseen records in test set lead to weak performance. Also, SMOTE based approach in some cases with presence of outliers degrades classification performance and extremely change data distributions like resampling methods.

As we know data mining is a process that extracts hidden predictive information from large databases. So having a data set with a large number of records is important for presenting useful knowledge. One question arises, whether the solution for missing value that was presented in previous research is efficient for a large volume of data? And the result can be used in real and practical situations? Since in [24] missing value problem was solved for a data set with 399 records.

Finally, it should be noted that for building a more accurate predictive model for breast cancer survivability these challenges should be considered and solved by an approach that takes into account all aspects. Because the model without solving these challenges might not practicable for real conditions.

REFERENCE

[1] N. Lavrac, "Selected techniques for data mining in medicine," Artif Intell Med, vol. 16, pp. 3-23, 1999.

[2] M. R. Parsaei, M. Salehi, "E-mail spam detection based on part of speech tagging", 2015 2nd International Conference on Knowledge-Based Engineering and Innovation (KBEI), Proceeding of the 2015 IEEE, pp. 1010 – 1013, November 2015.

[3] M. R. Parsaei, R. Javidan, and M. J. Sobouti, "Optimization of Fuzzy Rules for Online Fraud Detection with the Use of Developed Genetic Algorithm and Fuzzy Operators," Asian Journal of Information Technology, vol. 15, no. 11, pp. 1856-1864, 2016.

[4] M. R. Parsaei, R. Taheri and R. Javidan, R, "Perusing The Effect of Discretization of Data on Accuracy of Predicting Naïve Bayes Algorithm," Journal of Current Research in Science, (1), pp. 457-462, 2016.

[5] M. R. Parsaei, S. M. Rostami and R. Javidan, "A Hybrid Data Mining Approach for Intrusion Detection on Imbalanced NSL-KDD Dataset," International Journal of Advanced Computer Science & Applications,vol. 7, no. 6, pp. 20-25, 2016.

[6] S. S. Parsa, M. Sourizaei, M. M. Dehshibi, R. E. Shateri, M. R. Parsaei, "Coarse-grained correspondence-based ancient Sasanian coin classification by fusion of local features and sparse representation-based classifier," Multimedia Tools and Applications, 1-26, 2016, doi:10.1007/s11042-016-3856-6.

[7] G. Richards, V.J. Rayward-Smith, P.H. Sonksen, S. Carey, and C. Weng, "Data mining for indicators of early mortality in a database of clinical records," Artif Intell Med, vol. 22, pp. 215-231, 2001.

[8] J. Cruz, and D. Wishart, "Applications of machine learning in cancer prediction and prognosis," Cancer informatics, vol. 2, pp. 59-77, 2007.

[9] H. Lotfnezhad Afshar, M. Ahmadi, M. Roudbari, and F. Sadoughi, "Prediction of breast cancer survival through knowledge discovery in databases," Glob J Health Sci, vol. 7, pp. 392-8, 2015.

[10] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," AI Magazine, vol. 17, pp. 37-54, 1996.

[11] N. Esfandiari, M. R. Babavalian, A.-M. E. Moghadam, and V. K. Tabar, "Knowledge discovery in medicine: Current issue and future trend," Expert Systems with Applications, vol. 41, pp. 4434-4463, 2014.

[12] WHO "Cancer". World Health Organization. . Available: http://www.who.int/mediacentre/factsheets/fs297/en/.

[13] Breast cancer Q&A/facts and statistics. http://www.komen.org/bci/bhealth/QA/q_and_a.asp.

[14] J. G.-R. J. Jerez-Aragone´s, G. Ramos-Jimenez, J. Munoz-Perez, E. Alba Conejo, " A combined neural network and decision trees model for prognosis of breast cancer relapse," Artificial Intelligence in Medicine, vol. 27, pp. 45-63, 2003.

[15] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," Artif Intell Med, vol. 34, pp. 113-27, Jun 2005.

[16] American Cancer Society "Report sees 7.6 million global 2007 cancer deaths" Reuters.

[17] U. Khan, H. Shin, J.P. Choi, M. Kim,, "wFDT - Weighted Fuzzy Decision Trees for Prognosis of Breast Cancer Survivability," presented at the AusDM, 2008.

[18] Ministry of Health and Medical Education, Center for Disease Management, Cancer Department. National Registration Cancer Cases Reported in 2010. Iran: Center for Disease Management; 2013. p.344-9.

[19] D. K. S. Gupta, and A. Sharma, "Data mining classification techniques applied for breast cancer diagnosis and prognosis," Indian Journal of Computer Science and Engineering, vol. 2, pp. 188-193, 2001.

[20] D. K. S. Gupta, and A. Sharma, "Data mining classification techniques applied for breast cancer diagnosis and prognosis," Indian Journal of Computer Science and Engineering, vol. 2, pp. 188-193, 2001.

[21] M. Movahedi, S. Haghighat, M. Khayamzadeh, A. Moradi, A. Ghanbari-Motlagh, H. Mirzaei, et al., "Survival rate of breast cancer based on geographical variation in iran, a national study," Iran Red Crescent Med J, vol. 14, pp. 798-804, 2012.

[22] A. Bellaachia, and E. Guven, "Predicting Breast Cancer Survivability using Data Mining Techniques," presented at the Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining, 2006.

[23] Ch-M. Chao, Y-W. Yu, B-W. Cheng, and Y-L. Kuo, "Construction the Model on the Breast Cancer Survival Analysis Use Support Vector Machine, Logistic Regression and Decision Tree," J Med Syst, vol.38, 2014.

[24] A. Endo, S. Takeo, and H. Tanaka, "Predicting Breast Cancer Survivability: Comparison of Five Data Mining Techniques," Journal of Korean Society of Medical Informatics, vol. 13, no 2, pp. 177 180, 2007.

[25] J. Thongkam, G. Xu, Y. Zhang, and F. Huang, "Toward breast cancer survivability prediction models through improving training space, " Expert Systems with Applications, vol. 36, pp. 12200-12209, 2009.

[26] Y-Q. Liu, Ch. Wang, L. Zhang, "Decision Tree Based Predictive Models for Breast Cancer Survivability on Imbalanced Data," 3rd International Conference on Bioinformatics and Biomedical Engineering, 2009.

[27] K.-J. Wang, B. Makond, K.-H. Chen, and K.-M. Wang, "A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients," Applied Soft Computing, vol. 20, pp. 15-24, 2014.

[28] Nabaei, A., Hamian, M., Parsaei, M. R., Safdari, R., Samad-Soltani, T., Zarrabi, H., & Ghassemi, A. (2016). Topologies and performance of intelligent algorithms: a comprehensive review. Artificial Intelligence Review, 1-25, doi:10.1007/s10462-016-9517-3.

[29] P. J. Garcia-Laencina, P. H. Abreu, M. H. Abreu, and N. Afonoso, "Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values," Comput Biol Med, vol. 59, pp. 125-33, Apr 2015.