Medical Data Mining and Predictive Model for Colon Cancer Survivability

Narges Alizadeh Noohi[#1], Marzieh Ahmadzadeh[#2],Majid Fardaei[#3]

#1 Department of IT,Shiraz University of Virtual Education ,Shiraz,Iran.
#2 School of IT and Computer Engineering Shiraz University of Technology,Shiraz,Iran.
#3 Genetics Department Shiraz University of Medical Sciences,Shiraz,Iran.

ABSTRACT

Health care industry is the first candidate for data mining and the colon cancer is one of the most prevalent cancers in the world that has enough potential to be studied by data mining techniques. The present study has analyzed the SEER data that pertained to patients suffering from colon cancer in order to extract an accurate model of patients` survival by using data mining techniques.
To achieve an accurate model, preprocessing steps such as deletion, correction, and segregation has been done, which resulted in selecting nine features out of all to be later used in the selected data mining algorithms including Decision Tree, Bayes Networks, and Neural Network data mining. The model created by these algorithms, we created models which can predict mortality rate in three categories including less than a year, between 1 to 5 years, and more than 5 years.
Running several experiments using three mentioned algorithms showed that the most accurate model in predicting survivability of Colon cancer is Neural Network.

Keywords: Colon Cancer Survivability , Data Mining, SEER, Decision Tree algorithm, Bayes Network, Neural Network, Classification.

Corresponding Author: Narges Alizadeh Noohi

INTRODUCTION

One of the applications of data mining is in health care industry because there is a huge amount of clinical data available for analysis. Although a small part of data out of huge available data is useful for cure and prevention, but still this small part is large enough to be analyzed by data mining techniques [1]. Cancer is one of the main reasons for mortality in the world. The universal statistics shows that in 2008, 12.7 millions cases of cancer were reported. More than 40% of recognized cases include lung cancer, breast cancer in women, colon and stomach cancer[2]. Most common cancers in men are prostate, lung, and colon cancer, and in women are breast, lung, and colon cancer. Colon cancer stands in third place in terms of prevalence and second place in mortality[3].
This study attempts to discover and derive useful model in order to forecast the survival of colon cancer's patients using data mining techniques and data that were presented in SEER database . Clementine Software was used as a tool for data mining process. There is enough information available in literature about the factors which cause this illness but little research if any at all is

available that could help discovering the survival period of patients who suffer from colon cancer. This is the focus of this study that we hope could help practitioners and specialists in their decision making.

Research that were presented in literature regarding in the survival of cancer patients are usually based on binary prediction, i.e. whether or not a patient is alive. This study however focuses on modeling the patients' survival period based on categorical data type, i.e. the model will predict whether a patient survives less than a year, between 1 to 5 years, and more than 5 years.

RELATED WORK

There are a few research reported regarding breast cancer using data mining techniques. One of this kinds of research was done by Kadam et al. in which two data mining algorithms (a hybrid of Artificial Neural Network and Decision Tree) and Regression Logistic method was used for their experiment to forecast the survival of breast cancer patients. In this study, 200000 records of data were used and 10-fold-cross validation method were applied. The results showed that Decision Tree has presented the highest accuracy amongst other methods. Artificial Neural Network and Regression Logistic's statistic stood in next stage with lower accuracy respectively [4].

Bellaachia and Guven used the SEER data in their study in 2006 and applied three data mining techniques to find the best algorithm in forecasting patients' survival of breast cancer. The three techniques that were used included simple Bayes, Neural Network, and C4.5 Decision Tree. The Weka software was used as a tool of data mining. The output of this study was based on binary result, i.e. predicting a patient will be dead or alive in next 60 months. Results showed that although the accuracy of Neural Network and C4.5 Decision Tree were comparable but C4.5 Decision Tree had a higher accuracy [5].

Endo et al. investigated patients' survival of breast cancer in a 5 year period in 2008 which was recorded in SEER database from 1992 to 1997. They used 7 data mining algorithms including Artificial Neural Network, simple Bayes, pure Bayes, Decision Tree with simple Bayes, ID3Decision Tree, J48Decision Tree, and Regression Logistic model. Results showed that Regression Logistic and Decision Tree have had the best accuracy and sensitivity. Both last studies showed that there were a significant imbalance in data records in terms of patients "being alive" and "not being alive". This affects the accuracy of predicting model[6].

Survival modeling in other cancers are not studied as much as what has been done in breast cancer. But a few studies in lung cancer can be pointed out. For instance, Chen et al. used clustering techniques on SEER database in 2009 on lung cancer's data which were recorded from 1988 to 1998. For the purpose of their study, they selected features such as cancer developing level, grade, type of tissue and material and formed seven different clusters out of this data[7].

F.D used SEER data related to lung cancer from 1988 to 2001 to study patients' survival in a 8-months period. They used Regression Logistic algorithm and SVM data analysis, which resulted in better performance of Regression Logistic. SVM model which has a significantly low speed, had a high learning speed[8].

Agrawal et al. analyzed information of SEER database related to lung cancer in order to present a model for forecasting patients' survival. They used the data which were recorded in SEER database from 1998 to 2001. For this study they used data mining techniques in forecasting survival of patients with cancer of respiratory tract at the end of 6-month, 9-month, 1-year, 2-

year, and 5-year periods. In this study 10 algorithms of data mining were used including SVM, Artificial Neural Network, J48Decision Tree, etc. The results of this study showed that five of the algorithms including J48Decision Tree, Random Forests , LogitBoost, Random Subspace, and Alternating Decision Tree worked in a timely manner. They used voting methods to evaluate the algorithms which resulted in better accuracy for Decision Tree algorithm. Since SVM and Neural Network, they suggested that these kinds of algorithm are not appropriate for large number of data sets [9].

There is also a few studies reported in colorectal cancer such as the study of Grumett et al. in 2003 in which They used Neural Network algorithms to forecast patients' survival who were suffering from colorectal cancer. In their study, they compared two methods, Regression Logistic and Neural Network and it was shown that Neural Network had %78 and Regression Logistic had %66 of accuracy. Neural Network algorithm had a higher accuracy in large amounts of data in comparison with Regression Logistic. Regression Logistic had the advantage of creating a more simple model which is easier to be applied , while modeling in Neural Network had some complexities[10].

Fathi created a model to forecast the survival of colorectal cancer by using Artificial Neural Network algorithm in 2011. Results showed that Neural Network could be an appropriate suggestion for forecasting the survival of patient suffering from colorectal cancer. Errors could be decreased by changing the hidden layers during training. Also it can be concluded that the number of input characteristics is not important, but the important thing is choosing characteristics which increase accuracy[11].

Studying the prior investigations had many results. First, data mining research in colon cancer has been of less interest to researchers. That is why colon cancer is chosen for more investigation in current research. Second, different algorithms will work differently on different types of data in various cancers. Therefore a different investigation is needed for colon cancer in order to recognize the most efficient algorithm for data type of colon cancer. Third, a number of algorithms found to have the best prediction ability in different cancer. Therefore these algorithms were chosen for this study, which includes Neural Network algorithm, Decision Tree and Bayes Network.

MATERIALS AND METHODS

In this section, the steps for forecasting survival time are explained. The explanation about data is given in part 1 and data selection and preparation is brought in part 2. Explanation about algorithm and the evaluation of the research is given in part 3 and 4 respectively.

3-1. Data sources

SEER, which is abbreviation for "Surveillance, Epidemiology and End Results"[12], is a national valid cancer institute which is considered as a source of cancer statistics in U.S.A[13].

SEER database was established by America's government to collect statistic information of cancer patients in this country. Legally, all of the hospitals, clinics, laboratories, surgery sections and organization related to diagnosis and treatment of cancer have to report the information to this institute, which will be then inserted to this database after evaluation Attributes of SEER data can be considered in several different sections. Demographic attributes(age of diagnosis, sex, location), diagnosis attributes (the first part of the risk, level), treatment attributes (surgery,

radiation) and result attributes (survival, mortality's cause) are the information which make SEER's information an appropriate data for scientific study and analysis. This includes several databases and we used the most complete and last updated data base for this study. This station is related to 2011 which concludes all colon cancer information in America from 1969 to 2010.

3-2. Data Selection and preparation

We have chosen attributes from each group of properties which are more important based on clinical studies in survival. Following characteristics are chosen from data that related to patients of colon cancer for this study:

- Demographic attributes(age of diagnosis, sex, race)
- Recognition attributes (behavior, grade, Site rec[1] with Kaposi[2] and mesothelioma[3])
- Treatment attributes(radiation, radiation sequence with surgery)
- Result attribute (patient's life status, mortality cause, survival time record)

Appropriate preprocessing in every kind of forecasting like cancer survival, is very vital. Generally we did following steps to clean up the data:

1- Deletion of patients' information those are dead because of a reason other than cancer.
2- Deletion of some rare properties which do not significantly affect forecasting. For instance, there was just one record related to colon cancer in age 10 to 14 which we deleted it. Because this record does not have much importance and does not affect the forecasting.
3- After categorizing data by two properties, mortality cause and life statue, we kept the records that the corresponding patient was still alive and omitted the rest. This study is about people who lost their life because of colon cancer and their death dates were known.
4- Deletion the record with missing values

The total number of collected data was 8138 cases which had been reduced to 5276 cases after cleaning up steps.
Cancer patients' survival is different based on the type of the cancer. For example, maximum patient's survival time with malignant brain cancer is 2 years[14].while other cancers such as colon cancer have more probability of survival, therefore their survival time category can place a larger period. Therefore for colon cancer, the best time division is as follows:

1- Survival time record for patient is 1 year (0-12 months)(class A).
2- Survival time record for patient is more than 1 year and less than 5 years (13-60 months)(class B) [15].
3- Survival time record for patient is more than 5 years (61-322 months)(class C).

3-3. Algorithms

Decision Trees: according to some rules, decision is used for forecasting and classification. In this study we considered the results of forecasting in three clusters, so using Decision Tree

[1] Recreation
[2] Kaposi sarcoma is a multicentric, malignant neoplasm of a complex and still unclear etiology
[3] Mesothelioma is a type of cancer that develops in the mesotheliom which is a protective layer covering most of internal body parts

algorithm seems efficient. Decision Tree models appear as a set of "if-then" rules which shows information in a complete form for several cases. Since the inputs of this study is categorical , the result is a categorical tree too. Out of all decision tree algorithms, CHAID and C5.0 algorithms are suitable for categorical for categorical class variable.

Neural Network: this algorithm which is another classification algorithm , processing takes place in several layers. Usually there are three layers in a Neural Network; an input layer, a hidden layer, and an output layer. Inputs are connected to each other by different weights. We used four Neural Network methods (Quick, Dynamic, Multiple, RFBN) which were suitable for our data.

Bayes Network: it allows us to create a probability model. This group of algorithms is presented as TAN and Markoveblanket nets in current version of Clementine, which is the software that was used in this study.

3-4. Evaluation level

In this study we used hold-out method to evaluate the accuracy of classification and forecasting. This method divides data in two independent sections (train, test) randomly. Generally, two third of data is related to training section, and the rest is used for testing.

features such as age, race, site rec with Kaposi and mesothelioma, grade, type of radiation, the order of radiation with surgery, and survival time were defined as a set type, because each of which includes more than two parameters. other features such as sex and type of cancer defined as flag type due to their binary nature. All of these attributes were chosen as an input to our model except for survival time, which is the output variable.

Using Clementine software, we created models by Decision Tree, Bayes Network, and Neural Network algorithms several times to get an optimized result. We then compared then models based on percentage of achieved accuracy in train and test data.

In each algorithm, we choose the model which has the best accuracy. The result of comparison of these three algorithms can be seen in fig.1. The vertical axis shows the percentage of accuracy of forecasting in each model and horizontal axis shows the result of each model in train and test data.

As shown in fig.1, C5.0 Decision Tree model showed %60.5 accuracy in test data and %59.53 accuracy in train data. Bayes Network algorithm in Markove method had %60.76 accuracy in test data and %59.09 accuracy in train data while Neural Network algorithm in Multiple approach showed %71.61 of accuracy in test data and %71.15 of accuracy in train data. Neural Network algorithm therefore has a higher accuracy in test and train data in comparison.

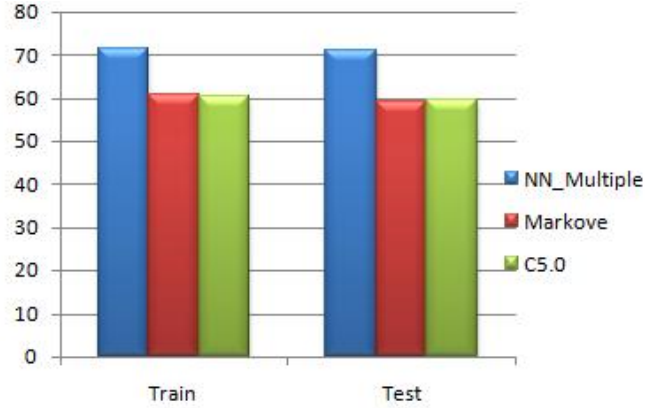

Fig 1: Comparison the percentage of three algorithms

The important variables should be involved in creating a model. Total weight of all variables' relative weight must be add up to 1.0 . Therefore, the importance of each variable should be between 0 to 1. It should be noticed that the importance of variables is not related to model accuracy.

The importance of each variable in forecasting the best model which is related to performed Neural Network algorithms by Multiple method, is shown in fig.2.

What is understood by importance order of features is that for all models the important variables for forecasting are age, Site rec with Kaposi and mesothelioma, radiation sequence with surgery, grade, type of radiation, race, sex and behavior respectively.

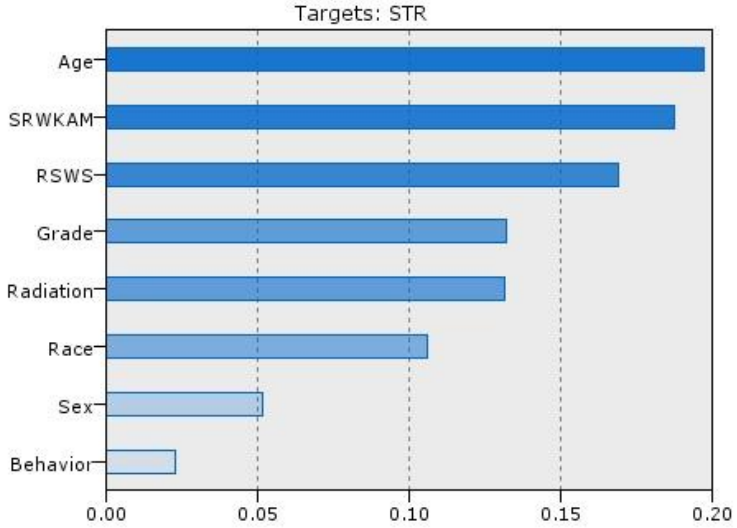

Fig 2: The diagram of importance properties in neurotic net performed by Multiple method.

CONCLUSION AND FUTURE WORK

We presented an useful model for forecasting patients' survival of colon cancer by using Neural Network algorithm in medicine data related to SEER database in this research.

It's understood from this study that in data related to colon cancer's patients which are recorded in SEER data base, Neural Network algorithms has a better accuracy in comparison with Decision Tree and Bayes Network algorithms. But this algorithm has lower learning speed and more complexity. Also the importance of variables to forecast the survival time of colon cancer patients in Neural Network model, which proved to be the best amongst all, are age, Site rec with Kaposi and mesothelioma, radiation sequence with surgery, grade, type of radiation, race, sex, and behavior respectively.

For future studies, one can create new models by other data mining algorithms on colon cancer data related to SEER database and compare it to the models of this study. The result of this investigation is that Neural Network algorithm is better than two other algorithms in this field and for this kind of data. But even this algorithm does not show very high accuracy. Therefore is the optimization of this algorithm in order to improve the accuracy to an acceptable level is suggested.

REFERENCE

[1] Illya.Mowerman, Data mining in the health care industry, University of Rhode Island, 2007.

[2]International Agency For Research On Cancer, World Cancer Factsheet,.World Health Organization, www.cancerresearchuk.org, 2012.

[3] AJ.Sasco, MB.Secretan andK .Straif, Tobacco smoking and cancer: a brief review of recent epidemiological evidence, Lung Cancer 45 Suppl 2: S3–9, 2004.

[4] Dursun.Delen, Glenn.Walker and Amit.Kadam, Predictingbreast cancer survivability:a comparison of three data mining methods, Department of Management Science and Information Systems, Oklahoma State University , 2004.

[5] Abdelghani.Bellaachia and Erhan.Guven, Predicting Breast Cancer Survivability Using Data Mining Techniques, Department of computer Science The George Washington University, 2005.

[6] A.Endo, T.Shibata and H.Tanaka, Comparison of seven algorithms to predict breast cancer survival,Biomedical Soft Computing and Human Sciences,2008.

[7] D.Chen, K.Xing. D.Henson, L.Sheng, A.Schwartz and X.ChengDeveloping prognostic systems of cancerpatients by ensemble clustering, Journal ofBiomedicine and Biotechnology, 2009.

[8] F. D, Machine learning methods in the analysis of lung cancer survival data, DIMACS Technical Report, 2006.

[9] Ankit.Agrawal, Sanchit.Misra, Ramanathan.Narayanan, Lalith.Polepeddi and Alok.Choudhary, A Lung Cancer Outcome Calculator Using Ensemble Data Mining on SEER Data".Electrical Engg. and Computer ScienceNorthwestern University, 2011.

[10] Simon.Grumett, Pete.Snow and David.Kerr, Neural Networks in the Prediction of Survival in Patients with Colorectal Cancer, Clinical Colorectal Cancer, Vol. 2, No. 4, 239-244, 2003.

[11] SHERIF KASSEM.FATHY, A Predication Survival Model for Colorectal Cancer, Information System Department, College of Computer and Information TechnologyKing Faisal University SAUDI ARABIA, 2011.

[12] Surveillance, epidemiology, and end results (seer) program (www.seer.cancer.gov) limited-use data (1973-2006). National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, 2008.

[13] Overview of the seer program. Surveillance Epidemiology and End Results, URL: http://seer.cancer.gov/about/ accessed, 2010.

[14] Walter G.Bradly, Robert B.Danoff,Gerald M.Fenichel and Joseph.Jankowi, NEUROLOGY IN CLINICAL PRACTICE, Butterworth-iteinemann, an imprint of Elsiver inc.Fifth edition, 2008.

[15]Michael.J.Zinner and Stanley.W.Ashley, Maingot's Abdominal Operations, McGraw-Hill companies,Inc.Printed in the United States of America,11 th edition,2007.