Acta Technica *62* (2017), No. 6A, 415–422

Comparing 8 different learning algorithms to predict the survival of colorectal cancer for 5 years

Fan Qi[1], Xu Zhengtong[1]

Abstract. Medical prognosis is an important field of medical research, which predicts the patient's survival state around the concurrent and recurrence of disease. In the traditional statistical analysis, the medical prognosis takes the life table, the Kaplan Meier test and constructs the Cox-Proponal risk model to carry on the research. Nowadays, with the extensive application of data mining in medical field, data mining has been applied to the field of medical prognosis gradually. As one of the main causes of human morbidity and mortality, cancer prognosis is a particularly important study. It is very important to predict the survival state of cancer patients in a certain period. In this paper we compare 8 different learning algorithms to predict five-year survival of colorectal cancer and to provide the assistant decision-making effect for medical treatment of cancer prognosis. The results of this study show that decision tree (C 5) had the highest accuracy of 84.07%, and it is superior to other algorithms in the sensitivity and specificity of colorectal carcinoma data.

Key words. Data mining, colorectal cancer, prognosis prediction.

1. Introduction

Colorectal cancer is one of the common cancers occurring in the digestive tract, and its morbidity and mortality rates are among the highest in many cancers, with 774,000 deaths worldwide in 2015 (WHO, 2017). Even though colorectal cancer is a relatively high rate of morbidity and mortality, there has been a significant downward trend in the number of new cases and deaths since 1992-2013, according to the statistics of National Cancer Center of the United States in recent years. This has important relationship with early screening, early diagnosis and treatment of colorectal cancer.

Colorectal cancer is a kind of cancer that occurs in the lower part of the digestive tract. The large intestine is mainly composed of rectum, colon, cecum, appendix

[1]School of Computer Science and Technology, Huaibei Normal University, Huaibei, Anhui 235000, China

and anus Tube; however, colorectal cancer is a general term for colon and rectal cancers.

The incidence of colorectal is related to individual dietary habit, family heredity and intestinal inflammatory disease, and the prevalence rate of the city is much higher than that in rural area. Early screening and treatment of colorectal cancer can significantly improve the survival rate and cure rate of colorectal cancer patients. According to the analysis of the SEER database of national Cancer Research in the United States, the proportion of patients receiving chemotherapy in the stage I and stage II of colon cancer was increased from 1987 to 2010 (National Cancer Institute, 2017).

According to the analysis of 5-year relative lifetime of patients with colorectal cancer from 2006 to 2012 in SEER database, we found that local colorectal cancer survival rate was up to 90.1% (National Cancer Institute, 2017). The treatment methods of colorectal are mainly surgical operation, chemical treatment, target therapy and radiotherapy.

Survival analysis is an important part of cancer prognosis therapy. It is of great significance for doctors to provide individualized diagnosis and treatment for the diagnosed cancer patients by constructing the survival prediction model of cancer patients and analysing the factors that influence the survival period. The traditional statistical analysis method has the life table, Kaplan Meier test and constructs the Cox-Proponal risk model. With the improvement of knowledge discovery and data collection and analysis technology, data mining technology has been involved in the field of medical prognosis, which promotes the development of cancer survival prediction.

Cancer survival is the percentage of cancer patients who survive a certain period of time. The cancer survival rate is usually discussed in the five-year relative survival period for colorectal carcinoma. The five-year relative survival is also a reference basis for the analysis of cancer patients' survival period (Dursun,2005).

2. Related work

In 2004 Delen used decision tree, artificial neural network and logistic regression method to construct survival prediction model of 1973-2000 American breast cancer data collected by SEER database, and found that decision tree has the highest predictive accuracy (Dursun, 2005). The detailed description and unique insights of data mining techniques in predicting the survival of breast cancer patients have an important guiding and learning effect on later researchers. The Clinical Information Data Warehouse (CDW), developed by IBM, combines data warehousing and OLAP analysis tools to realize the multidimensional analysis and deep excavation of medical information in medical institutions, which has been successfully implemented in the HIWAS project of Guangdong Provincial Hospital in China (IBM, 2017). IBM Watson supercomputer, which is trained by IBM and the Memorial Sloan-Caitlin Cancer Center, uses machine learning techniques to extract vast amounts of medical literature as an adjunct to providing cancer solutions for physicians. In August 2016, 21 local hospitals in China planned to use IBM Watson's cancer solution (IBM

Watson for Oncology) (Cognitive care, 2016), one of the first commercial available projects in health. AI has developed rapidly in adjuvant cancer therapy.

From research literature of data mining and colorectal cancer, it is found that the current research in this field mainly involves several aspects: taking advantage of data mining techniques to track the effects of colorectal cancer markers; classification from the perspective of clinical data and gene data, which could optimize the diagnosis effect; predicting the risk of colorectal cancer; predicting the high-risk factors after operation, and constructing individualized treatment system to optimize the function of computer-aided decision-making.

In the study of marker tracking of colorectal cancer, Tan and Hu constructed three kinds of artificial neural networks to predict colorectal cancer by using three different markers in 106 serum samples. It was concluded that the artificial neural network established by combining protein markers and tumor markers, which shows high throughput, sensitivity and specificity in predicting colorectal cancer (Tan, 2011). In the molecular pathology study of colorectal cancer in Li [Li, 2011], the 251 serum proteins were screened out, then the artificial neural network was established to diagnose the prognostic model, and the accuracy of the identified proteome models in diagnosis and prognostic analysis was higher.

From the view of clinical data classification, based on the nearest neighbour algorithm of the search tree, the auto fluorescence spectra of the large intestine early cancer were classified, which reduces the computational complexity and optimize the storage mode (Fan, 2017). Cao, Zhuang and Lian used decision tree to classify the liver CT images of 60 patients with colorectal cancer, which are helpful for physicians to improve the efficiency of diagnosis and treatment (Cao, 2014). From the point of view of genetic data classification, Meng, Liu, and Wang applied SVM to classify the cancer data in the microarray, which method has an excellent improvement in the accuracy and speed of data classification (Meng, 2007). Huang took advantage of the characteristics of gene selective expression, applied KNN, SVM, ANN and other methods in five classical gene expression databases to determine the genetic category, and found that SVM improves the accuracy and interpretation of cluster analysis in gene expression classification in some degree (Huang, 2009).

In the study of risk prediction of the related factors of colorectal cancer, Zhang Yongjing used stratified analyses and classification and regression tree model to discover the complex interactions of genes to genes and genes to the environment, which affect the risk of colorectal cancer (Zhang, 2009). In the predictive study of high-risk factors in patients with colorectal cancer after operation, Zeng Zijie, Liu Xiufeng, Shi Jiayun analyzed the data of 292 patients who underwent colorectal cancer surgery. The main factors affecting lymphatic metastasis of colorectal carcinoma were obtained by logistic regression and artificial neural network algorithm, and the extent of its impact was sorted (Zeng, 2016). Zhaobin, Chen Fujun, Liu Yang and others used the retrospective analysis method to construct logistic regression model for 933 cases of patients receiving colorectal cancer surgery, and to discuss the risk, influencing factors and predictive value of predicting anastomotic leakage after colorectal surgery (Zhao, 2016).

Wang and Chen proposed the research of individualized diagnosis and treatment

system of colorectal cancer based on data mining technology, which helps doctors solve the problem of semi-structured and unstructured diagnosis (Wang, 2011).

David used 3 data mining algorithms to predict the 5-year survival of colorectal cancer, and found that ANN had the highest accuracy rate of 70% (David, 2007). However, their papers did not clearly describe the source and quantity of data used. Sherif and other people used neural network model to predict the survival of colorectal cancer, in which the accuracy rate is 84.73% (Sherif, 2011). The accuracy rate of using the minimal subset of attributes is 86.51%. Reda AI and other people using 1973-2009 years of SEER data, using ensemble learning to predict the 5-year survival period of colorectal cancer, the accuracy is 85.13%, but they did not compare them with the independent model, but also did not list other important performance evaluation indicators (Reda, 2013).

To sum up, we found that data mining techniques are used in many aspects of colorectal cancer research, but the amount of data used in the study is inadequate, and the study of five-year survival prediction for colorectal cancer patients is not comprehensive enough.

In this paper, we investigated 8 data mining algorithms to construct the five-year survival prediction model for colorectal cancer. The latest 1973-2014-year colorectal cancer data from U.S. National Cancer Institute, which is open, credible, and data-large, is used as research data, which contains 983,807 records and 133 attributes (National Cancer Institute, 2017).

3. Methodology

3.1. Dataset

The SEER project of the National Cancer Center of the United States aims to provide national-led cancer surveillance science, analytical tools and methodological expertise in the collection, analysis, interpretation and dissemination of reliable demographic data to reduce the burden of cancer in the American population [17]. SEER currently collects and publishes cancer morbidity and survival data, and covers about 28% of the United States population based on population-based cancer registries, regularly collects demographic data for patients, primary tumor sites, tumor morphology and diagnostic stages, first course of treatment, and follow-up of important status [16]. The project collects rigorous and reliable seer data sources that are used by a wide range of scientific researchers.

3.2. Data mining tool

This article uses the SPSS modeler (17.0) to complete the work. SPSS Modeler is an industrial level data mining software platform. Data can be integrated, analyzed, processed, and developed on this platform. The SPSS modeler encapsulates the complex data processing operation in the module, and carries on the processing. Nodes are connected sequentially through pipelines to build the data stream. In a visual graphical interface, the development of a data mining process can be

accomplished by simply dragging and setting.

3.3. Data preprocessing

Before constructing the cancer survival Prediction model, we pre-processed the SEER data. We removed the attributes with the missing ratio above 70% and the coefficient of variation below 0.1. In different time periods, there were four kinds of coding schemes stored in different attribute values for the fields of Tumor Size, Extension of Tumor, and the Highest Specific Lymph Node Chain, and we integrated them together to represent them in a coding scheme. We removed the records with the null or unknown value for the Tumor Size. To predict the survivability of a patient, we generate the Target attribute of the survivability. If Survival Months >=60 and Vital Status Recode is alive then the Target attribute of the record is marked for 'T' (survival); If Survival Month < 60 and Cause of Death to SEER Site Recode is rectal cancer or colon cancer then the Target attribute of the record is marked for 'F'(not survival). We removed extraneous attributes for the forecast target: such as Patient ID Number. We removed the records with the null or unknown value for the Target attribute, Extension of Tumor, the Highest Specific Lymph Node Chain, Regional Nodes Positive, and Regional Nodes Examined. After the above data pre-processing process, the final dataset contains 286, 269 Records, 31predictor variables.

Table 1. Predictor variables

Age at diagnosis	AYA site recode /WHO 2008	CS Schema v0204+	Diagnostic Confirmation
EOD—Extension	EOD—Lymph Node Involv	EOD—Tumor Size	First malignant primary indicator
Grade	Histology Recode—Broad Groupings	ICCC site rec extended ICD-O-3/WHO 2008	ICCC site recode ICD-O-3/WHO 2008
Laterality	Marital Status at DX	NHIA Derived Hispanic Origin	Origin recode NHIA (Hispanic, Non-Hisp)
Primary by international rules	Race recode (W, B, AI, API)	Race recode (White, Black, Other)	Race/Ethnicity
Reason for no surgery	Recode ICD-O-2 to 10	Regional Nodes Examined	Regional Nodes Positive
SEER historic stage A	SEER Record Number	Sequence Number —Central	Sex
State-county recode	Total Number of Benign/Borderline Tumors for Patient	Total Number of In Situ/malignant Tumors for Patient	

3.4. Model

In this article, we tested 7 classical independent algorithms: ANN(MLP), Decision Tree (C5), Bayesian Networks, Logistic Regression, Discriminant, Decision Tree

(CRT), and SVM. Then we took advantage of Ensemble Learning to combine the results of these 7 models.

4. Experimental results

To evaluate the performance of the predictive models, we separated the data from the whole data set, where 70% of the data was used as a training set to extract the data mining model, and 30% of the data were used as test sets to test the pros and cons of models and evaluation models. The detailed results are presented in the form of a confusing matrix to evaluate the results of a predictive model. In this experiment, the main performance evaluation index is accuracy, sensitivity, specificity, precision, and recall.

As shown in table 2, decision tree algorithm has the highest accuracy in the prediction model of colorectal cancer survival, and its sensitivity and specificity are superior to other algorithms.

Table 2. Results for all model types

Classification Technique	Accuracy(%)	Sensitivity	Specific	Class	Precision	Recall
ANN(MLP)	80.78	0.78	0.84	T	0.78	0.84
				F	0.84	0.78
Decision Tree (C5)	84.07	0.82	0.86	T	0.82	0.87
				F	0.86	0.82
Bayesian Networks	76.73	0.74	0.80	T	0.74	0.80
				F	0.80	0.73
Logistic Regression	78.85	0.77	0.81	T	0.77	0.82
				F	0.81	0.76
Discriminant	78.34	0.72	0.86	T	0.72	0.84
				F	0.86	0.73
Decision Tree (CRT)	79.46	0.79	0.80	T	0.79	0.81
				F	0.80	0.78
SVM	81.26	0.80	0.82	T	0.80	0.83
				F	0.82	0.79
Ensemble Learning	81.34	0.78	0.85	T	0.78	0.85
				F	0.85	0.78

5. Conclusions and future work

In this paper, we report the results of data mining techniques in colorectal cancer five-year survival prediction. We used 8 classical algorithms. In the future, we can test the current more popular algorithm models, such as deep learning.

This study shows the enormous potential of data mining technology in cancer medicine. However, the medical diagnosis and treatment of disease is very prudent, and it is the key to construct a reliable data mining model that meets the clinical

needs. The accuracy, sensitivity and specificity of the model as the basis for the diagnosis of specific diseases, and the credibility of the results of the data model prediction are also considered.

Acknowledgements

This study was sponsored by the Anhui University Natural Science Research Project of the Education Department of Anhui Province of P.R. China (grant no. KJ2017A390)

References

[1] Cognitive care, *21 Chinese hospitals will apply IBM Watson Cancer Solutions (IBM Watson for Oncology) to help physicians provide personalized cancer diagnosis.* http://www.cognitivecare.cn/archives/210, (2016).

[2] T. W. Cao, Q. L. Zhuang, Y. B. Lian. *Application of decision tree model in CT image classification of colorectal cancer patients.* CT Theory and Application Research *23*, (2014), No. 2, 275–283. (In Chinese)

[3] D. Chhieng, M. Hardin, B. Anderson, U. Manne. *Predicting 5-Year Survival of Colorectal Carcinoma Patients Using Data Mining Methods.* AMIA 2007 Symposium Proceedings, *907*, (2007).

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

[5] X. P. Fan, Z. F. Liao, Y.Z. Chen. *A new method for data processing of autologous fluorescence spectra of colorectal carcinoma*, 26th Session of China Control Conference, (2017). (In Chinese)

[6] D.S. Huang. *Application of gene expression data in tumor diagnosis and gene function prediction.* Chinese Medical University, (2009). (In Chinese)

[7] IBM. *CDW/Hospital BI Solution.* https://www.cancer.gov/types/colorectal/patient/colon-treatment-pdq, (2017).

[8] X. Q. Li. *Establishment of diagnostic and prognostic model of serum differential proteome in colorectal cancer*, Lu Zhou Medical College, (2011). (In Chinese)

[9] F. J. Meng, Y. H. Liu, H. G. Wang. *Application of SVM in the classification of gene microarray cancer data.* Shandong Institute of Light Industry, Institute of Information Science and technology, (2007). (In Chinese)

[10] National Cancer Institute. *About the SEER Program.* https://seer.cancer.gov/about/, 2017.

[11] National Cancer Institute. *Cancer Stat Facts: Colon and Rectum Cancer*, https://seer.cancer.gov/statfacts/html, 2017.

[12] National Cancer Institute. *Overview of the SEER Program.* https://seer.cancer.gov/about/overview.html, (2017).

[13] R. AI-Bahrani, A. Agrawal, A. Choudhary. *Colon cancer survival prediction using ensemble data mining on SEER data.* In Proceedings of 2013 IEEE International Conference on Big Data, (2013), 9–16.

[14] S. K. Fathy. *A predication survival model for colorectal cancer.* In Proceedings of the 2011 American conference on applied mathematics and the 5th WSEAS international conference on Computer engineering and applications, (2011), 36–42.

[15] S. H. Tan, L. M. Hu. *Preliminary study on 3 kinds of artificial neural network models for predicting colorectal cancer.* Laboratory Medicine and Clinical *8*, (2011), No. 8, 941–942. (In Chinese)

[16] C. Wang, X.L. Chen. *Study on personalized diagnosis and treatment system of colorectal cancer based on data mining technology.* Medical Information *24*, (2011), No. 8, 3528–3529. (In Chinese)

[17] WHO *CANCER.* `http://www.who.int/mediacentre/factsheets/fs297/zh/` (2017).

[18] Z. J. Zeng, X. F. Liu, J. J. Xie. *Prediction of high-risk factors of lymph node metastasis after colorectal cancer operation.* Chinese Journal of Gerontology *14*, (2016), 048. (In Chinese)

[19] Y. J. Zhang. *Epidemiological study of TGF-11 pathway gene polymorphism, environmental exposure factors and risk of colorectal cancer.* Zhejiang University, (2009). (In Chinese)

[20] B. Zhao, F. J. Chen, Y. Liu. *Regression analysis of risk factors and prediction of anastomotic leakage after resection of colorectal cancer.* Hei Long Jiang Medical Science *39*, (2016), No. 4, 162–163. (In Chinese)

Received September 12, 2017