Poster: A Lung Cancer Mortality Risk Calculator Based on SEER Data

Ankit Agrawal, Sanchit Misra, Ramanathan Narayanan, Lalith Polepeddi, Alok Choudhary
Dept. of Electrical Enginering and Computer Science,
Northwestern University
Evanston, USA
{ankitag,smi539, ran310,choudhar}@eecs.northwestern.edu, lpolepeddi@u.northwestern.edu

***Abstract*— We analyze the lung cancer data available from the SEER program for developing survival prediction models using data mining techniques. The prototype mortality risk calculator developed as a result of this study is available at info.eecs.northwestern.edu:8080/CancerMortalityRiskCalculator**

Keywords- data mining; lung cancer; mortality; risk calculator.

I. INTRODUCTION

Lung cancer is the second most common cancer, and the leading cause of cancer-related deaths among men and women in the USA. Survival rate for lung cancer is estimated to be 15% after 5 years of diagnosis [1]. Data mining techniques can be useful to estimate risk of mortality due to lung cancer based on diagnostic and treatment attributes.

II. SEER DATA

The Surveillance, Epidemiology, and End Results (SEER) Program [2] of the National Cancer Institute (NCI) is an authoritative repository of cancer statistics in the US. The SEER data attributes can be broadly classified as demographic attributes (e.g. age, gender, location), diagnosis attributes (e.g. primary site, histology, grade, tumor size), treatment attributes (e.g. surgical procedure, radiation therapy), and outcome attributes (e.g. survival time, cause of death), which makes the SEER data ideal for performing outcome analysis studies.

There have been numerous statistical studies using the SEER data, and also some studies for predicting breast cancer survival. Modeling survival for lung cancer is still in its in preliminary stage. Reference [3] applied ensemble clustering to SEER data to study survival patterns. Reference [4] used SEER data to study 8 months survivability using penalized logistic regression and SVM.

III. MODELING USING DATA MINING TECHNIQUES

We used the data from SEER November 2008 Limited-Use Data files [2] from nine SEER registries, which had a follow-up cutoff date of December 31, 2006, i.e., the patients were diagnosed and followed-up up to this date. In our experiments, we used the WEKA toolkit for data mining [5].

After several SEER-related (cancer-independent) and cancer-specific preprocessing steps, the set of predictor attributes was constructed (63 attributes), along with 5 binary outcome attributes - survival after 6 months, 9 months, 1 year, 2 years, and 5 years. We subsequently used several classification techniques along with various data mining optimizations and validations to model these five outcomes. We found that several of these techniques resulted in a high prediction performance (in terms of accuracy and area under the ROC curve).

IV. RISK CALCULATOR

Further, for the purpose of developing a risk calculator for mortality risk estimation, we used attribute selection techniques to identify a smaller non-redundant subset of predictor attributes. The goal here was to reduce the number of attributes for use in the risk calculator, while trying to retain the predictive power of the original set of attributes in the preprocessed data. We found that even with the much smaller number of attributes (15 in the current version), the quality of predictions was more or less maintained. Figure 1 shows a screenshot of the risk calculator.

We plan to build on the current work to identify optimal attribute-sets and classification techniques for accurate risk prediction for lung cancer mortality, apart from performing similar analysis for other cancers.

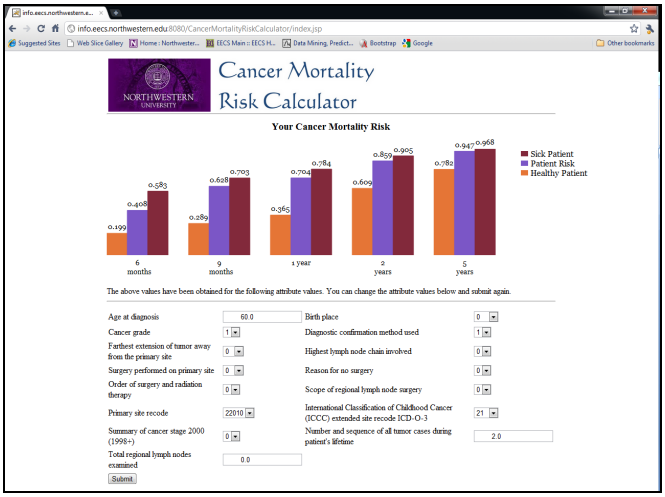

Figure 1 Screenshot of the prototype lung cancer mortality risk calculator

REFERENCES

[1] L. A. G. Ries and M. P. Eisner, Cancer of the lung. National Cancer Institute, SEER Program, 2007, ch. 9.

[2] "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, released April 2009, based on the November 2008 submission.

[3] D. Chen, K. Xing, D. Henson, L. Sheng, A. Schwartz, and X. Cheng, "Developing prognostic systems of cancer patients by ensemble clustering," Journal of Biomedicine and Biotechnology, vol. 2009, 2009.

[4] D Fradkin, "Machine learning methods in the analysis of lung cancer survival data," DIMACS Technical Report 2005-35 February 2006.

[5] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: An update," SIGKDD Explorations, vol. 11, no. 1, 2009.