ASSOCIATION RULE MINING BASED HOTSPOT ANALYSIS ON SEER LUNG CANCER DATA

Ankit Agrawal, Northwestern University, USA
Alok Choudhary, Northwestern University, USA

ABSTRACT

The authors analyze the lung cancer data available from the SEER program with the aim of identifying hotspots using association rule mining techniques. A subset of 13 patient attributes from the SEER data were recently linked with the survival outcome using prediction models, which is used in this study for segmentation. The goal here is to identify characteristics of patient segments where average survival is significantly higher/lower than average survival across the entire dataset. Automated association rule mining techniques resulted in hundreds of rules, from which many redundant rules were manually removed based on domain knowledge. Further, association rule mining based hotspot analysis was also conducted for conditional survival patient data, i.e., in cases where patients have already survived for a year after diagnosis. The resulting rules conform with existing biomedical knowledge and provide interesting insights into lung cancer survival.

Keywords: Association Rule Mining, Hotspots, Lung Cancer, Prediction Models, SEER

INTRODUCTION

Lung cancer ranks second in the list of most common cancers (National Cancer Institute, 2010), and first in the list of most deadly cancers (Centers for Disease Control and Prevention, 2010), with the survival rate being about 15% after 5 years of diagnosis (Ries & Eisner, 2007).

The Surveillance, Epidemiology, and End Results (SEER) Program (National Cancer Institute, 2008) of the National Cancer Institute is an authoritative repository of cancer statistics in the United States (National Cancer Institute, 2010). It is a population-based cancer registry which covers about 26% of the US population across several geographic regions and is the largest publicly available domestic cancer dataset. The data includes patient demographics, cancer type and site, stage, first course of treatment, and follow-up vital status. The SEER program collects cancer data for all invasive and in situ cancers, except basal and squamous cell carcinomas of the skin and in situ carcinomas of the uterine cervix (Ries & Eisner, 2007). The SEER limited-use data is available from the SEER website on submitting a SEER limited-use data agreement form.

DOI: 10.4018/jkdb.2011040103
Gloeckler Ries, Reichman, Lewis, Hankey, and Edwards (2003) present an overview study of the cancer data at all sites combined and on selected, frequently occurring cancers from the SEER data. 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.

With SEER data being available in the public domain, there is a mature literature on the statistics of SEER data (Yao et al., 2008; Rusthoven, Flaig, Raben, & others, 2008; Ries & Eisner, 2007; Coburn et al., 2008; Wang, Emery, et al., 2007; Wang, Fuller, Emery, & Thomas, 2007; Choi, Fuller, Thomas, & Wang, 2008), many of them using the SEERStat software provided by SEER itself. Statistical studies using the SEER data include demographic and epidemiological studies of rare cancers (Yao et al., 2008), assessing susceptibility to secondary cancers that emerge after a primary diagnosis (Rusthoven et al., 2008), performing survival analysis (Ries & Eisner, 2007), studying the impact of a certain type of treatment on overall survival (Coburn et al., 2008), studying conditional survival (measuring prognosis of patients who have already survived a period of time after diagnosis) (Wang, Emery, et al., 2007; Wang, Fuller, et al., 2007; Choi et al., 2008), amongst many others. There also have been scattered applications of data mining using SEER data for breast cancer survival prediction (Lundin et al., 1999; Delen, Walker, & Kadam, 2005; Bellachia & Guven, 2006; Endo, Shibata, & Tanaka, 2008) and a few studying lung cancer survival (Chen et al., 2009; Fradkin, 2006; Agrawal, Misra, Narayanan, Polepeddi, & Choudhary, 2011), but to the best of our knowledge there is no association rule mining analysis on lung cancer data.

Recently, lung cancer data from SEER was used to construct predictive models for lung cancer survival after 6 months, 9 months, 1 year, 2 years, and 5 years of diagnosis using several data mining techniques (Agrawal et al., 2011). Carefully designed preprocessing steps were shown to enhance prediction accuracy, where a total of 63 attributes were used to build the models, and the results of the top five models were combined using an ensemble voting scheme, which also improved prediction accuracy. Further, a lung cancer outcome calculator was designed (available at http://info.eecs.northwestern.edu:8080/LungCancerOutcomeCalculator/) using a subset of 13 attributes, which was carefully chosen using attribute selection techniques, in an attempt to reduce the number of attributes while preserving the prediction accuracy. In this work, we use same dataset with 13 predictor attributes as used for the lung cancer outcome calculator in Agrawal et al. (2011) for association rule mining analysis. Further, a conditional survival dataset was created using the above-mentioned dataset, which included only those cases where the patient survived at least for 12 months. Association rule mining analysis was subsequently conducted on the resulting conditional survival dataset as well.

The rest of the paper is organized as follows: We give an overview of association rule mining and describes the HotSpot algorithm used in this work, followed by a description of the data and its attributes. Experiments and results are presented, and finally the conclusion and future work.

ASSOCIATION RULE MINING

Association rule mining is often stated as follows (R. Agrawal, Imielinski, & Swami, 1993): Let $I$ be a set of $n$ binary attributes called items. Let $T$ be a set of transactions. Each transaction in $T$ contains a subset of the items in $I$. A rule is defined as an implication of the form $X \rightarrow Y$
Word Sense Disambiguation in Biomedical Applications: A Machine Learning Approach
Torsten Schiemann, Ulf Leser and Jörg Hakenberg (2009). Information Retrieval in Biomedicine: Natural Language Processing for Knowledge Integration (pp. 142-161).
www.igi-global.com/chapter/word-sense-disambiguation-biomedical-applications/23059?camid=4v1a

Knowledge-Driven, Data-Assisted Integrative Pathway Analytics
www.igi-global.com/chapter/knowledge-driven-data-assisted-integrative/76062?camid=4v1a