Expert Systems with Applications 42 (2015) 6168–6176

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Data mining in lung cancer pathologic staging diagnosis: Correlation between clinical and pathology information

Haofan Yang, Yi-Ping Phoebe Chen *

Department of Computer Science and Computer Engineering, La Trobe University, Melbourne, Victoria 3086, Australia

ARTICLE INFO

Article history:
Available online 9 April 2015

Keywords:
Lung cancer
Cancer staging diagnosis
Data mining
Association rule mining

ABSTRACT

Lung cancer is one of the leading cancers for both genders all over the world. It is the most common cause of cancer death and almost reaches 20% of the total. The incidence of lung cancer has significantly increased from the early 19th century. In this manuscript we have discussed various data mining techniques that have been utilised for cancer diagnosis. The lung cancer pathologic staging is set based on the pathology report to describe the size and/or the extent of the original tumour and whether the cancer has spread (metastasis). Being aware of the lung cancer pathologic staging is important because it can be used to estimate a patient's prognosis and also can help physicians plan a suitable treatment. A sample of tissue from the patient's lung is required in order to complete the pathology report for the lung cancer pathologic staging diagnosis. In this procedure, a surgery biopsy is necessary but it may put the patient's health in jeopardy. Therefore, this study focuses on taking the clinical information which can be obtained without surgery to replace the pathology report. The data mining techniques are used to find the correlation between the clinical information and the pathology report in order to support lung cancer pathologic staging diagnosis.

1. Introduction

In recent decades, cancer has risen dramatically. Lung cancer is one of the leading cancers for both genders all over the world. It is the most common cause of cancer death and reaches 19.4% of the total (Bray, Ren, Masuyer, & Ferlay, 2013; International Agency for Research on Cancer, 2013; Siegel, Naishadham, & Jemal, 2013). The incidence of lung cancer has significantly increased from the early 19th century. About 90% of cases of lung cancer are related to exposure to tobacco smoke due to cigarettes and cigarette smoke contains over 70 cancer-causing chemicals (Duaso & Duncan, 2012). Smokers dramatically increased and so did lung cancer follow this innovation.

The most effective way to reduce lung cancer deaths is to detect it earlier. It requires an accurate and reliable diagnosis procedure that physicians can use it to distinguish and assign patients to either a benign (non-cancerous) or a malignant (cancerous). In this procedure, screening exams and pathological tests are applied. Screening exams, including physical exams, history of smoking, tests of sputum, chest X-rays, and spiral CT scans, etc., can provide physicians some initial information to judge whether it is a suspected lung cancer (Wender et al., 2013). After a lung cancer is suspected based on screening exams, a sample of tissue is taken from the patient's lung for the pathological tests. In general, lung biopsies are obtained by invasive surgeries, such as needle biopsy (Greif, Marmor, Schwarz, & Staroselsky, 1999), bronchoscopic biopsy techniques (Newton et al., 2011), or surgical removal of tissue. Pathologists will examine the patient's lung cells and physicians can rely on the results shown in the pathology report to confirm the diagnosis and determine the cancer pathologic staging. The lung cancer pathologic staging is assigned based on the results of pathology report. It describes the size and/or the extent of the original tumour and whether the cancer has spread (metastasis) (Sihoe & Yim, 2004; Vu-Quang et al., 2011).

Being aware of the lung cancer pathologic staging is important because it can be used to estimate a patient's prognosis and also can help physicians plan a suitable treatment (Edge & Compton, 2010). However, it often requires a lot of time and resources to get the pathology report in order to diagnose the lung cancer pathologic staging. If we can apply the clinical information to diagnose the cancer pathologic staging, many advantages are able to be acquired, for instance, saving medical resources, enhancing the efficiency of medical treatment, promoting early diagnosis and early prevention, and reducing the medical waste, etc. Therefore, this study focuses on taking the clinical information which can be obtained without the need of surgery to replace the pathology

* Corresponding author.
E-mail addresses: h16yang@students.latrobe.edu.au (H. Yang), phoebe.chen@latrobe.edu.au (Y.-P.P. Chen).

http://dx.doi.org/10.1016/j.eswa.2015.03.019
0957-4174/

report. The experimental data was obtained from the Cancer Genome Atlas (TCGA). In the experiment, the data mining techniques are used to find the correlation between the clinical information and the pathology report in order to support lung cancer pathologic staging diagnosis.

The goal of this paper is to demonstrate the feasibility of applying the clinical information to replace the pathology report especially in diagnosing the lung cancer pathologic staging. The remainder of this paper is organised as follows. We review various data mining applications in the area of cancer diagnosis in Section 2. We explain the framework of proposed cancer pathologic staging system in Section 3. We evaluate the performance of our design and analyse the results in Section 4. Finally, we conclude the paper in Section 5.

2. Related work

Data mining involves several techniques, such as decision tree, association rule mining, and neural networks, etc. Each of these techniques analyses data in different ways. In this section, we review several data mining applicants in the area of cancer diagnosis.

In Mukti & Ahmed (2013) the authors explored the applicability of AprioriTid (Chai, Yang, & Cheng, 2007) and decision tree (Ben-Haim & Tom-Tov, 2010) to discover significant frequent patterns. The goal was to using the significant patterns to develop a lung cancer prediction system. The prediction system was able to detect a person's predisposition for lung cancer. 400 cancer and non-cancer patients' data were collected and evaluated. With the proposed methodology, the authors showed that it is possible to find statistically significant associations from the gathered data set. However, the result evaluated through the proposed methodology does not show a high degree of statistical confidence.

Zubi and Saad (2011) used classification and generalised, neural network and association rule mining induction methods (Aggarwal & Zhai, 2012; Liao, Chu, & Hsiao, 2012) in order to classify the digital X-ray chest films into normal and abnormal categories. The main goal was to develop an image classification system for supporting physicians in decision making. They considered taking 300 X-ray chest films multimedia database as a training dataset. The basic pre-processing tasks were explained and the data mining induction methods used in the paper were indicated. They provided a detailed explanation of their proposed data mining methods. Nevertheless, the authors do not provide the evaluation results of their work.

Agrawal and Choudhary (2011) performed association rule mining analysis on lung cancer data from SEER. SEER is a source of information on cancer incidence and survival in the United States. The HotSpot algorithm (Williams, 1999) was used to identify the higher and lower target value of patient survival time. Since a large number of rules were generated by association rule mining, the authors used a 2-stage semi-manual procedure to remove redundant rules. In the first stage, the generated rules corresponding to the leaf nodes were retained and the rules to the non-leaf nodes were discarded in order to keep the best target value compared to all the nodes on its path. In the second stage, the domain expertise was participated to manually remove the redundant rules. Through the 2-stage procedure, more than 90% of the redundant rules were removed. The experiment results showed that the obtained rules were able to provide interesting insights into lung cancer survival. However, the authors only used lift to measure the generated rules, which cannot supply a strong convincingness.

In Ramachandran, Girija, Bhuvaneswari, Enathur, and Ponneri (2013) the authors applied classification and generalised mining method with WEKA tool to extract the cancer spread patterns and classify them according to the pre-defined attributes. The goal of this research was to provide valuable knowledge on the cancer spread pattern and its factors to enhance the cancer control program. Approximately 15,000 cancer cases were collected and classified into different age groups and different types of cancer. The experiment results showed that the cancer spread patterns were extract as expected, and some previously unknown knowledge was mined by the classification and generalised mining method. In conclusion their research provided an efficiently way and valuable observation to plan and enhance the cancer control program by apply the basic data mining technique.

3. Framework of proposed cancer pathologic staging system

In this section, we discuss the data mining techniques and methodology used in our research. The framework of proposed cancer staging system is shown in Fig. 1.

3.1. Data pre-processing and pathology dataset

Data pre-processing consumed the biggest portion of this study. We downloaded approximately 500 lung cancer cases, including 497 pathology reports and 458 clinical reports, from the Cancer Genome Atlas (TCGA). TCGA is a project co-managed by scientists and managers from the National Cancer Institute and the National Human Genome Research Institute, two of the 27 Institutes and Centres of the National Institutes of Health, US Department of Health and Human Services. From TCGA Data Portal, researchers can search, download, and analyse data sets generated by TCGA.

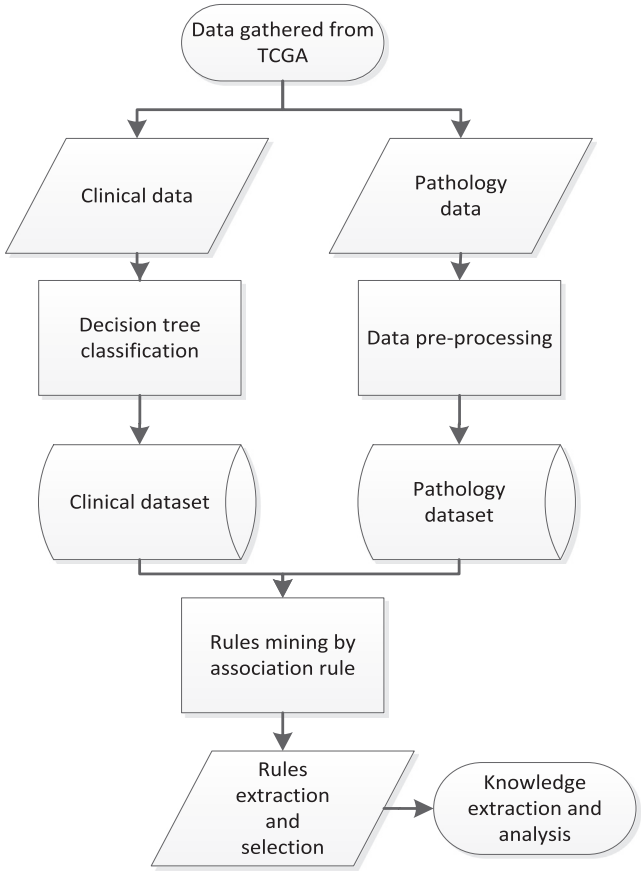

Fig. 1. Framework of proposed cancer staging system.