Analyzing complaint data for healthcare

Introduction

The provision of high-quality healthcare service is an increasingly difficult challenge, on the other hand the effective governance is increasingly recognized as pivotal to improvements in healthcare quality[ref]. The Australian Health Practitioner Regulation Agency (AHPRA) plays a crucial role in this aspect; it works with 14 National Health Practitioner Boards in implementing the National Registration and Accreditation Scheme. AHPRA also accept consumers or patients’ complaints, the AHPRA categorize the complaints based on the nature of the practice and then distribute them to the corresponding board. The board members will review the complaints in regular meeting to identify the issues and make suggestions; set new standards to improve the quality of service. Previous study[ref] shows the effectiveness of the boards is effected by insufficient resource and inadequate information received. Manually processing and analyzing this type of data is time consuming and inefficient. Moreover, even highly professional people could put subjective thinking into analyzes. Modern data mining technology already been employed in many area. For example market database system will analyze customers, categorize them in different groups and forecast their behavior. There is a huge potential to introduce data mining system into healthcare service. Not only reducing the labor required to process the massive amount of data, also tend to produce more accurate and objective analysis. Forecast the possibility of complaint against a medical practitioner based on historical data is a very valuable indicator for board members when they making the decision.

Although data mining techniques for marketing, manufacturing, security etc. is ripe for deployment, to apply data mining techniques to healthcare data is not a simple matter of switching content due to the complexity and sensitivity of the data. It is not a single task to develop one data mining package to cover all 14 boards, the variables and attributes affecting the result are different from board to board. There won’t be sufficient time and resource to develop algorithms to cover them all, however we are trying to find a generic data mining package to analyze the common variables among 14 boards, and develop an in depth analyze against one or two specialized boards, the similar method can be adapt to other boards with minimum amount of work in the future.

Related Work

A statistical study of healthcare complaint data[ref]

Unfortunately, data miming against healthcare complaint hasn’t attracted much attention, some studies has been done recently to discover the relationship between complaint data against particular or particular group ( gender, age, specialty ) of doctors with one decade worth of official complaint data within healthcare system. The study analyzed the distribution of complaints among practicing doctors. Then used recurrent-event survival analysis to identify characteristics of doctor at high risk of recurrent complaints, and to estimate each individual doctor’s risk of incurring future complaints. The study shows some very interesting and inspirational ideas, it coded specialty into 13 categories, for example: General Practice; Surgery; Psychiatry etc. The issues or complaints also categorized into 20 sections, for example Medication, Treatment in Clinical care; Consent, Information in Communication; Cost, Billing, Sexual Contact in Conduct etc. The result then been project to groups of doctors based on their gender, age and location (Urban/Rural).

Project Plan

Initially the project was designed to use official data from AHPRA. Due to the sensitivity of the data, the process of applying access of the data is time consuming, and the final decision of AHPRA is uncertain at stage. We will start the project with collecting unofficial data from some medical related forums. The data will be analyzed and a data mining package will be developed. At a later stage with access of official data, the similar process will be applied to the new data set. We will also try to identify a link between two different data sets.

Phase 1 – April, May, Jun 2014

‘Scrape’ data from medical related forums, use a third party tool or develop our own crawler to collecting web data. Some supervised learning algorithm will be used to train the crawler to collect correct data over the forums. Then we will start to analyze once we have handful of data.

It will also be useful if we can link the crawler and data mining package together to automate the analyze process.

Phase 2 – July, August 2014

Hopefully we got access to the official data from AHPRA at this stage, analyzing this data should be much easier compare to phase 1, the official data could be well formatted and will be less time spend on analyzing free-format text.

Phase 3 – September 2014

A study across official and unofficial data will be conducted

Phase 4 – October 2014

Finalizing the search paper and wrap up the thesis.

Tools will be used

Weka - is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

The main programming language will be used is Java, because it’s easy to run on different platforms, and rich library to choose from.

Evaluation

[References]