

Analyzing complaint data for healthcare

Bin Lu, S3360014

Supervisor: Xiuzhen Jenny Zhang, Daryl D'Souza, Amanda Kimpton

Program: Minor thesis

April 13, 2014

1 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 [4]. 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[3] 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[2][5], 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.[10][9][11][6] There wont 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.

2 Related Work

2.1 A statistical study of healthcare complaint data

[3] Unfortunately, data mining against healthcare complaint hasnt 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 doctors 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).

2.2 Data mining techniques in Customer Relationship Management (CRM)

The data mining techniques that used in other fields or industry can be used as a guide line for healthcare system, for example data mining in CRM. According to Swift (2001, p. 12), Parvatiyar and Sheth (2001, p.5) and Kracklauer, Mills, and Seifert (2004, p. 4), CRM consists of four dimensions:

1. Customer Identification;
2. Customer Attraction;
3. Customer Retention;
4. Customer Development.

These four dimensions can be seen as a closed cycle of a customer management system (Au & Chan, 2003; Kracklauer et al., 2004; Ling & Yen, 2001). They share the common goal of creating a deeper understanding of customers to maximize customer value to the organization in the long term. Data mining techniques, therefore, can help to accomplish such a goal

by extracting or detecting hidden customer characteristics and behaviours from large databases. The generative aspect of data mining consists of the building of a model from data (Carrier & Povel, 2003). Each data mining technique can perform one or more of the following types of data modelling:

1. Association;
2. Classification;
3. Clustering;
4. Forecasting;
5. Regression;
6. Sequence discovery;
7. Visualization.

The above seven models cover the generally mentioned data mining models in various articles (Ahmed, 2004; Carrier & Povel, 2003; Mitra, Pal, & Mitra, 2002; Shaw, Subramaniam, Tan, & Welge, 2001; Turban et al., 2007). There are numerous machine learning techniques available for each type of data mining model. Choices of data mining techniques should be based on the data characteristics and business requirements (Carrier & Povel, 2003). Here are some examples of some widely used data mining algorithms:

1. Association rule;
2. Decision tree;
3. Genetic algorithm;
4. Neural networks;
5. K-Nearest neighbour;
6. Linear/logistic regression.[7]

3 Project Plan

Initially the project was designed to use official data from AHPRA.[1] 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 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 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 and wrap up the thesis.

3.1 Tools will be used

Weka[8] - 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 its easy to run on different platforms, and rich resource of libraries to choose from.

4 Evaluation

As stated before, there are not much research been done for this area, we wont have some existing result to compare to. Basically the result will be evaluated in two ways. First, the result will be compared within same dataset but with different configuration. For example, change the algorithm used to analyze data or change the weight of attributes of input data, for example doctors age, gender. Second, compare result between different dataset (official data from AHPRA and unofficial data from forum) with similar configuration, this is the most important and interesting part of the project. By finding the relationship between two dataset, we are able to predict the data accuracy of data from public channel. If the data been approved reliable , it will expend the traditional healthcare feedback system in a cheaper and efficient way. As we are not having any data in hand, we are not absolutely

certain how the data will look like hence we are not able to give detailed evaluation plan at this stage.

References

- [1] Ahpra annual report 2011/12, 2012.
- [2] Deng Bin, Shao Peiji, and Zhao Dan. E-commerce reviews management system based on online customer reviews mining. In *Innovative Computing & Communication, 2010 Intl Conf on and Information Technology & Ocean Engineering, 2010 Asia-Pacific Conf on (CICC-ITOE)*, pages 374–377. IEEE, 2010.
- [3] Marie M Bismark, Matthew J Spittal, Lyle C Gurrin, Michael Ward, and David M Studdert. Identification of doctors at risk of recurrent complaints: a national study of healthcare complaints in australia. *BMJ quality & safety*, 22(7):532–540, 2013.
- [4] Marie M Bismark and David M Studdert. Governance of quality of care: a qualitative study of health service boards in victoria, australia. *BMJ quality & safety*, pages bmjqs–2013, 2013.
- [5] Anaya-Sanchez.H Garcia-Moya.L and Berlanga-Llavori.R. Retrieving product features and opinions from customer reviews. *Intelligent Systems*, 28(3):19–27, 2013.
- [6] Vivian Lin and Debra Gillick. Does workforce regulation have the intended effect? the case of chinese medicine practitioner registration. *Australian Health Review*, 35(4):455–461, 2011.
- [7] Eric WT Ngai, Li Xiu, and Dorothy CK Chau. Application of data mining techniques in customer relationship management: A literature review and classification. *Expert systems with applications*, 36(2):2592–2602, 2009.
- [8] Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.
- [9] Charlie C Xue, Wenyu Zhou, Anthony L Zhang, Kenneth Greenwood, Cliff Da Costa, Alex Radloff, Vivian Lin, and David F Story. Desired chinese medicine practitioner capabilities and professional development needs: a survey of registered practitioners in victoria, australia. *BMC health services research*, 8(1):27, 2008.
- [10] Charlie CL Xue, Anthony L Zhang, Vivian Lin, Ray Myers, Barbara Polus, and David F Story. Acupuncture, chiropractic and osteopathy

use in australia: a national population survey. *BMC Public Health*, 8(1):105, 2008.

- [11] Wenyu Zhou, Anthony Lin Zhang, Brian H May, Vivian K Lin, Anne-Louise Carlton, and Charlie Changli Xue. The victorian experience of transitional registration for chinese medicine practitioners and its implications for national registration. *Australian Health Review*, 36(1):61–67, 2012.