Analyzing complaint data for healthcare

Bin Lu, S3360014
Supervisor: Xiuzhen Jenny Zhang, Daryl D'Souza, Amanda Kimpton
Program: Minor thesis

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 affected by insufficient resource and inadequate information received. Manually processing and analyzing this type of data is time consuming and inefficient. Moreover, even highly trained professional 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.
The availability of official data from AHPRA is subject to approval, so we planned a backup source for the data. The stories or opinion from www.patientopinion.org.au. Patient Opinion Australia(POA) was established in 2012 and, similar to its UK counterpart, is registered as an independent not-for-profit charitable institution. As the data from AHPRA remains uncertain, the later one will be used as primary data source for this project. The opinion data will be different from formal complaint data in many ways. For example, the opinions wont be clearly categorized by specialty, they could be either positive or negative, people who make opinion can be patients, carers or relatives. Since all the opinions are available to public through web pages, we will develop a simple crawler to retrieve all the opinions from website. Then apply some machine learning techniques to classify the free formatted data. The classifier should be able to predict an opinion is positive or negative, finally statistical analyze to find distribution of the opinion across specialty, location and autho role. An unsupervised machine learning technique will be used for clustering the whole data set. By clustering the opinion, we hope to discover some hidden patterns within data set.
There is a potential problem for the POA data. A single story could contain multiple aspects (treatment, doctor, hospital) with different sentiment. To solve this problem, an application called aspect-based summarization[5] will be used.The set of aspects will be extracted from opinion and the sentiment of each aspect will be analyzed.
The search question or the goal of the project will be: 1) Is the classification of opinion data sufficient to identify the distribution of the positive/negative feedbacks against specialty, location, service or even hospital. 2) Is the classification of opinion data sufficient to rank the specialty and hospital. 3) Discover the similarity between opinions by clustering. 4)Is aspect-based summarization more suitable to opinion websites, blogs Internet forums and social networks.

2 Related Work

2.1 Aspect-Based Summarization

One of the most relevant applications of opinion mining and sentiment analysis is aspect-based summarization. Broadly speaking, given a collection of opinion posts, this task is aimed at obtaining relevant aspects (such as product features), along with associated sentiment information expressed by customers (usually an opinion word and/or a polarity score). Aspect-based summarization is usually composed of three main tasks: aspect identification, sentiment classification, and aspect rating. Aspect identification is fo-

cused on extracting the set of aspects or product features from the source collection. The word aspect is intended to represent the opinion or sentiment targets, which are also referred to as product features when the collection of poststypically, customer reviewsis about products or services. For example, given the sentence, The bed was comfortable in a review about a hotel room, the aspect being referred to is bed and the opinion is positively expressed by means of the opinion word comfortable. The sentiment classification task consists of determining the opinions about the aspects and/or their polarities, whereas aspect rating leverages the relevance of aspects and their opinions to properly present them to users.[5]A similar approach will be used in analyzing patient opinions. Each opinion can be classified as positive/negative by sentiment classification. In advance mode, aspects will be identified from each opinion. The feedback will be extended to each aspect. For example, if both practitioner and a hospital mentioned in an opinion, but practitioner get a positive feedback but hospital is negative. The opinion level rating isnt accurate enough to describe all the aspects.

2.2 A statistical study of healthcare complaint data

Unfortunately, data miming 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.[3] 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).Figure 1 plots the cumulative distribution of complaints among doctors in six jurisdictions over a decade. (New South Wales data was not included in these plots because the complaints window there spanned only 5 years.) The curve on the left side of the figure shows the distribution of complaints among doctors who experienced one or more complaints in the decade. Fifteen percent of doctors named in complaints accounted for 49% of all complaints, and 4% accounted for a quarter of all complaints. The curve on the right side of the figure shows the distribution of complaints across the full population of practicing doctors, not just those who experienced complaints. Three percent of all doctors accounted for 49% of all complaints, and 1% accounted for a quarter of all complaints. The study also concluded it is feasible to predict which doctors are at high risk of incurring more complaints in the near future.[3]

2.3 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.[11] According to Swift (2001, p. 12)[14], Parvatiyar and Sheth (2001, p.5)[12] and Kracklauer, Mills, and Seifert (2004, p. 4)[8], CRM consists of four dimensions:

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

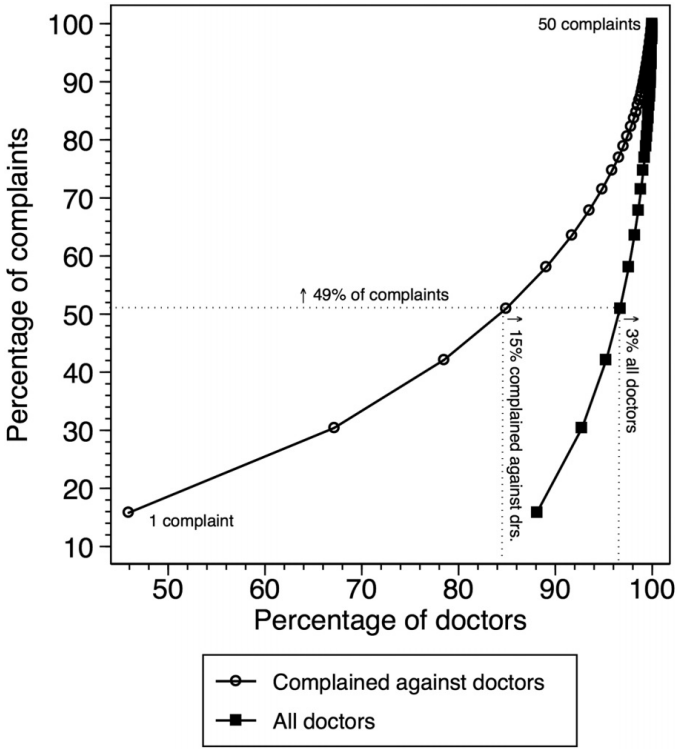

Figure 1. Cumulative distribution of complaints and doctors named in complaints.

These four dimensions can be seen as a closed cycle of a customer management system (Au & Chan, 2003;[2] Kracklauer et al., 2004;[8] Ling & Yen, 2001[9]). 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)[6]. 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;[1] Carrier & Povel, 2003;[6] Mitra, Pal, & Mitra, 2002;[10] Shaw, Subramaniam, Tan, & Welge, 2001;[13] Turban et al., 2007[15]). 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[6]. 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.

3 Project Plan

Due to the uncertainty of the availability of the complaint data from AHPRA, we wont include it in the project plan at this stage. The plan will be based on opinion/story from POA.

- Phase 1 May, Jun 2014
  Scrape data POA. The data will be stored as XML file with sufficient tags to retain original information. For example: title; specialty; location. Develop a program to efficiently mark each opinion/story as positive or negative and store the extra information on the disk. Develop a program to parse the stored data and output a valid WEKA file.

- Phase 2 July 2014
  WEKA will be used to analysing the data from phase 1. Two learning methods will be used to develop the classifier: C4.5 Decision tree and Basic Nearest-neighbor. Select different features as training data to develop a variety of classifiers. Experiment the clustering with Expectation Maximization(EM) method.

- Phase 3 August 2014
  Develop a program to extract aspect and sentiment from each opinion and calculate the distribution of the opinion across aspects.

- Phase 4 September 2014
  Evaluation of the classifiers, aspect-based summarization model.

- Phase 5 October 2014
  Finalizing the search and wrap up the thesis.

3.1 Tools will be used

Weka[16] - 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. We will use publicly available English opinion lexicon originally proposed by M Hu[7](www.cs.uic.edu/ liub/FBS/opinion-lexicon-English.rar). The set contains 6800 opinion words.

4 Preliminary Work

The first phase of the project will be focusing on collecting and post processing data. The data structure of POA is analyzed. The entry page will be www.patientopioion.org.au/opinions, the total number of page could be located from section ¡div class=pagination clearfix¿, the ¡li class=last¿ item contains a link to the very last page. The page number can be extracted from text string. Each page will be visited to retrieve individual

article IDs. Page URL looks like:
www.patientopinion.org.au/opinions?page=[PageNumber].
Article URL looks like:
www.patientopinion.org.au/opinions/[ID]. All the HTML text will be stripped. The useful information will be retained and formatted before store on the disk for future use. The information will be stored include title; story body; author role; service location; submit time and a list of related tags. By today, 629 stories from website had been collected.

5 Evaluation

There wont be existing data from other studies for us to compare to. To evaluate our project, we consider to use 3-fold and 10-fold cross-validation. The data will be divided randomly into 3 parts, each held out in turn and the learning scheme trained on the remaining two-thirds; thus the learning procedure is executed a total of 3 times on different training sets. We will do a quick calculation on the error rate. Then we will perform the 10-fold validation, which is similar to 3-fold but the data set is divided into 10 parts, thus the learning procedure will be executed a total of 10 times. The later process will give a smoother average of error rate.

Patient Opinion established Australian site in 2012, its ancestor patientopinion.org.uk was founded in 2005. It has collected more than 80,000 opinions/stories over almost a decade. Both websites use same data structure, so the data mining technique developed for Australian should also work for UK. The distribution of the data for UK will also be analyzed and compared to our first result set of Australian.

References

[1] Syed Riaz Ahmed. Applications of data mining in retail business. In *Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. International Conference on*, volume 2, pages 455–459. IEEE, 2004.

[2] W-H Au and Keith CC Chan. Mining fuzzy association rules in a bank-account database. *Fuzzy Systems, IEEE Transactions on*, 11(2):238–248, 2003.

[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] C Giraud-Carrier and Olivier Povel. Characterising data mining software. *Intelligent Data Analysis*, 7(3):181–192, 2003.

[7] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177. ACM, 2004.

[8] Alexander H Kracklauer, D Quinn Mills, and Dirk Seifert. Customer management as the origin of collaborative customer relationship management. In *Collaborative Customer Relationship Management*, pages 3–6. Springer, 2004.

[9] Raymond Ling and David C Yen. Customer relationship management: An analysis framework and implementation strategies. *Journal of Computer Information Systems*, 41(3):82–97, 2001.

[10] Sushmita Mitra, Sankar K Pal, and Pabitra Mitra. Data mining in soft computing framework: A survey. *IEEE transactions on neural networks*, 13(1):3–14, 2002.

[11] 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.

[12] Atul Parvatiyar and Jagdish N Sheth. Customer relationship management: Emerging practice, process, and discipline. *Journal of Economic & Social Research*, 3(2), 2001.

[13] Michael J Shaw, Chandrasekar Subramaniam, Gek Woo Tan, and Michael E Welge. Knowledge management and data mining for marketing. *Decision support systems*, 31(1):127–137, 2001.

[14] Ronald S Swift. *Accelerating customer relationships: Using CRM and relationship technologies*. Prentice Hall Professional, 2001.

[15] Efraim Turban, Ramesh Sharda, Dursun Delen, and Turban Efraim. *Decision support and business intelligence systems*. Pearson Education India, 2007.

[16] Ian H Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.