Knowledge Discovery and Data Mining in Health Care

Osagioduwa Edo-Osagie

1 Introduction

The term *knowledge discovery* refers to the process of obtaining useful information from analyzing large sets of data. It employs data mining, a field in computer science that involves the computational process of discovering patterns in large data sets, to do this (Wikipedia (2016)). Data mining is an intersection of a number of other computer science fields as it applies ideas and methods from machine learning, pattern recognition, databases, statistics and artificial intelligence. It is a multidisciplinary subject and is widely used in diverse areas such as financial data analysis, the retail industry, the telecommunications industry, the insurance industry, biological data analysis, intrusion detection etc. Because of its ability to aid in the analysis of customer and sales data for trends in order to maximize profit, data mining is well known as a wonderful financial tool. However, although this is lesser known, it is a also a useful tool in the fields of medicine and health care. This paper reviews the current application of knowledge discovery techniques in the area of medicine/health Care, the methods and techniques employed and the uniqueness of medical data as well as the challenges it poses.

2 What is Data Mining in Health Care

Data mining holds great potential for the health care industry in that it could enable health care systems to systematically use and analyze data to identify inefficiencies and best practices that improve care and reduce costs. Some experts believe the opportunities to improve care and reduce costs concurrently could apply to as much as 30% of overall health care spending (Crockett (2014)). It could also be applied to the medicinal and biological aspects of health care in order to predict the outbreak of diseases or monitor drug performance and side effects. This could be a win/win overall, but due to the complexity of medicine and health care, coupled with a slower rate of technology adoption, the health care industry is slightly behind others in implementing effective data mining and analytic strategies.

Moving forward, it is important to note that not all analyses of large amounts of data constitute data mining. Crockett and Eliason (2014) categorize data analysis as follows:

1. Descriptive analytics — Describing what has happened
2. Predictive analytics — Predicting what will happen
3. Prescriptive analytics — Determining what to do about it

While the knowledge discovery process involves all three categories at one point or another, data mining itself falls under the first and second categories - *descriptive analytics* and *predictive analytics*.

One example of a pragmatic application of data mining in health care today is that of monitoring and predicting fee-for-service volumes. Crockett and Eliason (2014) report that they worked

with a health care organization whose revenue is mostly generated by providing fee-for-service contracts via referrals to its top-rated facilities. Because of this, the organization wanted to make sure that these fee-for-service contracts remained in place and brought in a steady stream of income. To do this, they decided to monitor the process by building a data warehouse and an analytics system. The data warehouse aggregated multiple data sets, such as payer, financial, and cost data, and then displayed a dashboard of information such as referral patterns for each payer and the margins associated with those payers. With this in place, the organization was able to see if referrals from a certain source were slowing down. They could then react quickly through outreach, advertising, and other methods.

3 Machine Learning and Data Mining

In 1959, Arthur Samuel defined machine learning as "a field of study that gives computers the ability to learn without being explicitly programmed" (Samuel (1959)). Machine learning explores the study and construction of algorithms that can learn from and make predictions on data. As such, it is the foundation for the idea of descriptive and predictive analysis through the construction of models in data mining. It can roughly be categorized into two groups - ***Supervised Learning*** and ***Unsupervised Learning*** (Simon (2013)).

In supervised learning, data usually consists of examples (records of given attribute values) which are labeled by the class to which they belong (Kotsiantis et al. (2006)). The task here, is to find a model - known as a classifier - that will enable a newly encountered instance to be identified as one of the classes to which the data might belong to. On the other hand, in unsupervised learning, the data is not labeled and the task is to perform the labeling, uncovering some hidden structure from the unlabeled data.

Predictive analytics (also known as predictive data mining) falls under supervised learning as it is used to induce models or theories from class-labeled data to be used for future predictions. Examples of machine learning methods used for predictive data mining are *Decision Trees* and *Rule Set Induction*. Descriptive analytics (also known as descriptive data mining) is a form of unsupervised learning as it is used to induce interesting patterns from unlabeled data. Such patterns are useful in exploratory data analysis. Examples of machine learning methods used for descriptive data mining are *Clustering* and *Association Rule Learning* (Fürnkranz et al. (2012)).

4 Uniqueness of Medical Data and the Challenges It Presents

Data mining in health care is unique. Medicine is primarily a patient-care activity, and only secondarily a research resource. The whole idea and purpose for collecting medical data is to benefit the individual patient rather than for analytical purposes. This poses a problem when faced with the task of data mining. Below are some of the major points of uniqueness of medical data as proposed by Cios and Moore (2002):

- Heterogeneity of medical data
- Poor mathematical characterization of medical data
- Privacy and security of human data
- Special status of medicine

4.1 Heterogeneity of medical data

Medical data is highly voluminous and heterogeneous. This means that the data is from a number of different sources, recorded by different individuals in different forms and formats. The clinician's interpretation of clinical information is recorded in unstructured free-text language. This makes it difficult to standardize its mining process. Not only do different names describe the same disease, but the task is made even more daunting by using different grammatical structures to describe relationships between medical elements and concepts. It is thought that computer translation may hold part of the solution for tackling this problem with medical data

4.2 Poor mathematical characterization of medical data

Another unique feature of medical data is that it cannot really be represented mathematically, as compared to many areas of the physical sciences. In traditional data mining problems, the data collected can be put into formulas, equations, and models that reasonably reflect their nature and relationships. Medical data on the other hand, consists of word descriptions (and maybe images) with only a few formal constraints on the vocabulary (Zadeh (1979)).

4.3 Privacy and security of human data

The clinician-patient relationship exists such that the patient is completely open and honest with the clinician, with the expectation that any information generated is private and will never be made public. However, when being processed for data mining, such data might need to be sent to other parties and possibly be transferred over the internet. How secure is this? The data could be anonymized but what happens when an important, possibly life threatening diagnosis is made. If the data is completely anonymous, there will be no way to inform the patient on any new discoveries regarding their health.

4.4 Special status of medicine

The field of medicine is special in that, unlike other fields, it often involves life-or-death situations. Also, when medical data is published, it is expected that the researchers will maintain the dignity of the individual patient, and that the results will be used for socially beneficial purposes (Sweeney (2001)). It has been suggested that scientific truths are fundamentally amoral; they can be used for good or evil (Cios et al. (2012)). Because of this, there are certain tests that may not be performed, certain questions that may not be asked, and certain conclusions that may not be drawn, because of medicine's special status.

5 Applications of Data Mining in Medicine

A lot of research is being carried out to apply data mining and knowledge discovery techniques to health care. Two rather interesting applications are in ***syndromic surveillance*** as well as ***drug surveillance***. These applications are special in that they incorporate user web data and social media data instead of relying on just organizational data. This section reviews recent attempts at the applications of data mining to twitter data for the purposes of monitoring disease outbreaks and drug safety respectively.

5.1 Content Analysis of Tweets during the 2009 H1N1 Outbreak

Chew and Eysenbach (2010) carried out a study to show that "*tweets*" - user posts to the Twitter website - can be used for real-time content analysis and knowledge translation research into

allowing health authorities to respond to public concerns. The study monitored the use of the medical term "H1N1" versus the common vernacular "swine flu" over time and conducted a content analysis of tweets to validate Twitter as a real-time public attention trend-tracking tool.

An open-source infoveillance system, *Infovigil* (Eysenbach (2009)), was developed which continuously gathered and mined textual information from Twitter using the Twitter Application Programming Interface (API). To analyse the tweets, a tri-axial coding scheme was employed to reflect the tweet's content, how it was expressed and the type of link posted if any. Tweets were categorized as ***Humour***, ***Concern***, ***Frustration***, ***Downplayed Risk***, ***Relief***, ***Misinformation***, ***Personal Experience***, ***Personal Opinions***, ***Resources*** and ***Questions***. The chi-square test for trend was used to determine if the proportion of content, qualifiers, or links tweeted changed linearly over the analysis timeframe. Public attention was also measured as sharp increases in the volume of tweets related to H1N1 were observed to coincide with major H1N1 news events.

H1N1 related tweets on Twitter were primarily used to broadcast information from credible sources to the public, but were also a rich source of opinions and experiences. These tweets provide an opportunity for near real-time content and sentiment analysis, allowing health authorities to become aware of and respond to real or perceived concerns raised by the public.

5.2 Digital Detection of Pharmaceutical Side Effects

Sometimes drugs and other pharmaceutical products lead to unexpected negative side effects. In medicine and pharmacy, these side effects are known as *adverse events*. Traditional systems for reporting AEs have been slow in adapting to online reports from patients, and instead rely on clinicians and drug safety groups to verify each potential event. Freifeld et al. (2014) carried out a study to evaluate the level of concordance between Twitter posts mentioning AE-like reactions and spontaneous reports received by a regulatory agency.

Tweets mentioning medical products were collected (using the Twitter API) and manual and semi-automated techniques were applied to them in order to identify posts that pointed to AEs. A tree-based dictionary-matching algorithm was used to extract both product and symptom mentions from the tweets. Further natural language processing was applied to identify the AE in each post, converting common internet vernacular to a standardized regulatory dictionary, namely Medical Dictionary for Regulatory Activities (MedDRA). AEs were identified from public FDA Adverse Event Reporting System (FAERS) data for the products of interest using exact name matching.

In the end, AEs identified in Twitter appeared to have a similar profile to spontaneous reports received by the FDA. Furthermore, it was observed that patients tend to passively report AEs in social media, as evidenced by hashtags and mentions such as #accutaneprobz and @EssureProblems.

The use of web and social media for health surveillance is a relatively recent phenomena. Both the tools themselves and the methodologies for exploiting them are evolving over time. Health surveillance has the potential to improve in sensitivity and specificity advances in data mining and knowledge discovery.

6 The Future of Data Mining

It is thought that in the future, text mining will be made more powerful and image mining and graph mining will be more established and commonplace (Coenen (2011)). Image mining is concerned with the representation of two-dimensional images such that mining techniques may be applied. Current research in image mining continues to be focused on how best to represent images so that data mining techniques can be applied.

Knowledge discovery is closely linked with data mining. Currently, the relationship between data mining and knowledge discovery has been used in retail marketing (e.g. identifying customers' buying patterns, market basket analysis), banking (e.g. detecting credit card fraud, identifying loyal customers, etc.) and insurance (e.g. claim analysis, predicting customers who buy new policies) (Le (2015)). In the future, organizations will begin harnessing more practical methods of extracting knowledge using data mining techniques. This could result in interesting things such as dynamic web technology fine tuned to each individual.

However, I believe that before some of these predictions can come to pass, advancements in other areas of computer science must be made. An example of this is the issue of image mining. Computer Vision and Image Analysis will need to develop a concise representation for images that can be mined efficiently before new image mining methods can be developed.

7 Conclusion

Data mining has come to prominence over the last twenty years or so, as a discipline in its own right which offers benefits with respect to many other disciplines and domains, both commercial and academic. Broadly, data mining can be viewed as an application domain, as opposed to a technology. Today, data mining in health care takes place primarily in an academic setting rather than it being applied in practice. However, as can be seen from the outcomes of many research projects carried out in the area, data mining in health care has some very powerful benefits in the real world. It not only puts forward novel approaches to certain medical processes (as in syndromic surveiilance), but can also be used to augment existing ones as in (pharmaceutical adverse event reporting). But as far as we have come, we are still interested in techniques to mine ever larger data sets and an ever greater variety of data. Therefore, while the discipline is in a good place, it still requires a significant amount of research input.

References

Chew, C. and Eysenbach, G. (2010). Pandemics in the age of twitter: content analysis of tweets during the 2009 h1n1 outbreak. *PloS one*, 5(11):e14118.

Cios, K. J. and Moore, G. W. (2002). Uniqueness of medical data mining. *Artificial intelligence in medicine*, 26(1):1–24.

Cios, K. J., Pedrycz, W., and Swiniarski, R. W. (2012). *Data mining methods for knowledge discovery*, volume 458. Springer Science & Business Media.

Coenen, F. (2011). Data mining: past, present and future. *The Knowledge Engineering Review*, 26(01):25–29.

Crockett, Johnson, E. (2014). What is data mining in healthcare?

Eysenbach, G. (2009). Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. *Journal of medical Internet research*, 11(1).

Freifeld, C. C., Brownstein, J. S., Menone, C. M., Bao, W., Filice, R., Kass-Hout, T., and Dasgupta, N. (2014). Digital drug safety surveillance: monitoring pharmaceutical products in twitter. *Drug safety*, 37(5):343–350.

Fürnkranz, J., Gamberger, D., and Lavrač, N. (2012). Machine learning and data mining. In *Foundations of Rule Learning*, pages 1–17. Springer.

Kotsiantis, S. B., Zaharakis, I. D., and Pintelas, P. E. (2006). Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26(3):159–190.

Le, T. (2015). What does the future hold for data mining? `https://www.linkedin.com/pulse/what-does-future-hold-data-mining-thiensi-le`.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229.

Simon, P. (2013). *Too Big to Ignore: The Business Case for Big Data*. John Wiley & Sons.

Sweeney, L. (2001). Computational disclosure control. *A Primer on Data Privacy Protection*.

Wikipedia (2016). Data mining. `https://en.wikipedia.org/wiki/Data_mining`.

Zadeh, L. A. (1979). Fuzzy sets and information granularity. *Advances in fuzzy set theory and applications*, 11:3–18.