Data Mining in Healthcare:

A Literature Survey of Current Applications and Issues

Ruben D. Canlas, Jr.

Wilson A., Thabane, L., Holbrook A (2003). Application of data mining techniques in pharmacovigilance. *British Journal of Clinical Pharmacology*, *2*(57), 127-134.

Witten, I. H. and Frank, E. (2005). *Data mining: practical machine learning tools and techniques: Morgan Kaufmann series in data management systems*. Boston, MA: Morgan Kaufman.

References

The Occasional Paper Series (OPS) is a regular publication of the Ateneo Graduate School of Business (AGSB) intended for the purpose of disseminating the views of its faculty that are considered to be of value to the discipline, practice and teaching of management and entrepreneurship. The OPS includes papers and analysis developed as part of a research project, think pieces, and articles written for national and international conferences. The OPS provides a platform for faculty to contribute to the debate on current management issues that could lead to collaborative research, management innovation and improvements in business education.

The views expressed in the OPS are solely those of the author(s) and do not necessarily reflect the views of AGSB or the Ateneo de Manila University.

Published by the Ateneo de Manila University
Graduate School Business
Ateneo Professional Schools Building
Rockwell Drive, Rockwell Center, City of Makati Philippines
Tel.: (632) 899-7691 to 96 or (632)729 2001-2003
Fax: (632) 899-5548
Website: http://gsb.ateneo.edu

Limited copies may be requested from the AGSB Research Unit.
Telefax: (632) 898-5007
Email: submit@agsbresearch.org

Kou, Y., Lu, C.-T., Sirwongwattana, S., and Huang, Y.-P. (2004). *Survey of fraud detection techniques*. Proceedings of the 2004 IEEE International Conference on Networking, Sensing and Control, 749-754.

Nightingale, F. (1858). Notes on Matters Affecting the Health, Efficiency and Hospital Administration of the British Army.

Shillabeer, A. (2009, 29 July). *Lecture on Data Mining in the Health Care Industry*, Mellon University, AUS.

Shillabeer, A. and Roddick, J. (2007). Establishing a lineage for medical knowledge discovery. *ACM International Conference Proceeding Series 70* (311), 29-37.

Tandoc, E.S (14 October 2006). DOH order probe after Rizal hospital tragedy - Sanitation regulations stressed. *Philippine Daily Inquirer*, p. A19.

Thangavel, K., Jaganathan, P. P. and Easmi, P.O. Data mining approach to cervical cancer patients analysis using clustering technique. *Asian Journal of Information Technology 4*(5), 413-417.

Tufte, E. (1997). *Visual Explanations: Images and Quantities, Evidence and Narrative*. Connecticut: Graphics Press.

Wong, W.K., Moore, A., Cooper, G., and Wagner, M. (2005). What's strange about recent events (WSARE): An algorithm for the early Detection of disease outbreaks. *Journal of Machine Learning Research 6*, 1961-1998.

Data Mining in Healthcare:

A Literature Survey of Current Applications and Issues

Ruben D. Canlas Jr.

Introduction

The successful application of data mining in highly visible fields like e-business, marketing and retail have led to its application in knowledge discovery in databases (KDD) in other industries and sectors.. Healthcare management is one sector that is just discovering data mining.

This paper provides a literature survey of current techniques in KDD using data mining techniques that are in use today in medical research and public health. It also discusses some critical issues and challenges associated with the application of data mining in the health profession and medical practice in general.

More specifically, the paper aims to: (i) enumerate current uses and highlight the importance of data mining in medicine and public health management; (ii) find data mining techniques used in other fields that could also be applied in the health sector; (iii) identify issues and challenges in data mining as applied to the medical practice; and (iv) outline some recommendations for discovering knowledge in electronic databases through data mining.

The literature survey covered journals and publications in the fields of medicine, computer science and engineering. The research focused on more recent publications, with 2000 as the cut off year.

References

Audain, C. (2007). *Florence Nightingale*. Retrieved July 30, 2009 from http://www.scottlan.edu/lriddle/women/nitegale.htm

Ayres, I (2008). *Super Crunchers*. New York: Bantam Books.

Bailey-Kellog, C., Ramakrishnan, N., and Marathe, M. (2006). Spatial data mining to support pandemic preparedness. *SIGKDD Explorations 8*(1), 80-82.

Cao, X., Maloney, K.B., and Brusic, V. (2008). Data mining of cancer vaccine trials: A bird's-eye view. *Immunome Research 4*(7). Available online at http://immunome research.com/content/pdf/1745-7580-4-7.pdf

Cheng, T.H., Wei, C.P., and Tseng, V.S. (2006, June). *Feature selection for medical data mining: Comparisons of expert judgment and automatic approaches*. 19th IEEE International Conference on Computer-Based Medical Systems (CBMS '06), USA.

Health Grades, Inc. (2007). *The Fourth Annual Health Grades Patient Safety in American Hospitals Study*. Available online at http://www.eurekalert.org/images/release_graphics/pdf/PatientSafetyInAmericanHospitalsStudy2007Embargoed.pdf

cholera in certain areas that are supposed to have eradicated these diseases. PhilHealth could also apply data mining to find and stop anomalous insurance claims.

Before embarking on data mining, however, an organization must formulate clear policies on the privacy and security of patient records. It must enforce this policy with its partner-stakeholders and its branches and agencies.

Public health concerns like rapid pandemic outbreaks, the need to detect the onset of disease in a non-invasive, painless way, and the need to be more responsive to its customers - all these add up to an increasing need for health organizations to integrate data and apply data mining to analyze these data sets.

The practice of using concrete data and evidence to support medical decisions (also known as evidence-based medicine or EBM) has existed for centuries. John Snow, considered to be the father of modern epidemiology, used maps with early forms of bar graphs in 1854 to discover the source of cholera and proved that it was transmitted through water supply (Tufte, 1997).

Figure 1. Early Map Depicting the Source of Cholera

Source: Tufte, 1997

Snow counted the number of deaths and plotted the victim's addresses on the map as black bars. He discovered that most of the deaths clustered towards a specific water pump in London (center of the red circle in the map in Figure 1).

Florence Nightingale invented polar-area diagrams in 1855 (Figure 2) to show that many army deaths could be traced to unsanitary clinical practices and were, therefore, preventable. She used the diagrams to convince policy-makers to implement reforms that eventually reduced the number of deaths (Audain, 2007).

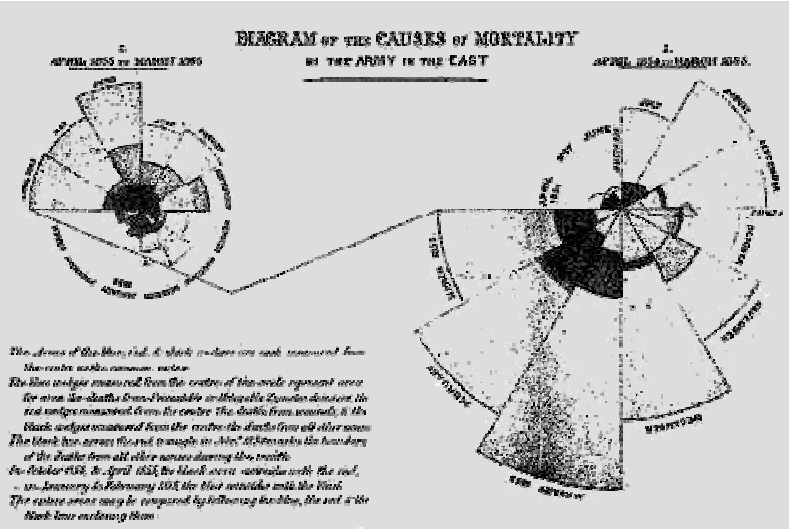

Figure 2. Cause of Mortality in the Army in the East, April 1854 to March 1855

Conclusion and Recommendations

The survey of data mining applications in medicine and public health provided only an overview of current practices and challenges. It highlighted a number of key data mining applications that could be useful in the Philippines. Health care organizations and agencies could look into these possible applications to extract knowledge from their own database systems in order to improve management and policymaking in the fields of medicine and public health.

For example, DOH could coordinate with government-operated hospitals, PhilHealth and the National Statistics Office to collate and analyze public health indicators. They could apply data mining techniques to find trends in disease outbreaks or deaths (e.g., infant mortality), per region and per hospital.

DOH could uncover hidden patterns in deaths or disease that could lead to better health policies like better vaccination planning, identification of disease vectors like malaria, prevention of hospital errors and the inexplicable but sporadic outbreaks of flu and

Snow and Nightingale were able to personally collect, sift through and analyze the mortality data during their times because the volume of information was manageable. Today, the size of the population, the amount of electronic data gathered, along with globalization and the speed of disease outbreaks make it almost impossible to accomplish what the pioneers did.

This is where data mining becomes useful to healthcare. It has been slowly but increasingly applied to tackle various problems of knowledge discovery in the health sector.

Data mining and its application to medicine and public health is a relatively young field of study. In 2003, Wilson, and associates began to scan cases where KDD and data mining techniques were applied in health databases. They found confusion in the field regarding what constituted data mining. Some authors refer to data mining as the process of acquiring information, whereas others refer to data mining as utilization of statistical techniques within the knowledge discovery process. (Wilson, Thabane and Holbrook, 2003)

Data mining and its application to medicine and public health is a relatively young field of study.

Because of misconceptions still ongoing in the medical community, “the term” data mining must first be defined. The generally accepted definition of data mining today is the set of procedures and techniques for discovering and describing patterns and trends in data (Witten and Frank, 2005). This definition shall be used throughout the paper.

The Importance and Uses of Data Mining in Medicine and Public Health

Despite the differences in approaches, the health sector has more need for data mining today. There are several arguments that could be advanced to support the use of data mining in the health sector, covering not just the concerns of public health but also of the private health sector (which in fact, as shown later, are also stakeholders in public health).

Data overload. There is a wealth of knowledge to be gained from computerized health records. Yet the overwhelming bulk of data stored in these databases makes it extremely difficult, if not impossible, for humans to sift through it and discover knowledge (Cheng, Wei and Tseng, 2006).

Even if data mining results are credible, convincing the health practitioners to change their habits based on evidence may be a bigger problem. Ayres (2008) reported a couple of cases where hospital doctors refused to change hospital policy even when confronted with evidence. In one case, it was found that doctors coming out of autopsy without washing hands led to a high probability of deaths in the patients they treated after the autopsy. Presented with this evidence, doctors still refused to change their habits until only much later.

Shillabeer (2009) also reported that most doctors (at least in Australia) prefer to listen to a respected opinion leader in the medical profession, rather than to the result of data mining. Shillabeer’s observation can be validated by us, since we have worked with doctors in a medical school in our capacity as a management consultant.

Privacy of records and ethical use of patient information is also one big obstacle for data mining in healthcare.

Privacy of records and ethical use of patient information is also one big obstacle for data mining in healthcare. For data mining to be more accurate, it needs a sizeable amount of real records. Healthcare records are private information, and yet, using these private records may help stop deadly diseases.

For example, anthrax and influenza share the same symptoms of respiratory problems. Lowering the threshold signal in a data mining experiment may either raise an anthrax alarm when there is only a flu outbreak. The converse is even more fatal: a perceived flu outbreak turns out to be an anthrax epidemic (Wong, Moore, Cooper and Wagner, 2005). It is no coincidence that we found, in most of the data mining papers on disease and treatment, that the conclusions were almost-always vague and cautious. Many would report encouraging results but recommend further study. This failure to be conclusive indicates the current lack of credibility of data mining in these particular niches of healthcare.

In most of the data mining papers on disease and treatment, many would report encouraging results but recommend further study.

The confusion about the definition of data mining also complicates the issue. For example, we found a couple of papers with the keywords “data mining” in their titles but turned out to be the simple use of graphs. Shillabeer (2009) said that this misunderstanding is prevalent in the relatively young existence of data mining in healthcare.

Some experts believe that medical breakthroughs have slowed down, attributing this to the prohibitive scale and complexity of present-day medical information.

In fact, some experts believe that medical breakthroughs have slowed down, attributing this to the prohibitive scale and complexity of present-day medical information. Computers and data mining are best-suited for this purpose (Shillabeer and Roddick, 2007).

Evidence-based medicine and prevention of hospital management errors. When medical institutions apply data mining on their existing data, they can discover new, useful and potentially life-saving knowledge that otherwise would have remained inert in their databases. For instance, an ongoing study on hospitals and safety found that about 87% of hospital deaths in the United States could have been prevented, had hospital staff (including doctors) been more careful in avoiding errors (Health Grades Inc., 2007). By mining hospital records, such safety issues could be flagged and addressed by hospital management and government regulators.

Policy-making in public health. Lavrac et al (2007) combined GIS and data mining using, among others, Weka with J48 (free, open source, Java-based data mining tools), to analyze similarities between community health centers in Slovenia. Using data mining, they were able to discover patterns among health centers that led to

policy recommendations to the Institute of Public Health. They concluded that data mining and decision support methods, including novel visualization methods, can lead to better performance in decision-making.

The preceding factors remind us of an incident in the Philippines at the Rizal Medical Center in Pasig City in October 2006. Failing to implement strict sanitation and sterilization measures, the hospital contributed to the death of several new-born babies due to neonatal sepsis (bacterial infection). No one really knew what was going on until the deaths became more frequent. Upon examining hospital records, the Department of Health (DOH) found that 12 out of 28 babies born on October 4, for example, died of sepsis (Tandoc, 2006). With an integrated database and the application of data mining, the DOH could detect such unusual events and curtail them before they worsen.

More value for money and cost savings. Data mining allows organizations and institutions to get more out of existing data at minimal extra cost. KDD and data mining have been applied to discover fraud in credit cards and insurance claims (Kou, Lu, Sirwongwattana and Huang, 2004). By extension, these techniques could also be used

Data mining allows organizations and institutions to get more out of existing data at minimal extra cost.

Issues and Challenges

Applying data mining in the medical field is a very challenging undertaking due to the idiosyncrasies of the medical profession. Shillabeer and Roddick's work (2007) cited several inherent conflicts between the traditional methodologies of data mining approaches and medicine.

In medical research, data mining starts with a hypothesis and then the results are adjusted to fit the hypothesis. This diverges from standard data mining practice, which simply starts with the data set without an apparent hypothesis.

Also, whereas traditional data mining is concerned with patterns and trends in data sets, data mining in medicine is more concerned with the minority that do not conform to the patterns and trends. What heightens this difference in approach is the fact that most standard data mining is concerned mostly with describing but not explaining patterns and trends. In contrast, medicine needs these explanations because a slight difference could change the balance between life or death.

Whereas traditional data mining is concerned with patterns and trends in data sets, data mining in medicine is more concerned with the minority that do not conform to the patterns and trends.

The model analyzed the pixels and their RGB (please spell-out) content to find sufficient patterns to distinguish between malignant and benign tumors. Then the team applied the resulting model to other cases. They found that their model resulted to high accuracy in diagnosis with only a small standard deviation.

Adverse drug events (ADEs). Some drugs and chemicals that have been approved as non-harmful to humans are later discovered to have harmful effects after long-term public use. Wilson, Thabane and Holbrook (2003) revealed that the US Food and Drug Administration uses data mining to discover knowledge about drug side effects in their database. This algorithm called MGPS or Multi-item Gamma Poisson Shrinker was able to successfully find 67% of ADEs five years before they were detected using traditional ways.

Administration uses data mining to discover knowledge about drug side effects in their database.

We have seen how data mining applications could be used in early detection of diseases, prevention of deaths, the improvement of diagnoses, and even detecting fraudulent health claims. However, there are caveats to the use of data mining in healthcare.

to detect anomalous patterns in health insurance claims, particularly those operated by PhilHealth, the national healthcare insurance system for the Philippines.

By using data mining and visualization, medical experts could find patterns and anomalies better than just looking at a set of tabulated data.

Early detection and/or prevention of diseases. Cheng, Wei and Tseng (2006) cited the use of classification algorithms to help in the early detection of heart disease, a major public health concern all over the world. Cao, Maloney and Brusic (2008) described the use of data mining as a tool to aid in monitoring trends in the clinical trials of cancer vaccines. By using data mining and visualization, medical experts could find patterns and anomalies better than just looking at a set of tabulated data.

Early detection and management of pandemic diseases and public health policy formulation. Health experts have also begun to look at how to apply data mining for early detection and management of pandemics. Kellogg and associates (2006) outlined techniques combining spatial modeling, simulation, and spatial data mining to find interesting characteristics of disease outbreak. The analysis that resulted from data mining in the simulated environment could then be used towards more informed policy-making to detect and manage disease outbreaks.

Wong, Moore, Cooper and Wagner (2005) introduced WSARE, an algorithm to detect outbreaks in their early stages. WSARE, which is short for "What's Strange About Recent Events" is based on association rules and Bayesian networks. Applying WSARE on simulation models have been claimed to result to relatively accurate predictions of simulated disease outbreaks. Of course, claims of this nature always come with warnings to take precaution when applying these models in real life.

Non-invasive diagnosis and decision support. Some diagnostic and laboratory procedures are invasive, costly, and painful to patients. An example of this is conducting a biopsy in women to detect cervical cancer. Thangavel, Jaganathan and Easmi (2006) used the K-means clustering algorithm to analyze cervical cancer patients and found that clustering found better predictive results than existing medical opinion. They found a set of interesting attributes that could be used by doctors as additional support on deciding whether or not to recommend a biopsy for a patient suspected of having the cervical cancer.

Some diagnostic and laboratory procedures are invasive, costly, and painful to patients.

Gorunescu (2009) described how computer-aided diagnosis (CAD) and

endoscopic ultrasonographic elastograph (EUSE) were enhanced by data mining to create a new non-invasive cancer detection. In the traditional approach, doctors look at the ultrasound movie and decide on whether a patient is to be subjected to a biopsy.

The physician's judgment is primarily subjective, depending mostly on the interpretation of the ultrasound video. Gorunescu (2009) approached this problem in a different way, using data mining. He did not study patient demographics. Instead his team focused on the ultrasound movies. They first trained a classification algorithm using a multi -layer perceptron (MLP) on known cases of malignant and benign tumors.

Figure 3. EUSE Sample Movie Frame with Corresponding Histogram

Source: Gourunescu, 2009