1-1-2000

Data Mining in Health-Care: Issues and a Research Agenda

Monica Adya
Marquette University, monica.adya@marquette.edu

Published version. *AMCIS 2000 Proceedings*, 2000: 94-97. Permalink.

Data Mining in Healthcare: Issues and a Research Agenda

Niya Werts[1]
Monica Adya

Department of Information Systems, University of Maryland Baltimore County
E-mail: nwerts1@gl.umbc.edu;
adya@umbc.edu

Abstract

While data mining has become a much-lauded tool in business and related fields, its role in the healthcare arena is still being explored. Currently, most applications of data mining in healthcare can be categorized into two areas: decision support for clinical practice, and policy planning/decision making. However, it is challenging to find empirical literature in this area since a substantial amount of existing work in data mining for health care is conceptual in nature. In this paper, we review the challenges that limit the progress made in this area and present considerations for the future of data mining in healthcare.

Data-Mining in the Current Healthcare Climate

In the age of managed care where the cost-effectiveness and necessity of medical procedures are closely scrutinized by insurance providers, it is not surprising that data mining has begun to play an increasingly important role as data repositories swell with valuable information. Likewise, in the federal and state governments the proponents of health care reform demand comprehensive data-analysis of the effectiveness of current services, and the predictive value of new services. Not exempt from the sting of downsizing, hospitals also must review the use of the resources for optimum efficiency and reduced costs. One study noted that in 1998 one third of hospital respondents were performing some form of data-mining (Health Management Technology, 1998)

A Case for Data Mining

Data mining lends itself well to domains that manage large datasets or warehouses for several reasons.

- **Speed:** Both traditional statistical tools and AI-based data-mining tools aim to produce information by identifying valid, novel, potentially useful correlations and patterns in existing data (Chung and Gray, 1999). However, one of the key advantages of data-mining tools is their speed in working with large data sets (Chen and Sakaguchi, 2000). Considering that data is collected on a per encounter basis from each patient, medical care provider (including pharmacists), insurance payer and/or government agency, the massiveness of health care databases requires faster and more flexible analysis tools that can query in multiple dimensions. The Health Care Finance Administration's (HCFA) MEDPAR database alone contains 14 million discharge abstracts of Medicare funded acute care hospital stays annually (Goodall, 1999). The MEDPAR is not unique in its size as payers, providers, employers, and government agencies make care tracking at every level a priority. Faster analysis of large data sets and quicker report preparation can increase operating efficiency and reduce operating costs (Chen and Sakaguchi, 2000).
- **Variety of Applicable Models:** Data mining can employ a wide variety of techniques for predictive analysis. AI techniques such as neural networks, decision tree, genetic algorithms, nearest neighbor method, rule induction, and machine learning are some of the techniques for data mining (Lavrac, 1999; Mitchell, 1999).
- **Prospective Analysis:** Data-mining extracts predictive information from large databases. This ability has reaped tremendous benefits for the business sector, and can potentially do the same in the health care arena.

Challenges of Data Mining

With these advantages, what limits the broad applicability of data mining techniques for health care issues? The most challenging aspect of data mining is the very nature of this technique – its reliance on data. Data-mining insights are only as "accurate" as the data being mined. In the health care arena, this becomes an even more challenging problem.

[1] Corresponding Author

- **Data Quality:** Data quality in health care is a significant challenge. For one, it is hard to find data accurate and complete data. The problem becomes more pronounced when inter- and intra-agency data standards vary greatly or are not enforced. A case in point is the Minimum Data Set (MDS) maintained by HCFA. HCFA requires all hospitals to record data for the MDS. This data set requires a patient to respond to over 300 questions at the time of check-in. Since filling out the MDS is largely a manual process, it is challenging for both the patient and the care provider to respond to and record such a large volume of responses. Consequently, the MDS suffers from several problems including missing information, incorrect entries. Additionally, there are significant differences between the quality of MDS data sets from various states in the US. Since the hospitals and HCFA are independently functioning agencies, it is challenging to enforce data standards and consequently the MDS suffers from the classic problems found in health care data.
- **Data Sharing**: Privacy is another challenge to the situation. Organizations that are able to maintain large data warehouses on health related factors are unwilling to share the data for reasons of maintaining patient privacy. This problem is further magnified because proprietary restrictions can be an obstacle to mining large datasets with potentially valuable "hidden" information**.** HMOs and health insurance organizations are unwilling to provide even an insight into their data sets for proprietary reasons. This poses a significant challenge to health care fraud detection studies that require inputs from HMOs about patient billing transactions.
- **Start up Costs-:** Building and maintaining a data-warehouse for the most efficient and accurate mining can be an expensive challenge. Design errors early on the process can have a far- reaching negative impact. A reported 85% of first generation data-warehouses have failed or have not met desired goals (Winans, 1999). Clearly, downsized hospital budgets may not be able to adequately support or sustain such an effort.
- **No Concrete Answers:** While data-mining is a valuable tool for discovering patterns in large databases, the sensitivity and the specificity of data-mining tools will impact the predictive value of the gathered information. Sensitivity and specificity are particularly important in medical data mining Lavrac (1999) defines sensitivity as the number of positive cases that are classified as positive and specificity as the number of negative cases classified as negative. Sensitivity is an indicator of detection rate, while specificity can be interpreted as a "false alarm" rate. Sensitivity and specificity are high criteria for all forms of biomedical testing. For example, the DNA micro-array technology which uses the self-organizing map (SOM) technique has shown great promise in the area of increasing the sensitivity and specificity of cancer diagnosis even at a molecular level (Stephenson,1999). Clinical practitioners in the healthcare arena may also be resistant to the value of data warehousing and data mining when point of service care is guided primarily by data produced in the midst of present circumstances (lab tests, physiological monitors, etc.) and not by trend analysis based on historic data (Allen, 1997).

  There is also some debate as to whether information obtained from data mining has real world value (Wasserman, *et al.*, 1999). In their use of the Data Mining Surveillance System (DMSS), which monitors infection control, Moser, *et al*. (1999) noted that applying filtering tools could reduce a number of "uninteresting" trends. Noteworthy trends do not explain causation, and further investigation by persons with domain knowledge is crucial.

Current Applications of Data-Mining in Health Care and Medicine

Data mining applications are currently being applied to two main branches in health care and medicine: medical decision support, and policy planning/decision making.

Medical Decision Support

The broad goal of an MDSS is to provide either direct or indirect assistance in medical decision-making by exploring an extensive medical knowledge base (Degoulet and Fieschi, 1997). Data mining takes on a key role in knowledge discovery in medical databases. For example, data-mining techniques have been utilized to pinpoint the key indicators of myocardial infarction in 12 lead EKG's (Burn-Thorton and Edenbrandt, 1998), as well as to help determine the best ways to stop the spread of infections in a health care setting (Brosset, *et al.*, 1998). With the rising popularity of gene therapy, data mining has been used extensively in the genome mapping process (Regalado, 1999). Identifying disease-causing genes is a primary objective of pharmaceutical companies, and data mining is considered by some to be a "core part" of drug discovery.

Health Policy Planning

Data mining has the potential to strongly impact health policy planning as massive government, state, and private sector health care databases are analyzed in order to increase efficiency and quality while reducing costs. The COREPLUS and SAFS systems projects were designed to analyze outcomes of hospital care and model resource consumption including cost and length of stay

(Goodall, 1999). The Health Insurance Commission of Australia data-mines to aid in the development of automated risk management systems (Bushell,1999).

In addition to resource management, and quality control systems, data-mining shows promise in aiding in the prevention of health care-fraud. Health care fraud and abuse have become primary concerns for a number of government agencies as its deleterious impacts on public health and government fiscal spending have come under increasing scrutiny. In 1997, US taxpayers were estimated to have lost more than $500 in improper payments due to fraudulent billing for every one of the 38.5 million Medicare beneficiaries (Health Care Fraud Report,1997). Fraudulent billing practices may also disguise improper and inadequate medical care.

The Medicare Integrity Program and Payment Safeguards (MIP) Program, the Corrective Coding Initiative, comprehensive compliance program guidelines, and educational efforts all aim to prevent fraudulent claims from being made and/or paid (CIC-3, 1999). However, accurate and efficient fraud detection techniques still must be available to identify post-payment fraudulent claims and the offenders. The need for integrated fraud detection technologies becomes particularly acute as Medicare/Medicaid and other public funds are increasing disbursed through automated payment systems (Sparrow, 1997).

Data mining has already shown promise in furthering the fraud prevention agenda in health care. The state of Texas estimates that by the year 2000 it will recover 14 million dollars a year from fraudulent Medicaid billers due to the "suspect-generating" capabilities of neural network software produced by the EDS information technology company (Intelligent Systems Report, 1998).

Future Issues

At present, data mining is still considered to be in its infancy. As new applications continue to emerge, certain issues for the data mining in the healthcare field will need to be considered .

- **Improved Data-Sharing Between Agencies –** Several organizations such are overcoming privacy issues that limit data sharing by blocking out critical patient identification information such as Social Security Numbers. For instance, HCFA has established a data availability link on their web site (www.hcfa.gov**)** and a hotline to support data exchange for research purposes. The challenge will be to overcome propriety restrictions imposed by private organizations. Researchers may want to develop contractual relationships with such organizations which may limit the publication of explicit findings but will provide the opportunity to work with real data instances. Finally, researchers will increasingly find public data becoming available on the web. A case to point is the gene expression banks that are being maintained on the Internet.
- **Integrated Mining Tools for the Web-** Text Mining has recently come into focus for mining on the web (Zorn and Emanoli, 1999). However, the web encompasses a large amount of non-text based information that may need to be considered in the future, especially as online telemedicine begins to flourish. Another aspect that is gaining significant prominence is the automation of billing transactions via the Internet. This change will provide a great opportunity to use data mining techniques to detect fraudulent transactions online.
- **Data/ Data Warehouse Standardization-** While a uniform data warehouse standards may take a while to emerge, there needs to be a "bridge" to help facilitate mining of data from various sources. The development of inter-agency, flexible standards may mitigate the need for extensive cleaning tools.
- **Data Warehouse Compression-** As data warehouses continue to grow, the problems for mining a massive data set will continue. A process of "compression" without compromising data quality would mitigate some of those issues. Scaling a wider range of existing tools for large datasets is met with a number of obstacles including visualization, data and computational complexity, and storage requirements (Huber, 1999). This is particularly a concern if mining is going to continue to be a driving force behind desktop decision support.
- **Focus on representation and interpretation of findings:** Early, formative research on data mining and more broadly Knowledge Discovery in Databases addressed the need to conceptualize the discovery of new knowledge in databases as a process with data mining being one important component of this (Fayyad, *et al*., 1996). In our experience, most literature on data mining focuses on the application of data mining techniques for pattern extraction and reduces the significance of pattern interpretation. This problem becomes magnified due to lack of sufficient empirical studies. Consequently, researchers have not adequately examined issues that relate to presentation and interpretation of these patterns. Data mining can potentially benefit from research in the area of data visualization but we have yet to find a study in health care that addresses data visualization and pattern interpretation techniques. To truly understand the nature of outcomes from data mining techniques, it is important to think of data mining as one component in an overall decision support environment that integrates data cleaning, data visualization, and interpretation.

Conclusions

Empirical research in data mining for health care is limited. Several factors, most importantly those related to data quality and availability, have limited research in this area. In our assessment, addressing these two issues will give significant impetus to research in this arena. In particular, health care organizations and governing bodies need to establish strong data quality standards before the environment can be conducive to productive research. Secondly, a positive partnership must be established between organizations maintaining data warehouses and institutions wanting to further the research in the field. Establishing privacy requirements and standards can address some concerns in this regard.

References

Allen, H.G. "The Healthcare Data Warehouse," Data Management Review, (7:3), 1997, pp. 50-51.

Brosset, S. E., Sprague, A. P., Hardin, J. M., Waites, K. B., Jones, W. T. and Moset, S. A. "Association Rules and Data Mining in Hospital Infection Control and Public Health Surveillance," Journal of the American Medical Informatics Association, (5:4),1998, pp. 373-381.

Burn-Thorton, K. and Edenbrandt, L. "Myocardial Infarction- Pinpointing the Key Indicators in the 12-Lead ECG Using Data Mining, "Computers and Biomedical Research, (31:1998, pp. 293-303.

Bushell, S. "Healing by Numbers," Newsweek, (117:6174),1999, pp. 94-96.

Chen, L.-d. and Sakaguchi, T. "Data Mining: Information storage and retrieval systems," Information Systems Management, (17:1),2000, pp. 65-71.

Chung, H. M. and Gray, P. "Special Section: Data Mining," Journal of Management Information Systems, (16:1),1999, pp. 11-17.

Degoulet, P. and Fieschi, M. Introduction to Clinical Informatics, Springer, New York,1997.

Fayyad, U., Haussler, D. and Stolorz, P. "Mining Scientific Data," Communications of the ACM, (39:11),1996, pp.

Frederique, B. "Fool's gold?, "Sales & Marketing Management, (150:6),1998, pp. 58-62.

Goodall, C. R. "Data Mining of Massive Datasets in Healthcare," Journal of Computational & Graphical Statistics, (8:3),1999, pp. 620-635.

Health Care Fraud Report. "Health Care Fraud Report Fiscal Year 1997," www.usdoj.gov/01whatsnew/hcffraud2.htm

Health Management Technology. "Hospital Y2k strategic planning lags, survey says," Health Management Technology, (19:8),1998, pp. 8-11.

Intelligent Systems Report. "Neural networks helps Texas detect Medicaid Fraud," Available online: http://lionheartpub.com/ISR/ISRsubs/isr-2-98/neuralnetworks.html

Huber, P. "Massive Datasets Workshops: Four Years After," Journal of Computational & Graphical Statistics, (8:3),1999, pp. 635-638.

Lavrac, N. "Selected techniques for data mining in medicine," Artif Intell Med, (16:1),1999, pp. 3-23.

Mitchell, T. " Machine Learning and Data Mining," Communications of the ACM, (42:11), 1999, pp. 30-36.

Moser S, T., Warren, J. T. and Brossette, S. "Application of Data-mining to Intensive Care Unit Microbiologic Data," Emerging Infectious Diseases, (5:3),1999, pp. 454-459.

Regalado, A. "Mining the Genome," Technology Review, 1999 September/October, pp. 56-62.

Sparrow, M. "Automation Fosters Health Care Fraud," http://www.govtech.net/1997/gt/feb/feb97-automationfosters.../feb97-automationfostershea.shtml

Stephenson, J. "Lab on a Chip Shows Promise in Defining and Diagnosing Cancers," Journal of the American Medical Association, (282:19), 1999, http://jama.ama-assn.org/issues/v282n19/full/jmn1117-1.html.

Wasserman, L., Bruce, P. and Brodley, C. "Something from Nothing," American Scientist, (87:2), 1999, p.100

Winans, C. "Building from the bottom up can result in dirty datamarts," Best's Review/ Life Health Insurance Edition, (99:9),1999, pp. 87-89.

Zorn, P. and Emanoli, M. "Mining Meets the Web," Online, (23:5),1999, pp. 16-25.