APPLICATION OF DATA MINING TO HEALTH CARE

Nakul Soni
Information Technology Department
Birla Vishvakarma Mahavidayalaya
V.V.Nagar, Gujarat, India
nakul_soni911@yahoo.com

Chirag Gandhi
Electronics Department
Birla Vishvakarma Mahavidayalaya
V.V.Nagar, Gujarat, India
chirag.gandhi@gmail.com

Abstract— Data mining has been used extensively in many fields like retail, e-business, marketing, etc. and has provided pioneering results. This paper presents the application of data mining in health care.

The paper compares data mining with traditional statistics, gives some advantages of automated data systems, enumerates the steps involved in data mining process. A growing number of data mining applications in health care have been discussed in this paper. Also the issues and challenges faced by data mining in health care are presented.

Keywords- Norwalk virus; West Nile virus; Listeriosis; GIS; computer-aided diagnosis (CAD); endoscopic ultrasonographic elastography (EUSE); multi-layer perceptron (MLP); malignant tumor; benign tumor.

I. DATA MINING PERSPECTIVE

Data mining is the process of extracting knowledge from database/data mart/data warehouse. A database is a collection of data that is organized so that its contents can easily be accessed, managed, and updated. Databases usually include a query facility, and the database community has a tendency to view data mining methods as more complicated types of database queries. For example, standard query tools can answer questions such as, "How many operated patients stay in hospital longer than 10 days?" Data mining is valuable for more complicated queries such as, "What are the important preoperative indicators of excessive length of stay?"

Data mining techniques can be implemented retrospectively on massive data in an automated manner, whereas traditional statistical methods used in epidemiology require custom work by experts. Traditional methods generally require a certain number of predefined variables, whereas data mining can include new variables and accommodate a greater number of variables. Application of either statistical or data mining techniques requires substantial human effort, and collaboration, rather than competition, between the two fields would contribute to each other more effectively by building on each other's strengths to create synergy[1].

There are advantages of an automated surveillance system, regardless of whether based on data mining or statistical methods. These include

(1) A reduction in time and effort by the end user;
(2) The ability of the system to examine multiple areas simultaneously;
(3) A decreased potential for human error;
(4) Presentation of data in the correct format; and
(5) Accessibility of data anytime and anywhere; [1].

An example of automated surveillance system is as follows. Military medical researchers in the United States are using such a system that gathers data from military medical facilities worldwide as well as from other healthcare sources. The system detects outbreaks (e.g., the Norwalk virus in San Diego in 2002) when individual healthcare practitioners may not be able to see the big picture, and it monitors progression of diseases (e.g., West Nile virus and Listeriosis)[2].

II. DATA MINING PROCESS

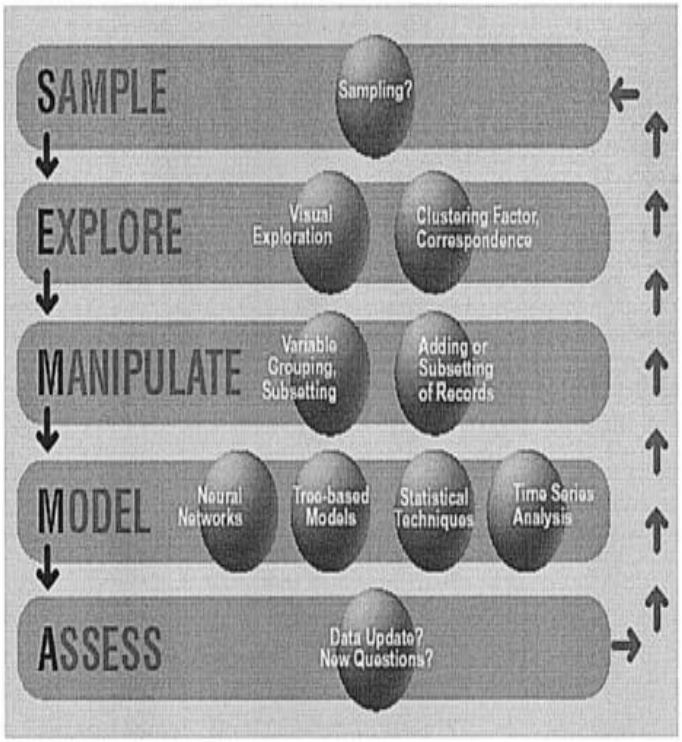

Figure 1. The SEMMA data mining process.

Data mining can be viewed as a process rather than a set of tools, and the acronym SEMMA (sample, explore, modify, model, and assess) refers to a methodology that clarifies this process. Figure 1 shows the steps followed in SEMMA process.

The SEMMA method divides data mining into five stages:

1)Sample: to draw a statistically representative sample of data;
2)Explore: to apply exploratory and statistical visualization techniques to data;
3)Modify (or manipulate): to select and transform the most significant predictive variables;
4)Model: to model the variables to predict outcomes;
5)Assess: to confirm a model's accuracy.

III. WHY DATA MINING IN HEALTH SECTOR?

For centuries concrete data and evidence have been used for taking medical decisions (in medical terms it is known as evidence-based medicine or EBM). John Snow, considered to be the father of modern epidemiology, used maps in conjunction with early forms of bar graphs in 1854 to discover the source of cholera and proved that it was transmitted through the water supply [3].

Snow counted the number of deaths caused by cholera and plotted the victims' address on the map of London as black bars. He discovered that most of the deaths clustered towards a specific water pump in London (shown by the red circle in Figure 2).

Figure 2. Bar chart plotted on map of London.

Snow was able to personally collect, and analyze the data during his times because the volume of information was manageable. Today, the size of the population, the volume of data, and the speed of disease outbreaks make it almost impossible to accomplish what Snow 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.

IV. THE IMPORTANCE AND USES OF DATA MINING IN MEDICINE AND PUBLIC HEALTH

From the above discussion we can say that use of data mining in health care is the need of the hour. To support this notion we enumerate several arguments below.

Data Overload - The vast amount of data present in large databases/data warehouses/data marts contains a wealth of knowledge to be discovered. But it is extremely difficult, if not impossible, for a human being to do this. Hence we need an automated data mining system to mine the information.

Evidence-based medicine and prevention of hospital 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 [4]. By mining hospital records, such safety issues could be discovered and addressed by hospital management and government regulators.

Policy-making in public health - To analyze similarities between community health centers in Slovenia, GIS and data mining were employed. Using data mining, patterns among health centers were discovered, that led to policy recommendations to Slovenia's Institute of Public Health. It was concluded that "data mining and decision support methods, including novel visualization methods, can lead to better performance in decision-making."

More value for money and cost savings - By using data mining the organizations can extract a lot of useful knowledge from their existing data at a minimal cost. Data mining has been applied to discover fraud in credit cards and insurance claims. Similarly, these techniques can also be used to detect anomalous patterns in health insurance claims, and thus fraudulent claims can be detected.

Early detection and management of diseases and pandemics - Data mining has been used as a tool to aid in monitoring trends in the clinical trials of cancer vaccines [5] and in early detection of heart disease [6]. By using data mining and visualization, medical experts could find patterns and anomalies better than just looking at a set of tabulated data.

Health experts have also begun to look at how to apply data mining for early detection and management of pandemics. Experts have 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, these sorts of claims always come with warnings to take precaution when applying these models in real life [7].

Non-invasive diagnosis and decision support - Some laboratory procedures used for diagnosis are invasive, costly and painful to patients. As an example, conducting a biopsy in women to detect cervical cancer.

Gorunescu described how computer-aided diagnosis (CAD) and endoscopic ultrasonographic elastography (EUSE) were enhanced by data mining to create a new noninvasive cancer detection method. In the traditional approach, doctors look at the ultrasound movie and decide on whether a patient is to be subjected to a biopsy [8].

The doctor's judgment is primarily subjective, depending mostly on their interpretation of the ultrasound video. Gorunescu 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 [8].

The model analyzed the pixels and their RGB 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 a small standard deviation [8].

Adverse drug events (ADEs) - Some drugs and chemicals that have been approved as non-harmful to humans are later discovered to have harmful effects on humans after long-term public use. Wilson 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 than they were detected using traditional ways [9].

V. 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 cite 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 that hypothesis. This diverges from standard data mining practice, which simply starts with the data set without a hypothesis.

Also, traditional data mining is concerned about patterns and trends in data sets, whereas data mining in medicine is more interested in the minority that do not conform to the patterns and trends. All the more, most standard data mining results are concerned mostly with describing but not explaining the patterns and trends. In contrast, medicine needs those explanations because a slight difference could change the balance between life and death.

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 [10]. The failure of data mining to provide conclusive results indicates the current lack of credibility of data mining in these particular niches of healthcare.

Even if data mining results are credible, convincing the health practitioners to change their habits based on evidence may be a bigger problem. There are a couple of cases where hospital doctors refused to change hospital policy even when confronted with evidence [11]. In one case, it was found that doctors coming out of autopsy without washing hands and 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.

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 [8].

CONCLUSION AND RECOMMENDATIONS

This survey paper gives an overview of how data mining is being used in crucial fields like that of health care and medicine. The health care organizations, which could not reap out advantages of data mining still, should look into these applications and find ways of extracting information from their databases using data mining techniques.

As an example, Indian Ministry of Health should coordinate with government and private hospitals to collect and analyze public health indicators. Then the Ministry may apply data mining techniques to find trends in disease outbreaks or deaths (e.g., infant mortality), per region and per hospital.

Also, Ministry of Health will be able to uncover hidden patterns in deaths or diseases that may lead to better health policies like better vaccination planning, identification of disease vectors like malaria, prevention of hospital errors, etc. The insurance companies may also use data mining to find and stop anomalous insurance claims.

Along with this, the challenges faced by data mining should be tackled effectively. For example, an organization must formulate clear policies on the privacy and security of patient records.

Also the medical practitioners should be taken into confidence about the effectiveness of data mining results. So that they follow the results and act in the required manner

Thus we conclude that data mining can be of great help in health care if applied effectively..

REFERENCES

[1] Mary K. Obenshain, "Application of Data Mining Techniques to Healthcare Data**",** Infection Control and Hospital Epidemiology, Vol. 25, No. 8 (August 2004), pp. 690-695.

[2] Johnston G. System adds to biodefense readiness. *Bio-IT World*. November 1, 2002, Available at www.bio-itworld.com/news/110102_report1436.html.

[3] Tufte, E. (1997), "Visual Explanations, Images and Quantities, Evidence and Narrative", Connecticut: Graphics Press.

[4] Health Grades, Inc. (2007), The Fourth Annual Health Grades Patient Safety in American Hospitals Study.

[5] Cao, X., Maloney, K.B. and Brusic, V. (2008)," Data mining of cancer vaccine trials: a bird's-eye view", Immunome Research, 4:7, DOI: 10.1186/1745-7580-4-7.

[6] Cheng, T.H., Wei, C.P., Tseng, V.S. (2006), "Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches", Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06).

[7] 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.

[8] Ruben D. Canlas Jr., "Data Mining in Healthcare: Current Applications and Issues".

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

[10] 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.

[11] Ayres, I (2008). Super Crunchers. New York: Bantam Books.