Heuristics in Medical Data Mining

Susan E. George
University of South Australia, Australia

HISTORICAL PERSPECTIVE

Deriving—or discovering—information from data has come to be known as data mining. Within health care, the knowledge from medical mining has been used in tasks as diverse as patient diagnosis (Brameier et al., 2000; Mani et al., 1999; Cao et al., 1998; Henson et al., 1996), inventory stock control (Bansal et al., 2000), and intelligent interfaces for patient record systems (George at al., 2000). It has also been a tool of medical discovery itself (Steven et al., 1996). Yet, it remains true that medicine is one of the last areas of society to be "automated," with a relatively recent increase in the volume of electronic data, many paper-based clinical record systems in use, a lack of standardisation (for example, among coding schemes), and still some reluctance among health-care providers to use computer technology. Nevertheless, the rapidly increasing volume of electronic medical data is perhaps one of the domain's current distinguishing characteristics, as one of the last components of society to be "automated."

Data mining presents many challenges, as "knowledge" is automatically extracted from data sets, especially when data are complex in nature, with many hundreds of variables and relationships among those variables that vary in time, space, or both, often with a measure of uncertainty, as is common within medicine. Cios and Moore (2001) identified a number of unique features of medical data mining, including the use of imaging and need for visualisation techniques, the large amounts of unstructured nature of free text within records, data ownership and the distributed nature of data, the legal implications for medical providers, the privacy and security concerns of patients requiring anonymous data used, where possible, together with the difficulty in making a mathematical characterisation of the domain.

Strictly speaking, many ventures within medical data mining are better described as exercises in "machine learning," where the main issues are, for example, discovering the complexity of relationships among data items, or making predictions in light of uncertainty, rather than "data mining," in large, possibly distributed, volumes of data that are also highly complex. Large data sets mean not only increased algorithmic complexity but also often the need to employ special-purpose methods to isolate trends and extract "knowledge" from data. However, medical data frequently provide just such a combination of vast (often distributed) complex data sets.

Heuristic methods are one way in which the vastness, complexity, and uncertainty of data may be addressed in the mining process. A heuristic is something that aids discovery of a solution. Artificial intelligence (AI) popularised the heuristic as something that captures, in a computational way, the knowledge that people use to solve everyday problems. AI has a classic graph search algorithm known as A* (Hart et al., 1968), which is a heuristic search (under the right conditions). Increasingly, heuristics refer to techniques that are inspired by nature, biology, and physics. The genetic search algorithm (Holland, 1975) may be regarded as a heuristic technique. More recent population-based approaches have been demonstrated in the Memetic Algorithm (Moscato, 1989), and specific modifications of such heuristic methods in a medical mining context can be noted (Brameier et al., 2000).

Aside from the complexity of data with which the medical domain is faced, there are some additional challenges. Data security, accuracy, and privacy are issues within many domains, not just the medical (Walhstrom et al., 2000). Also, while ethical responsibility is an issue in other contexts, it is faced by the medical world in a unique way, especially when heuristic methods are employed. One of the biggest ethical issues concerns what is done with the knowledge derived combined with a "forward-looking responsibility" (Johnson et al., 1995). Forward-looking responsibility is accountable for high-quality products and methods and requires appropriate evaluation of results and justification of conclusions.

George (2002) first identified and proposed a set of guidelines for heuristic data mining within medical domains. The proposed guidelines relate to the evaluation and justification of data-mining results (so important when heuristic "aids to discovery" are utilised that "may" benefit a solution) and extend to both where and how the conclusions may be utilised and where heuristic techniques are relevant in this field. The remainder of this article summarises some heuristic data-mining applications in medicine and clarifies those proposed guidelines.

BACKGROUND

First, we will explain some of the heuristic methods that have been employed in medical data mining, examining a range of application areas. We broadly categorise applications as clinical, administrative, and research, according to whether they are used (or potentially used) in a clinical context, are infrastructure related, or are exploratory, in essence. We also note that with the exception of some medical imaging applications and mining of electronic medical records, the databases are small.

There is a wide variety of automated systems that have been designed for diagnosis—systems that detect a problem, classify it, and monitor change. Brameier and Banzhauf (2000) described the application of linear genetic programming to several diagnosis problems in medicine, including tests for cancer, diabetes, heart conditions, and thyroid conditions. Their focus was upon an efficient algorithm that operates with a range of complex data sets, providing a population-based heuristic method that is based upon biological principles. Their heuristic method is based on an inspiration from nature about how "introns" (denoting DNA segments with information removed before proteins are synthesised) are used in generating new strings. They suggest that introns may help to reduce the number of destructive recombinations between chromosomes by protecting the advantageous building blocks from being destroyed by crossover. Massive efficiency improvements in the algorithm are reported.

An interesting administrative application of data mining in a medical context comes in the area of interfaces for electronic medical records systems that are appropriate for speedy, accurate, complete entry of clinical data. At the University of South Australia, George et al. (2000) reported on the use of a data-mining model underlying an adaptive interface for clinical data entry. As records are entered, a database is established from which predictive Bayesian models are derived from the diagnosis and treatment patterns. This heuristic is used to predict the treatment from new diagnoses that are entered, producing intelligent anticipation. The predictive model is also potentially incremental and may be re-derived according to physician practice. This application addresses issues in incremental mining, temporal data, and highly complex data with duplication, error, and nonstandard nomenclatures.

One interesting ongoing database mining project at Rutgers is the development of efficient algorithms for query-based rule induction, where users have tools to query, store, and manage rules generated from data. An important component of the research is a facility to remember past mining sessions, producing an incremental approach. They are using heuristics for efficiently "re-mining" the same or similar data in the face of updates and modifications. In their trials, a major insurance company was trying to explore anomalies in their medical claims database. The new data-mining techniques aided the isolation of high-cost claims and scenarios in each disease group that would lead to high-cost claims. They also identified characteristics of people who were likely to drop out of their health plans and locations where there were higher dropout rates. This is a general approach to mining, where information from prior mining is utilised in new mining to prevent the need to compute relationships from scratch every time data is added to the database. This is, naturally, a general approach to mining large-scale changing databases that may be considered in a variety of fields.

Medical data mining is a natural method of performing medical research, where new relationships and insights are discovered in human health. The University of Aberdeen address the problem of mammographic image analysis using neural nets together with conventional image analysis techniques to assist in the automated recognition of pathology in mammograms (Undrill, 1996). The group also addresses the use of genetic algorithms for image analysis, applying this powerful general optimisation technique to a variety of problems in texture segmentation and shape analysis in two-dimensional and three-dimensional images (Delibassis, 1996). Mining information from the data in these tasks must address many of the problems of finding patterns within large volumes of highly complex data.

Banerjee et al. (1998) described the use of data mining in medical discovery. They reported on a data-mining tool that uncovered some important connections between diseases from mining medical literature. The data-mining tool compared the article titles in various medical journals. Medical discoveries were made, such as the connection between estrogen and Alzheimer's disease, and the relationship between migraine headaches and magnesium deficiency. Ngan et al. (1999) reported on medical discovery using data mining based upon an evolutionary computation search for learning Bayesian networks and rules. They were able to discover new information regarding the classification and treatment of scoliosis as well as knowledge about the effect of age on fracture, diagnoses, and operations and length of hospital stays.

Kargupta and colleagues (1999) were interested in an epidemiological study that involved combining data from distributed sources. Their study investigated what affects the incidence of disease in a population, focusing upon hepatitis and weather. They illustrated the collective data-mining approach, emphasising the importance within medicine of merging data from heterogeneous sites. Their solution minimises data communication using decision-tree learning and polynomial regression. As more hospitals and general practitioners, pharmacists, and other health-care-related professions utilise electronic media, mining ventures are going to have to cope with mining across data sources. They will have to address issues such as those addressed by this study, such as minimising data exchange and adopting suitable heuristic approaches.

GUIDELINES FOR HEURISTIC MEDICAL DATA MINING

Responsibility is clearly an issue in medical data mining given the unique human arena in which the conclusions are outworked. If medical data-mining products are ever produced by "professionals" or are ever exploited "commercially," there may be serious legal consequences for their creators in the wake of harmful consequences from information produced. In the context of software engineering, the computer field seeks to promote high-quality software products, so too,