ISSUES IN DATA MINING FOR MEDICAL RESEARCH

Abbas Heiat, Montana State University-Billings, aheiat@msubillings.edu, 406-657-1627
Nafisseh Heiat, Montana State University-Billings, nheiat@msubillings.edu

INTRODUCTION

Today there is a mixture of technologies used to mine medical data, and all of them use some combination of statistical operations, artificial intelligence, machine learning, inference, neural networks, and information technologies.

Intelligent Agents- Agents use a set of rules to decide which action or actions they should take, and these are usually called production rules in Artificial Intelligence. The problem with this approach is that the user needs to recognize the opportunity for employing an agent, take the initiative in programming the rules, endow the agent with explicit knowledge specified in an abstract language, and maintain the rules over time, as habits or events change. The key problem with such systems is that they require a large amount of work from the knowledge engineers. Furthermore, the knowledge of the agent is fixed and cannot be customized to the habits of individual users. Even with the rule bases and knowledge bases mentioned above, there is still the requirement to incorporate new knowledge into the intelligent agents during their lifetime. Here the agents will have to adapt their own behavior and extend their own knowledge, instead of relying on users to constantly modify the rule bases and knowledge bases. Learning refers to this modification of behavior as a result of experience.

The simplest learning technique that an intelligent agent can use is statistical analysis, with the goal of finding any correlation among events of interest by periodically scanning and analyzing the logs of user actions to find repeated sequences of actions. Also EVA (evolving agent) technology uses statistical analysis to find terms that co-occur and should be added to a query. When an agent needs to reason with imprecise or incomplete information, fuzzy logic can be employed.

Neural Networks- are another technology for intelligent data mining. Neural networks handle unstructured data or noisy data effectively. These are often difficult to process using rigid reasoning techniques. Intelligent agents based solely on neural networks can only learn locally; that is, their learning experiences are restricted to the documents they have scanned or the Web sites they have traveled through. Due to the historical collection and growing nature of online medical information, a single agent can only investigate a tiny fraction of all available information. To expand the learning horizon and to create more intelligent agents, one needs a learning algorithm, such as a genetic algorithm, that can operate at a higher level and view things from an inter-agent perspective.

Genetic Algorithm- By approaching the learning algorithm from two different levels — the local level of individual agents and the global level of inter-agent operation — we can ensure the optimization of each agent from local knowledge, while genetic algorithms will act as a "driving force" to evolve the agents collectively based on global knowledge. The goal is to "breed" a new generation of agents that benefit from the learning experiences of individual "parent" agents and the collective learning experiences of previous generations. Evolutionary algorithms build on so-called neuro-genetic algorithms, mimicking natural selection by terminating poor performers and combining successful ones to produce new offspring and create an agent-based search-and-retrieval system for unstructured data. This combines neural nets with natural language processing and a genetic algorithm, enabling the system to learn at two levels — that of the individual agent and of the pooled knowledge of all the

agents. Because it builds with NLP technology, it can easily find information under conditions which constantly change, as the Internet does.

Medical Applications - In medical applications, individuals will have a Personal Profile that contains their personal information. The user's computer will store this profile (and can, at the user's option, be securely stored in a corporate-wide or global directory), and the first time that an individual visits a Web site that supports open profiling, the Web site will request information from the Personal Profile. The individual has the choice of releasing all, some, or none of the requested information to the Web site. In addition, if the Web site collects additional information about the individual's preferences, it can (with the individual's permission) store that information in the Personal Profile for future use. On subsequent visits, the individual can authorize the Web site to retrieve the same personal information without asking permission each time.

Open profiling can facilitate commerce and services on the Internet while enhancing personal privacy, as users want personalized information, entertainment, and services. Companies and service organizations worldwide want to take advantage of the 1-to-1 nature of communications on the Internet to provide their customers and visitors with such personalized information, entertainment, and services. While drug manufacturers will want to target their messages to the needs and wants of specific audiences.

To gather the information that makes this personalization and targeting possible, Web sites ask their visitors for information — who they are, where they live, what they do, etc. A single individual might provide the same information to dozens (or even hundreds) of Web sites over time. This can be complex, time-consuming, and inconvenient. With OPS, an individual only has to provide the most frequently requested information once, so it saves time and helps to eliminate frustration. Also, it enhances personal privacy by putting control for releasing the information in the hands of individuals, instead of Web sites.

Privacy And Security Issues - Possible negative aspects of open profiling result when you unleash an intelligent agent, as you entrust it with some information. In any case, you have given it something of yourself that should only be used under carefully circumscribed conditions. In its travels through cyberspace, your agent will encounter agents or servers that could steal or modify that information. Or, your medical information could be invaded by a virus, and your records changed or altered.

The security of intelligent agents, particularly mobile ones, is a complex problem, particularly since we haven't had enough experience with this technology to help us guess at the full range of possible misadventures. Authentication of both the agent and the server is critical for agents to roam comfortably from one site to another

Agents may require considerable computing power in order to carry out their tasks. Not only do they monitor and store information, but they may perform computations, either at the host or client site. Since they operate in the background, their use of computing resources may not be obvious. This concerns any site that agrees to host mobile gents, as well as the desktop owner.

Mobile agents raise all kinds of specters for system administrators. Since they save computing resources by performing their work at the site housing the data, they occupy the resources of the host computer, rather than the user's computer. Aside from using someone else's computing power; they might also wreak havoc with the remote computer system. How will systems differentiate between benign, well-

intentioned intelligent agents that gather information for remote users and pernicious attackers bent on making mischief? One approach would limit the kinds of activities a mobile agent can perform. Another would confine their interaction to a specified site with an interface that allows them to find information in specified locations, but prevents them from entering the computer proper.

Privacy of information is a second question that mobility raises. Intelligent shopping agents, for instance, may travel to several sites to gather prices. They may well transport the prices of several competitors as they make their rounds. Could a clever competitor extract that information? Or falsify it so they can make a sale? Many online sites are threatened by mobile agents, because they don't want to have their prices compared to others, or because they object to having their computing power occupied by robots instead of people. Implementations of artificial intelligence in medicine bring up some important issues:

- In complex, high-pressure situations, can we trust an agent system to offer the best potential solutions or will it miss something important?
- How much can we trust an agent-based system to automate our actions? Do we want them to check with us for every initiative they take? Can we delegate authority little by little, so that we edge into a partnership with them? Will they hijack our work, as MS Word does in renumbering my outlines?
- Do we trust someone else's agent to be who it says it is? Can we develop widely accepted and secure authentication mechanisms?
- No quality control, dubious credibility of documents
- Automatic indexing is language dependent with unusable subject descriptions of non-English documents
- Query construction with Boolean operators, with very long ranks of search hits.

Medical Data Mining Concerns - While the advancement of these new data mining processes is providing benefits, never the less there are negative aspects worth taking a concerned look at mainly quality data, privacy, and cost. The first item of concern in any data mining operation is the quality of raw data. While a data warehouse does perform data unification and cleansing activities, it is a management responsibility to insure that medical forms, such as protocols, do not contain misleading information or missing information. Another important component of medical data are images, and "a medium size hospital handles about one million images per year" [9], it stands to reason that with that large of number, errors will happen. It is the reduction of errors that is required to eliminate false data, or maybe even false diagnosis.

Another area of concern is the threat of "loss of privacy". Here with new technologies requiring more and more data, there are many negative concerns such as:

- Who has access to my medical records?
- How will my privacy be maintained?
- Will I be penalized by pre-conditioned medical conditions (Buying Health Insurance)?
- Will I be diagnosed with someone else's records?

By utilizing data mining techniques we are now able to sift through the mountains of medical research records that have accumulated over the last hundred years. As the potential benefits of new data mining processes are explored, the negative cost incurred will also have to be balanced. I believe that we have only mined the tip of the mountains of raw data, and that the new discoveries awaiting us are worth the cost; however we must always insure adequate healthcare for all, including the poor people.

References available upon request