Statement of Research Interests

Hamed Valizadegan

November 2, 2012

In recent years, advances in technology have lead to the collection of large-scale patient data in the forms of Electronic Health Records (EHR). Analyzing EHR data is extremely important because it is the key to healthcare quality improvement, healthcare cost reduction, a better understanding of various types of diseases, and personalized medicine. However, learning from this data is very challenging because of their special nature: EHR data instances are traces of complex behaviors characterized by large number and types of temporal data (e.g. lab results, medications, text reports, etc.) sampled at irregular points, and represented by both structured and unstructured forms. Moreover, a majority of medical conditions are not explicitly annotated by the treating clinicians. This creates a need for additional annotation by clinical experts, which is a challenging task.

The benefits of analyzing medical data and the practical challenges in effective mining of this data have created a number of research and funding opportunities. An appropriate analysis of medical data demands not only a knowledge of machine learning and information retrieval, but also knowledge and experience in working with medical data. Machine learning is concerned with the design of algorithms to automatically learn models from the data. Information retrieval is concerned with the extraction of useful information from different forms of data such as text, relational databases, and the World Wide Web. My expertise in machine learning and information retrieval, along with the experience I acquired during my post-doctoral research fellowship on working with medical data, places me in a relatively unique position to develop new research methods for analyzing this data. Below, I describe my

- research background in machine learning and information retrieval and its relevancy to the biomedical research,
- current contributions to biomedical informatics, and
- future research directions in biomedical informatics.

Relevant Research Background:

Information Retrieval: The everincreasing biomedical literature largely contains information in free text format. Developing methods to summarize this type of textual data is important in retrieving and classifying documents in response to complex queries and to extract useful information such as protein-protein interactions, protein functions, associations between genes and different diseases, equivalent drug names, drug and treatment relationships, and general information about various diseases. I have research expertise in the relevant areas of information retrieval [1, 2, 3, 4].

Machine Learning: The enormous volume of medical data (both the number of patient records and the amount of information per patient record) requires the development of specialized learning algorithms that make efficient use of computational resources, minimize labeling cost, and that are able to handle high-dimensional data. My machine learning background provides me with the tools to develop mining algorithms that work under these constraints: I have experience in developing techniques that reduce the amount of human agent involvement [5, 1] (e.g. in labeling/annotation process) on the one hand, while on the other hand dealing with the critical computational issues of memory management [2, 3], running time [6], and the challenges of learning from high dimensionality data [7].
Medical Informatics: Machine learning from EHR data requires using the clinical data collected from past patients which are often not complete and ready to use. There may be a need for annotation by human experts. While diagnosing a patient case, the physician might need to review the entire electronic health record, a very time consuming and costly task. To make the process more efficient, we 1) introduced intelligent sampling in [8] to obtain more informative samples to evaluate the diagnosis models, 2) utilized auxiliary diagnosis information in [9, 10] to reduce the number of examples to construct diagnosis models, and 3) developed a probabilistic approach to obtain more accurate labels when multiple experts label patient cases differently [11].

We also developed temporal methods for learning classifiers from the time-series data in EHRs [12, 13]. In a related work, we utilized past medical records to identify unusual clinical care, such as the omission of an important lab test [14, 15]. Detecting such anomalies in patient management decisions could help increase the quality and decrease the costs of patient care.

Future Directions:
The EHR data collected in a medical center remain largely private to that institution, including their use for research under HIPAA compliance and IRB approval. To improve our understanding of this data and consequently to improve healthcare, we need to locally study the relationships between different medical variables in each medical center and then pass (via a two-way message passing system) the localized knowledge to a central repository that aggregates the knowledge globally. The design of a distributed learning framework for medical knowledge and inference is one future direction of my research.

EHR data consist of different types of clinical variables, each represented differently. For example, medication variables are usually represented in interval-based formats that specify the time interval during which a patient was taking each medication. Conversely, lab variables are usually numerical time series that specify the patient’s laboratory results over time. Moreover, there are different types of notes and reports in text format that indicate how the state of the patient changes over time. Most existing machine learning techniques, however, are designed to work on a feature vector representation of data; those that work on time series data (e.g., Hidden Markov Models or kernel methods) are difficult to apply on EHR data because EHR data are irregularly sampled over time. Consequently, adapting existing machine learning techniques to handle EHR data often sacrifices one or more aspects of such rich data (e.g., converting temporal data to a static representation). Designing a new learning framework capable of handling different types of clinical variables is a difficult task. Such a framework not only needs to handle different types of clinical variables efficiently and effectively, but it should have the ability to 1) discover the relationships between different types of clinical decisions and patient conditions, 2) incorporate domain knowledge into the learning process, and 3) perform learning/inference from distributed data bases while preserving appropriate data privacy. These are areas of research that I plan to pursue.

As mentioned earlier, in building decision support systems, it is often necessary to annotate medical records in order to learn from them. Annotation is usually a time-consuming task. Often multiple physicians are recruited to annotate a subset of patient cases. The selection of the most informative cases depends on how they will be used, such as whether the annotation is for model building (e.g., alert-rule construction), model evaluation (e.g., alert-rule evaluation), or other objectives. The selection of the most informative patient records to annotate and the assignment of patient records to different experts for annotation are areas that I plan to pursue in future research.
References


