Preface

The importance of Information Technology in healthcare cannot be overestimated in today’s world. With the rapid development of Information Technology (IT) there is an ever-growing emphasis on the use of IT in the research and practice of healthcare. The research in health information technology has been motivated by several developments, including electronic medical records, emphasis on personalized medicine and research on genomic data, and development of computational techniques for finding patterns, to name a few. Along with these new privacy concerns for electronic medical records and their access to third parties has become an important issue. In this article we discuss some of the recent research on health information technology within the context of the following areas: pattern recognition in medicine, privacy and security issues in healthcare, and clinical decision support.

PATTERN RECOGNITION IN MEDICINE

Pattern recognition algorithms are powerful techniques that can identify hidden patterns in data that can be beneficial for medical diagnosis, treatment management, epidemiological studies, finding longitudinal patterns, information retrieval in medical literature, and integration of different types of data. There is vast potential for applying pattern recognition techniques including algorithms from data mining, machine learning, statistical analysis, and matrix-based methods such as singular value decomposition and spectral analysis in healthcare data. In this section we first provide an overview of some of the common methods that have been developed in the recent literature and then describe a few articles in more details. In our discussion we recognize the overlap between machine learning and data mining methods but differentiate the two fields by the fact that data mining has a more explicit focus on dealing with large databases. Statistical methods are hypothetico-deductive where the main focus is to test and either validate or refute, through rigorous mathematical treatment, a given hypothesis (Hand 1998). Data mining is useful in exploratory data analysis where patterns are found using inductive methods (see Han et al 2006, Tan et al 2006, Dunham 2000 for more details).

Association Rule Mining

Traditionally, data mining methods have been categorized into association rule mining, classification, clustering, and outlier or anomaly detection. Association rules have their roots in market basket analysis where the goal is to find frequently co-occurring items in retail store transactions. Since different transactions (market baskets) typically belong to different customers it is assumed that the transactions...
are independent of each other. In some applications, such as co-location analyses such assumption is not valid and hence some modifications to the algorithms for association rule mining are needed. Generating association rules using the popular *a-priori* algorithm requires multiple database scans and hence more efficient algorithms such as FP-tree (Han et al 2000) have been proposed. Association rules represent local patterns between items, terms, diseases, patients etc.

**Clustering**

Clustering is an exploratory technique used in finding homogeneous groups is clustering, which can be useful in finding global characteristics such as the main groupings in datasets. There are several challenges in finding clusters in datasets. Since the goal of clustering is to group similar data points in one group, an appropriate distance measure is needed. The popular distance measure of the Euclidean distance assumes an underlying flat-space geometric representation of the data. Other flat-space geometry measures such Mahalanobis distance (Mahalanobus 1936), Kullback-Leibler distance (Kullback et al 1951), are also used. There are many clustering algorithms, the most popular being the k-means clustering algorithm. However, others such as hierarchical and density-based clustering algorithms are also commonly used. Methods such as k-means and k-medoids tend to discover globular clusters whereas density-based clustering algorithms can discover clusters with arbitrary shapes including those with high aspect ratios.

**Dimensionality Reduction**

A fundamental challenge in data mining, particularly in clustering, is the high dimensionality of the data. In a high-dimensional space the data gets scattered which makes it difficult to find homogeneous clusters. Dimensionality reduction techniques such as Principal Components Analysis (PCA) (Joliffe 2002) can be used to project the data into a low dimensional subspace as a pre-processing step. The low dimensional subspace is spanned by a set of eigenvectors that correspond to the largest eigenvalues of the covariance matrix of the data. Since the covariance matrix is symmetric the eigenvectors are orthogonal. Hence projecting the original data into the new basis vectors (set of selected eigenvectors) also decorrelates the dimensions. In a recent study (Joshi et al 2010) PCA was applied to glomerular filtration rates of the kidneys for 110 patients tracked over a period of twenty-four months after kidney transplantation with the goal of identifying homogeneous clusters that would explain why some of the transplanted kidneys failed while others did not. The data in the original twenty-four dimensional space did not exhibit any consistent patterns while PCA reduced the dataset to a single dimension. Subsequent analysis showed statistically significant differences between patient clusters with implications such as possible intervention strategies that can be deployed more effective care.

Principal components analysis is a linear transformation of the data that may not preserve nonlinear relationships. Manifold techniques such as multi-dimensional scaling (MDS) (Lee at al 2003), locally linear embedding (LLE) (Saul et al 2003), and ISOMAP (Tenenbaum et al 2000) try to discover low dimensional manifolds for high-dimensional data. However, MDS tries to reconstruct the nonlinear relationships using Euclidean distance and may not be topology preserving. ISOMAP builds a graph using K nearest neighbors and looks for local properties in those neighborhoods. ISOMAP assumes that there are no holes in the data. It is also sensitive to noise. LLE, like ISOMAP also uses nonlinear embedding. However, LLE may not be able to handle high-dimensional data.
Classification

Classification is a supervised technique for pattern recognition where the goal is to classify an instance into one of several discrete pre-defined classes. A classification model is built from a set of existing instances whose class values are already known. A simplified example is to classify a patient into one of two classes (healthy or sick) from a set of observable symptoms. The classification model is created from a training set and its accuracy is estimated using a set of test set. In addition to the overall accuracy other measures of goodness of the classification model are used. Examples of classification algorithms include decision trees, support vector machines, neural networks, and Bayesian classifiers such as the naïve Bayes classifier and Bayesian networks. Examples of classification algorithms in healthcare include Ludwig et al (2010) and Yap et al (2010).

Mining Medical Data Streams

Biomedical signals generated by monitoring devices such as Electrocardiogram (ECG), Electroencephalogram (EEG), and Phonocardiogram (PCG) abound in today’s healthcare industry (e.g., Cowley 2006, Lu et al 2005). Systematic analysis of such data can help monitor and detect early signs of critical conditions and possible interventions. However, the data generated by such devices is typically very large and potentially unbounded. For example, an ECG can generate several million data points in a matter of minutes (Chaovalit 2010). Analyzing such data streams is a challenging task that requires efficient algorithms that can detect time-dependent changes efficiently and accurately. Chaovalit (2010) describes an algorithm called POD-Clus for clustering high-speed data streams using a linear time algorithm that can create and update clusters over a period of time. More specifically, the method has the following characteristics: it scans the data only once which is required given the high-speed, high-volume intake of data streams, it creates a compact model (cluster synopsis) of the data and updates it parameters incrementally, it uses a small and fixed amount of memory, it detects outliers, and it can handle cluster evolutions. There are several types of cluster evolutions that are handled by the method: migration of data points from one cluster to another, creation of new clusters either as a result of cluster spilt or mergers, emergence of new clusters, and dissolution of existing clusters. Two algorithms are proposed to handle cases with and without cluster evolutions respectively. If cluster evolution is not expected the algorithm starts with an initial cluster formation, computes the mean and standard deviation (or covariance matrix if the data is multi-dimensional), and updates these measures as new data points are generated. Obviously, the method assumes that the data is follows a Gaussian distribution, which is realistic for large data sets (Thompson et al 1990).

When cluster evolution is expected a number of additional steps are necessary. As new data points arrive, the algorithm checks for outliers that are not within the standard deviations from the centroids of any existing clusters. Outliers may form a new cluster if they are homogeneous. The method monitors the sparsity of the clusters using a threshold and uses a chi-square test to determine if the cluster needs to be split. Clusters can dissolve when a sufficient number of data points migrate to other clusters or groups. Clusters can also be merged when the overlapping region exceeds a predetermined threshold. The method was extensively tested against previous data stream clustering algorithms such as E-Stream (Udommanetanakit et al 2007) as with ground truth validation.
Information Retrieval

The practice of evidence-based medicine requires physicians to retrieve and analyze data from published research, which helps provide the latest treatment to their patients. An example is randomized clinical trials that provide evidence of the effectiveness of drugs on specific patient populations. There is a huge amount of medical information in the form of text documents including research papers in online databases, physician notes, discharge summaries, Web portals, etc. There is a vital need for effective search and retrieval of information from these sources as well as for the organization of these text corpuses into homogeneous groups or clusters. This remains a challenge even with the advanced technology found in today’s world. One important characteristic needed for information search and retrieval tools in this context is the ability to remove obsolete documents and narrow the search down to the set of newer documents or respond to queries with specific time windows. These methods should be incremental and dynamic in order to avoid expensive re-computation every time changes are made. A dynamic document clustering method is described in Ramesh (2011) that creates singular value decomposition (SVD) of a term-document matrix (TDM) that contains term weights as products of term and inverse document frequencies normalized by the document size. The document clustering is done using the left singular vectors of the TDM corresponding to the largest eigenvalues. Since only a few singular values are selected a truncated SVD is created consisting of the submatrices corresponding to the original left and right singular vector matrices and the submatrix of the original singular value matrix with only the selected subset of singular values. The challenge is to incrementally update the truncated SVD when new terms and/or documents are added to the existing text corpus. Re-computing the SVD for the entire TDM each time is computationally expensive. However, fold-in methods have been suggested for incremental updates of the singular vector matrices of the TDM. A new fold-in method and subsequent clustering are suggested in Ramesh (2011). When new documents arrive, the TDM for the newly added documents is created and the new TDM is projected into the truncated SVD space. In addition to folding-in the method also describes a method called “windowing” that updates the truncated SVD when users select a specific time window for document clustering or retrieval. The method is validated using medical abstracts from Pubmed.

PRIVACY AND SECURITY

The Health Insurance Portability and Accountability Act (HIPAA), enforced by the Office for Civil Rights, is the national standard for security of electronic health information and the protection of identifiable information of patients. With the growing digitization of health records the issues of patient privacy and security of electronic medical records are becoming increasingly important. In this section we first discuss the privacy issues in healthcare data and then the issues related to security.

Privacy

Research on health information technology requires collection of data from a variety of sources including government agencies, healthcare providers, insurance companies, etc. The sensitive nature of healthcare data and the social and legal implications of the disclosure of such data require that such data be protected against any possible violation of the privacy of individuals. However, it is also important to
share such data with researchers to advance the state of the art in medical research. It has been shown that publicly available datasets that remove all identifying information about individuals can be compromised using linkage attacks (Samarati et al 1998). This presents an interesting dichotomy that requires the dual objectives of having to share the data in its original or some perturbed or transformed version such that both data utility and privacy will be preserved. The area of privacy preserving data mining offers solutions to this dichotomy by providing perturbation methods that hide private information while preserving patterns in the data.

The research on privacy preserving data mining can be broadly categorized into the following groups: data anonymization methods, data perturbation methods, cryptographic methods, methods dealing with specialized data types such as genomic data, image data, spatial data and protection of location privacy, incremental methods for privacy preserving data mining, privacy preserving data imputation, and privacy preserving data integration. In this article we will provide an overview of the latest work in this area. Interested readers are referred to books such as Vaidya et al (2006) and Aggarwal et al (2008) for a more detailed treatment of this area. Data perturbation methods distort the data while preserving the underlying data distributions. Additive data perturbation methods were among the first data perturbation methods suggested in the literature. These methods add random noise to the data while preserving the data distributions (Agrawal et al 2001). However, such methods were shown to be vulnerable to attacks based on correlations (Kargupta et al 2003). Subsequently, random projection-based methods were suggested that would reduce the dimensions of the original data my multiplying it with a random matrix (Liu et al 2006). While the work described in Liu et al (2006) provided some privacy there was no provable worst-case privacy guaranty, which was achieved in a subsequent work described in Mukherjee at al (2008). The data perturbation and transformation based approaches suggested throughout the privacy preserving data mining literature have a fundamental problem when it comes to healthcare data in that the data utility is severely compromised in most cases. A more recent research describes a perturbation method for preventing 3D reconstruction of CT axial images used by radiologists for the diagnosis of various diseases such as carotid aneurysm (Chen et al 2011). The method is based on edge detection of individual DICOM slices, creating a map for each slice, and adding random voxels as mask to each slice. The mask is smoothed by making the voxel intensities similar to those of the closest skin surface. The diagnostic information below the skin surface is fully preserved.

Another area of research in privacy preserving data mining focused on secure mining of data with distributed inputs from a number of data owners (sites). This branch of research, also known as secure multi-party computation, typically use cryptographic techniques to securely compute mathematical functions such that all parties can see the final results but not the individual data which is sensitive (see Verkios et al 2004 for a survey). Li (2010, 2011) described a method for privacy preserving clustering for gene expression microarrays on distributed data. Three different data partitioning scenarios were considered: horizontal, where each site has a different set of genes but all sites have the same set of experimental conditions, vertical, where each site has the same set of genes but different experimental conditions, and mixed, where each site has a different set of genes and a different set of experimental conditions. This method also reduces data transmission cost that can be prohibitively large for genomic data. Experimental results showed the feasibility of the approach with high degree of accuracy in PCA and clustering.
Security

Electronic health record systems (EHRs) can provide many benefits such as reducing cost, improving the quality of care, and empowering patients to maintain their own health-related information. The ability to record and maintain health-related information provides patients to have more control over their own information and increases their awareness about their health condition. However, electronic storage of data also brings increasing threats for inappropriate access to such data with devastating consequences. Government agencies across the world (such as HIPAA in the United States) have developed directives to safeguard patient privacy. This necessitates research and development of methods to secure such systems against unauthorized access to data.

While legislations control the privacy and electronic communications of EHRs personal health record systems (PHRs) such as Google Health (Google 2007) and Microsoft Health Vault (Microsoft 2007) are still outside the scope of such legislation (Ibraimi et al 2010). One problem with PHRs is that a patient cannot create fine-grained and dynamic access control policy on their health data where the access is all or nothing. Another concern is that the sensitive health information such as history, drug-use, and sexual orientation are stored in a third-party server without any authority over the access control mechanisms established by the third party host. The major issues related to PHR systems are the ability for patients to provide differential access control and guaranteed security against unauthorized access to such data.

Ibraimi et al (2010) describe a public key encryption method for securing personal health records. Patients might want to share their health information with their primary care physicians that can be achieved using a proxy re-encryption mechanism. Public key encryption schemes create a pair of private and public keys. Anyone can encrypt a message with the public key. However, the private key is needed to decrypt the encrypted data (the cyphertext). The identity-based proxy re-encryption scheme (Green et al 2007) extends public key encryption by allowing the patient (delegator) to create a proxy re-encryption key that can be used by a designated physician (delegatee) to re-encrypt the cyphertext and subsequently decrypt the original data using their own private key. Thus, the delegator does not need to share his/her own private key with anyone else. This scheme is useful is PHR systems because (1) the data is secure in a third party server because it is encrypted, (2) the patient can share his/her data with designated physicians, and (3) the patient can dynamically change access policy with out having to decrypt the data. One potential drawback of the identity-based proxy re-encryption is that the proxy (in this case the physician) can re-encrypt all of the patient data and get access to them. Hence, this mechanism is still all-or-nothing, which does not let the patient establish differential access controls. This limitation is addressed in Ibraimi et al (2010) by extending the identity based encryption scheme (Boneh et al 2001) to a type-and-identity based proxy re-encryption algorithm. In this method the patient can categorize the data into different types corresponding to different levels of sensitivity using one key-pair and delegate the decryption key of each type to a different healthcare professional thorough a proxy. This has the advantage that if the proxy and the delegatee are corrupted only one a subset of the data will be compromised. Also, the patient can control access to his/her data based on the level of sensitivity they associate with different subsets. It also has the advantage of giving the patient more access control over their private data, which should enhance their trust in the system. In this mechanism the proxy maintains the list of re-encryption keys that should not be made public. However, the proxy cannot decrypt any of the data by themselves because they do not have access to the private keys of the delegates.
CLINICAL DECISION SUPPORT

Computational methods have long been studied for developing tools and techniques to support decision making in healthcare. However, recent advancement in information technology has generated tremendous opportunities sharing, designing, and integrating information among patients and healthcare professionals in the forms of record sharing, health information networks, and protocols (Feishi et al 2003). In this section we describe a few studies to illustrate the role of information technology in the practice of medicine.

Identifying, Screening, and Testing At-Risk Groups

Zhao et al (2010) describe a two-step branch and bound algorithm to help clinical managers make decisions to identify at-risk population groups and develop strategies for screening and treatment of Chlamydia Trachomatis (CT) and Neisseria Gonorrhoeae (GC), two of the most common sexually transmitted diseases in the United States. Screening for CT and GC is expensive and budget restrictions limit the extent to which clinics can perform expensive screening procedures. In order to be cost effective the Center for Disease Control (CDC 2010) and the U.S. Preventive Services Task Force (USPSTF 2005) have identified the subpopulation of women who are 25 or under for such screening. Although restricting diagnostic tests to a subpopulation reduces the cost of testing, it does not solve the problem of choosing between expensive but sophisticated testing options versus testing wider groups of the population with less expensive and less effective tests. Bundled tests as opposed to single-pathogen tests have also been offered as potential solutions. Zhao et al. (2010) proposed an optimization method to solve the resource allocation problem under cost constraints for two sexually transmitted diseases. The solution is a two-step branch and bound algorithm, and subsequently a dynamic programming algorithm that provides the globally optimal solution that is also computationally efficient. The cubic binary model proposed by the authors is applicable for solving resource allocation problems to manage budgets under a fixed number of groups, screening assays, and treatment regimens for CT and GC.

Clinical errors are the cause of a significant number of deaths in hospitals in the United States (Ludwig et al 2010). Examples of such errors include adverse drug effects and improper transfusions, falls, burns, pressure ulcers, wrong-site surgery, and mistaken patient identities. Errors in intensive care units, operating rooms and emergency department have the most adverse consequences (IM 1999). Clinical decision support systems (Kaplan 2001, Metaxiotis et al 2000) can be decision aids for medical practitioners in various ways such as matching specific patients using a computerized knowledge base, patient-specific assessments and recommendations, pattern recognition, and data visualization. Fritz et al (2010) identify three main areas as critical for the design of a clinical decision support system: data entry and decision algorithms, human factor and usability for data acquisition and data requests, and presentation of the output. Machine learning has been used in clinical decision support systems in mainly two areas: (1) diseases recognition from input sources such as cardiograms, Computed Axial Tomography, and Magnetic Resonance Imaging; (2) bioinformatics such as DNA analysis and interpretation of gene expression data (Li 2011). Ludwig et al (2010) compare machine learning techniques using data from adult intensive care units to predict two outcome measures in two different data sets: death and hours of ventilation needed by the patient. Eight classification algorithms including Bayesian networks, nearest neighbor classifier, decision tree, naïve Bayes classifier, and support vector machines were compared with two new algorithms automatic genetic programming and neural network with weight-elimination.
Implantation of Coronary Stents

Implantation of coronary stents is a regular non-clinical procedure applied to patients with coronary heart disease. Stents are metallic cylinders that can be classified as slotted tubes, coil, or mesh (Xiao et al., 2007). Computer simulations based on the Finite Element method have been developed to study the structural behavior of stents in terms of stresses and strains developed when the device is implanted. The finite element method is used to generate a description of a real structure in terms of discrete finite elements and develop a mathematical model to describe the behavior of the structure. Computational finite element methods generate an approximate numerical solution and the goal is to minimize the modeling error. The better the selected mathematical model describe the physical properties of the structure being analyzed the less the error. A few additional factors such as constraints and loads, coarseness of the mesh that discretizes the structure, mesh dimensions, the number of degrees of freedom, and the simplification of the geometry can affect the modeling error and also runtime accuracy.

Gigliardi (2010) describes a method for optimizing the mesh dimensions and simplification of the geometry using seven different mesh dimensions and three different geometries of a number of commercial stents. Convergence tests were applied to select the best element dimensions for each geometry in order to minimize analysis time. The Newton-Raphson method was used minimize the residual force and moments calculated from the stiffness matrix in each iteration. This work extends previous related research in several ways: by varying element dimensions (as opposed to Xia et al 2007, Feng et al 2008) and performing convergence tests on a larger range of element sizes (as compared to Narracott et al 2007). This paper used the finite element ANSYS software and the author suggests verifying the results with other software packages for finite element analysis. Furthermore balloon expandable stents and contacts between the stent and atherosclerotic plaque should be included in future studies. However, Gigliardi (2010) provides guidelines for correct mesh choice for computational analysis of bare coronary stents.

Aryya Gangopadhyay
University of Maryland Baltimore County, USA

REFERENCES


Han, J. Pei J., & Yon, Y. (2000). Mining Frequent Patterns Without Candidate Generation. In Proceedings of the ACM-SIGMOD International Conference of Data, 1-12, Dallas, Texas.


