CSE 300: Topics in Biomedical Informatics
Data Mining and its applications and usage in medicine
Radhika
Spring 2008
# Table of Contents

Introduction 3  
Data Mining 5  
  - Concept Description 5  
  - Associations 7  
  - Classification and Prediction 9  
  - Clustering 10  
Applications and Usage of Data Mining in Medicine 11  
  - HealthMiner 13  
  - Medical Data Mining on the internet: Research on a cancer Information System 15  
  - Rough Set Data Mining of Diabetes Mellitus Data 18  
  - Heart Disease Prediction with Association Rules 19  
Conclusion 20  
References 21
1. Introduction

Data mining is a knowledge discovery technique widely used in many domains including finance, commerce, geological surveys, weather pattern prediction and telecommunications. The major reason that data mining has attracted attention in the past few years is due to the wide availability of data in electronic form and the need for turning such data into useful information and knowledge. Sophisticated data mining tools have been developed over the past decade with the progress in database and information technologies.

Database and information technologies have been evolving since the 1960s from the primitive file processing systems to more powerful database systems. The development in databases since 1970s has progressed from the early hierarchical and network database systems to relational database systems, data modeling tools with entity-relationship model, and indexing and data organization techniques with B+ trees hashing, query languages such as SQL and the “Online transaction processing” (OTLP). The sophisticated data models of the 1980s include extended-relational, object-oriented, object-relational and deductive models. The concepts of data warehousing and data mining have existed from the late 1980s together with the Online Analytical Processing (OLAP) technology. In the 1990s, web-based database systems including XML-based database systems and systems that provide web based data mining were developed. Heterogeneous database systems and the Internet based global information systems such as the World Wide Web now play a major role in information retrieval and processing. Computer hardware technology has also progressed by leaps and bounds in the past three decades, leading to large supplies of powerful and affordable computers and data collection and storage devices. This has provided a boost to the information industry with huge number of databases and information repositories available for transaction management, information retrieval and data analysis.

A database architecture that has recently emerged with the intention of facilitating data mining is the data warehouse, defined by Han and Kambler[1] as “a repository of multiple heterogeneous data sources, organized under a unified schema at a single site in order to facilitate management and decision making”. Data warehouse technology includes data cleaning, data integration and OLAP which enables data to be viewed in an aggregated from in a multidimensional manner. Hence we see that data mining is indeed very relevant in today’s context with specific technology being developed to facilitate its usage.

Data mining is a synonym for another popularly used term “Knowledge Discovery in Databases” or KDD. Knowledge discovery is an iterative process consisting of data cleaning, to remove noisy and inconsistent data, data integration, to combine multiple heterogeneous or homogeneous data sources, data selection, to consider only data relevant to the task and data transformation where data is transformed into forms appropriate for mining functions such as aggregation or summarization. Then data mining algorithms are employed to extract interesting and meaningful patterns from the data and present the knowledge to the domain expert in an informative manner.

Based on this, it is intuitive that the typical data mining system has a multi-tiered architecture as shown in Fig 1.
Data from a set of databases, data warehouses, spreadsheets or other information repositories form the first tier. Data cleaning and integration techniques maybe performed on the data to make it more tuned for the user queries. A database or data warehouse server is then responsible for fetching the relevant data from the database based on the user’s mining request. A knowledge base supports the data mining engine that processes the user queries. This is the domain knowledge that will guide the search or evaluate the resulting patterns for knowledge. It can include concept hierarchies which are used to organize attributes and attribute values into different levels of abstraction. Domain knowledge can also include additional interestingness constraints and threshold values as well as metadata describing the data from multiple heterogeneous sources. The data mining or OLAP engine consists of a set of modules which contain the algorithms for different types of mining techniques such as association rule mining, classification and prediction or clustering. The front end of the system will contain the pattern evaluation module and the graphical user interface which will represent the mined data in easily to understand visualized forms such as graphs and figures.

Typical data mining functionalities can be classified into two categories: descriptive and predictive. Descriptive techniques mine for relationships between the different attributes and observations in the data. In this manner, they characterize the general properties in the data using as few variables as possible. Predictive technique use inference on given attribute values to perform classification of data into predefined classes or group given data together to similar classes. In most cases, the user does not know what patterns are interesting and what the outcome of a search should conclude. Hence, a data mining system should be able to mine multiple kinds of patterns to accommodate different user expectations or applications. Furthermore, these systems should be able to discover patterns at various levels of abstractions and allow users to specify hints to guide the search for interesting patterns. Since not all patterns found in the database maybe be supported by a large percentage of the data, it is also important to specify a measure of certainty of the patterns discovered, ie. what percentage of support and confidence
from the actual data can we say that the mined pattern is meaningful. The general data mining functionalities include concept or class description, association analysis, classification and prediction, cluster and outlier analysis. These techniques will be discussed in detail in the following sections.

The applications of data mining techniques exist in various domains. The interest in systems for autonomous decisions in the medical domain is growing as large amounts of records in electronic forms are becoming available. However, in the medical domain, data mining is a relatively new field due to certain challenges that are associated with this domain. Not only are the datasets involved large and complex, with heterogeneous types of data that include clinical and lab recordings and images such as MRI scans, medical data is also hierarchical, time varying and of varying quality. Also, there exists a need for substantial medical knowledge base that requires collaboration between the data miner and health professionals for useful information to be extracted. Subsequent sections describe the problems, its associated solutions and provide successful examples of data mining techniques in medical applications.

2. Data mining

Data mining has been defined as the process of discovery of interesting, meaningful and actionable patterns hidden in large amounts of data. Data warehousing and OLAP was mentioned in the previous section. So how does data warehousing and OLAP relate to data mining? Information processing, based on queries, can find useful information. However, answers to such queries reflect the information that is directly stored in the database or at most, computable by aggregate functions. They do not reflect patterns or hidden interestingness in the underlying data. Since OLAP systems can present the general description of data in data warehouses, OLAP functions are for user-specified summary and comparison. These are the basics of the data mining functionalities, which on a larger scale, include, under the broad categorizations of descriptive and predictive data mining, association classification, prediction, time-series analysis and other data analysis tasks.

2.1 Concept Description

The simplest form of descriptive data mining is concept description. A concept refers to a collection of data such as graduate_students or patients_with_high_blood_pressure. Concept description generates description for characterization and comparison of data. Given the large amounts of data stored in databases, it is useful to be able to describe the characteristics in the most concise manner that cover most of all the data. Allowing generalizations of data at each level facilitates users examining the data at different levels of abstraction. Data generalization is a process that abstracts a large-set of task relevant data from a low conceptual level to higher conceptual levels. This can be done using two approaches.

In the first approach, OLAP operations can be performed on the data in the data warehouse, which may be stored in relational manner or in a multidimensional data cube. The data cube enables the access of summarized data in a faster and more sophisticated manner. An example of a data cube is shown in Fig 2 below. Example of OLAP operations on data cube include roll-up and drill-down, which allow user to view the data at different levels of summarization as shown in the figure. For example, the user may drill down on the data
summarized per quarter to data by month to analyze the data patterns in more detail. Similarly, he may want to roll up on cities to view data at a higher level of abstraction, countries.

Figure 2: Multidimensional data cube commonly used for data warehousing.

The second approach for data generalization and summarization based characterization is the attributed oriented induction approach (AOI). The data cube approach can be considered as a data-warehouse based, pre computation-oriented, materialized view approach. It performs offline aggregation before a user query is posted. The AOI method, however processes data online when the query has been submitted. The general idea of AOI is to first collect the task-relevant data using a relational database query and then performing generalization based on the examination of the number of distinct values for each attribute in the set of data. The generalization is then performed by either attribute removal or attribute generalization. Aggregation is then performed by merging the results. This reduces the size of the general data set. The aggregation is then mapped to visualized tools that enable to user to seek patterns. For attributes that need a large number of distinct values to cover the set of data given, attribute removal is performed. For example, to mine patient records for the characteristics that predict heart failure, the name of the patient or the patient health id is irrelevant and consists of an infinite set with no generalization able to be performed. Hence, the attribute name of ID will be removed. Since gender might provide an insight and there are only two values, make or female, we can retain the attribute. Body Mass Index(BMI), is an important indictor. However, actual values may be too many. Hence we can perform attribute generalization on these values and classify them to very low, low, medium, high or very high. In this manner, we can modify every attribute in the data set and present the final dataset in a more concise manner where patterns between the different attributes and the final outcome can be easily predicted.
Since attribute generalization and removal requires prior knowledge as to which the attributes are in fact relevant to the task required, we may not be completely sure of which attribute to remove or generalize. Also, such intuition requires domain knowledge or manual discrimination of attributes. To facilitate this task and to automate it, attribute relevance analysis can be performed. Since this method mathematically chooses the more important attributes over the less prominent ones, it can be automated with no effort required by the domain expert. There have been many studies in machine learning, statistics, fuzzy and rough set theories on attribute relevance analysis. The idea is to basically compute the information gain by having each of the attributes that are present in the data describe the data individually. For example, if there is an attribute in the data whose values in a certain range can be directly linked to heart failure, we can remove all other attributes from the dataset and examine the data using only the relevant attribute. Measures to quantify the relevance of an attribute with respect to the given class or concept include information gain, Gini Index, uncertainty and correlation coefficients. As information gain is a relevance measure that is used in medical data mining, it is described next.

How does information gain quantify the relevance of an attribute? Let $S$ be a sample from the data set, which consist of tuples who class labels are known. Suppose there are $m$ class labels. In our earlier example, there are only 2 classes for the presence of heart failure, “yes” and “no”. So $m$ is two. An arbitrary sample belongs to class $C_i$ with the probability of $s_i/s$ where $s_i$ is the number of tuples with class labels $s_i$ and $s$ is the total number of samples in the set. The expected information needed to classify the given sample is $I(s_1,s_2,\ldots,sm) = -\sum_{i=1}^{m} \frac{s_i}{s} \log \frac{s_i}{s}$. An attribute $A$ with values $\{a_1, a_2, \ldots, a_v\}$ can be used to partition $S$ into the subsets $\{S_1, S_2, \ldots, S_v\}$, where $S_j$ contains the samples in $S$ that have the value $a_j$. Let $S_j$ contain the $S_{ij}$ samples of class $C_i$. The expected information based on this partition by $A$ is known as entropy of $A$. It is given as $E(A) = \sum_{j=1}^{m} \frac{S_{ij} + s_2 j + \ldots + Sm_j}{s} I(S_1 j, \ldots, S_m j).$

The information gain is then $G(A) = I(s_1,s_2,\ldots,sm) - E(A).$ This means that we would select the attribute with the least entropy each time. Intuitively, entropy can be seen as the number of bits needed to classify that data. The lesser the number of bits needed, the more concisely the dataset can be represented. By computing the information gain for each attribute, we can hence have a ranking of the importance of each attribute to the user query. With these measures, we can characterize the dataset in as few attributes as possible for comparison and the mining of patterns.

2.2 Associations

Association rule mining is another type of descriptive data mining which is more complex than concept description. It is widely used in finding patterns of association among the attributes or variables (e.g. demographics) and observations (e.g. patients). Unlike the methods to be described in the next section on predictive data mining, the data have not been segmented by the analyst into sets of particular interest. Associations among variables are more popular in datasets where there are many more variables than observations. An example of such a dataset is gene expression where there are tens of thousands of variables but only a few hundred observations. A typical example of association rule mining is market basket analysis. The process analyzes customer buying patterns by finding the associations between the different items that the buyer

7
places in his shopping basket. For example, in a transactional database of customer records, if milk and bread are items commonly purchased together, the department store might strategically place these two items together, or the store might want to place them in different aisles, next to sales items less commonly bought so that the shopper can be aware of the discount prices on these items. Such strategies can help retailers to increase their sales by performing selective marketing and plan their shelf space. The associations among the variables can be seen as rules.

For example, if in a given dataset, we see that for every tuple that has the condition heart failure, the condition high LDL (or bad cholesterol) and low HDL (or good cholesterol) are almost always recorded, we can form a rule of the type, \{High LDL, Low HDL\} $\rightarrow$ \{Heart Failure\}. Not all the tuples for which heart failure occurs might have a record for High LDL and Low HDL. Hence we need to augment the association rules with two rules of measure of interestingness, support and confidence. A support of 90% means that for 90% records under the analysis, both heart failure and High LDL and Low HDL occurred together. A confidence of 80% means that of all the records containing heart failure, 80% of them contained High LDL, Low HDL. Associations rules are of type \{High LDL, Low HDL\} $\rightarrow$ \{Heart Failure\}. [support 90%, confidence 80%]. A dataset may have millions of association rules generated from it. However, only a subset that pass a minimum value for the support and confidence are considered as interesting.

Depending on whether we are finding associations among variables or among observations, the methods of discovery vary. The singular value decomposition (SVD) is a most common and effective method for associating variables. The SVD can be described in many mathematical variations including principle components. Principle components look for linear combinations of the variables and these linear combinations are then the used to describe the original data. The linear combinations are selected such that they best describe the spread in the original data with minimum overlap as possible. The variance of each linear combination is thus maximized to cover the entire database. Typically, solutions are sought such that the number of principle components is less than the number of variables in the database, indicating that the procedure has identified similar variables and placed them in the same principle component. The results are particularly interesting when a small number of principle components describe a large amount of variance in the database.

Several algorithms have been developed for finding rules among observations. One of the most commonly used among these is the Apriori algorithm which was developed for market basket analysis. Apriori is an influential algorithm for mining frequent itemsets for Boolean association rules. The name of this algorithm is based on the fact that the algorithm uses prior knowledge of frequent itemset properties (the minimum support level) to do a level wise search, where k-itemsets are used to explore k+1th itemsets. The algorithm begins with the most frequently occurring 1-itemsets. These are the attributes that occur frequently individually. The count for each of the attributes occurrence is recorded and the attributes that do not pass a minimum threshold for support are removed. The algorithm then proceeds to do a join of the 1-itemsets to produce 2-itemsets. Then the occurrence of the 2 itemsets are recorded and those that do not occur too frequently, as specified by the minimum support are pruned again. This way, the algorithm proceeds to consider 3 itemsets, 4 itemsets and so on until none of the joined attributes in the kth itemset passes the threshold value. In this manner, the attributes that occur together frequently are found. Then the domain expert can now process these rules to mine information from them.
2.3 Classification and Prediction

Classification is the process of finding a set of models that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown [1]. Classification is a technique that falls under predictive data mining and is the most popular of the data mining techniques, especially in the medical domain. It encompasses prediction of values given the values of other variables. Prediction requires Apriori the set of class labels that we want to predict and values for the predicting variables. Although many methods attempt to identify the more useful predictor variables, no method can be used to find predictors that are not already given in the data. Prediction techniques, as association rule techniques, vary based on whether the response expected is numeric or categorical. Usually, the dataset is divided in a fraction, commonly two-thirds for training and one thirds for testing. The derived model is based on data used to train the model. The accuracy of the model is then tested based on the testing set and the model is then improved if the accuracy is not good enough for the purpose. In clinical decision support systems, prediction plays a very vital role and the required accuracy of the system is high.

Regression is a common technique used when predicting numerical values. The oldest and the most primitive method is least squares regression. This method finds a linear function of predictor variables that minimize the sum of square of differences from the response variable. Let $y_i$ be the response variable and $x_i$ be the vector of predictor variables for observations $i=1..n$. Let $f$ be the linear function that estimates the response and $\theta$ be the vector of parameters in this function. Hence the estimated parameters, $\theta'$ is given by $\theta' = \arg \min \{ \sum (y_i - f(\theta, x_i))^2 \}$. A convenient choice for $f$ is a linear form of all the predictor variables.

The derived model for classification of categorical values may be represented in various forms such as IF-THEN rules, decision trees, mathematical formula or neural networks. A decision tree is a flow-chart like structure where each node represents a test on an attribute and a branch represents the value of the outcome. The leaves represent the classes which are pre-defined. Decision trees can be easily converted to IF-THEN rules. The ID3 algorithm is used to build a decision tree. Neural networks, when used for this purpose consists of a collection of neuron-like processing units with weighted connection between the neurons. It is usually used to mimic the biological processes and pattern recognition properties of biological systems. The input layer to the neurons is usually the attributes and functions in the hidden layer neurons convert the input values via some non-linear logic to the output layer. The output layer consists of the classes, similar to the leaves of the decision tree. Werbos algorithm is usually used in the construction of neural networks. This algorithm starts by randomly initializing and assigning weights in the neural networks. An observation from the database is then presented into the system. The error in computing the output by the network is then back propagated through the network and the weights are appropriately altered. The algorithm proceeds in the manner by altering the weights until there is not much error in predicting the outcome, specified by a predefined threshold value. A third method, support vector machine(SVM), is a recent addition to the data mining techniques category and is used in the medical domain. The technique seeks to predict class labels by separating the database into mutually exclusive regions. Each tuple is represented by a point in the k dimensional space where each dimension is an attribute. The SVM then identifies points that lie in prospective decision boundaries. Then they transform the database to points that can be linearly separated. When unknown data is presented to the SVM, it
predicts the class the data point belongs to by checking its proximity to the nearest class of points. Trees are by far the most popular choice due to their ease of understanding and interpretability. Classification and prediction may need to be preceded by attribute relevance analysis mentioned in previous sections.

2.4 Clustering

Unlike classification and prediction which analyze data objects labeled by classes, clustering analyzes objects without the prediction of class labels. Objects whose class labels are not known with but which exhibit a wide range of characteristics are grouped into one cluster. Class labels may be generated later. An unknown tuples characteristics can be predicted by comparing it to the objects it is similar with. The clusters are formed so that objects within a cluster have high intraclass similarity and low interclass similarity. Clustering can also aid taxonomy formation which is the organization of objects into a hierarchy of classes that group similar events together. Clustering can also be used to identify outliers or the odd data tuple. In certain medical cases, such a the identification of anomalous cells which may be indication of cancer, rare events are most interesting than regularly occurring ones. Some data points or tuples which maybe be cell data may be clustered away from the normal cells.

Common clustering techniques include partitioning methods, hierarchical clustering methods, density based methods, grid based methods and model based methods. Given a database of n objects and a number k, the partitioning method clusters the n objects into k clusters. The clusters are formed according to a similarity function, such as distance so that objects within a cluster are similar in terms of the database attributes. The most well-known and commonly used partitioning methods are k-means, k-medoids and their variations. In k-means, cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed as the center of gravity. The method starts off with choosing a random k which represent the mean of the k clusters. For each of the remaining objects, each object is assigned to the cluster it is most similar to. Once grouped, the mean of all clusters is calculated and the objects are re-assigned based on the means they are closest to. Dissimilarity measures vary according to the data type of the attribute in the dataset. For interval-scaled values, mean absolute deviation is used; for binary variables a dissimilarity matrix is calculated; simple matching coefficient is used for nominal variables and Euclidean or Manhattan distance used to ordinal variables.

Hierarchical clustering method is used in the medical domain to group the dataset into similar clusters for analysis. This method works by grouping data objects into a tree of clusters. Hierarchical clustering can be further classified into agglomerative and divisive clustering methods. Techniques for agglomerative clustering include the single linkage, complete linkage or the average linkage. The plot of the clusters is a diagram known as the dendrogram which shows the combination of the different clusters and the dissimilarity between each cluster. An example of a dendrogram is shown in Figure 3. Single link uses a minimum distance updating. After two clusters are joined, the new dissimilarities between the clusters are the minimum dissimilarities between observations in one cluster and the observations in the other cluster. This procedure computes a minimum spanning tree through the observations. The complete link method computes the maximum dissimilarity between observations in one cluster from another while the average link method computes the average dissimilarity between two clusters. These techniques will be re-visited in the examples of later sections.
3. Applications and Usage of Data Mining in Medicine

The amount of clinical data stored electronically is making it possible to carry out large scale studies that focus on the interaction between genotype, phenotype and diseases at a population level. Such studies have the extraordinary potential to determine the effectiveness of treatment and monitoring strategies, identify subpopulations at risk for disease, define the real variability in the natural history of disease and co morbidities, discover rational bases for targeting therapies to particular patients, and determine the incidence and contexts of unwanted health care outcomes. Matching patient responses(phenotype) with gene expression and known metabolic pathway relationships across large number of individuals may be the best hope for understanding the complex interplay between multiple genes and environment that underlies some of the most debilitating health problems[3]. However, despite the myriad of problems data mining in medicine can solve, the application of data mining to this domain is still relatively new due to obstacles that exist with this domain. There are several special characteristics that makes it difficult to analyze medical data in an automated fashion by traditional techniques.

Firstly, the common nature of medical data is such that there is high dimensionality associated with it. There are many data elements, each representing a dimension that varies in value, characterize an itemset of interest such as patient disease or specimen. Usually there could be 50 to 100 different types of data elements. It is important to thus carry out attribute relevance to choose only a subset of the elements since the like hood of patients sharing coincidental data is high and that may lead to incorrect conclusions.

Medical data is also heterogeneous in nature, including textual data, various types of images and discrete values using multiple scales. Data mining requires consistent data, meaning that large amounts of data need to converted to compatible representations. Linking data variables to patient characteristics is not straightforward. Unlike market basket analysis where the buyer patterns can be directly studied, medical observations commonly indicate a probability that a condition exists based on their support and confidence (which are alternatively known as sensitivity and specificity in this domain). A given feature may be consistent for the condition or
the condition may exist without the feature. It is not always as easy to predict medical conditions with the expected confidence and support as it is in other domains. The domain’s data records are also in some cases incomplete and with inconsistency. Patients with the same conditions may have substantially different types of timings of observations. Since most of it requires human data entry, the data is prone to inconsistency and noise.

Physician’s interpretation generally is the diagnosis in medical conditions. Interpretations of different individuals may differ or even conflict and are often expressed in text that need to be transformed to other forms before data can be mined from them. Even specialists from the same discipline cannot agree on unambiguous terms to be used in describing patient data. There are many synonyms for the same disease and the process of mining data is even more daunting as different grammatical constructs can be used to describe the relationships among medical conditions. The major reason for this ambiguity is because medical data has no canonical form. In mathematics, canonical form is used to represent all equivalent forms of the same concept, for example \( \frac{1}{2}, \frac{4}{8}, \frac{5}{10} \). In the medical domain, canonical form for even a simple idea such as “adenocarcinoma of colon, metastatic to liver” has no canonical form. It can be equivalently represented as “Colonic adenocarcinoma with metastasis to liver” or “Large intestinal adenocarcinoma with hepatic metastasis”. If there no equivalent ideas in medicine, then how can indexes and statistical tables constructed and data mining depends on these equivalent concepts. Usually, algorithms that perform data mining are programmed to work on mathematical form of the data in the form of formula and equations in the area of physical sciences. The conceptual structure of medicine on the other hand, consists of word descriptions and images, with very few constraints on vocabulary, the composition of images or the allowable relationships among basic concepts.

Because medical data is collected on human subjects, there is enormous ethical, legal and social traditions designed to prevent the abuse of patients and misuse on their data. There is an open question on data ownership. The data collected annually in North America and Europe alone is thousands of terabytes. These data are buried in heterogeneous database and scattered throughout the medical establishments. Questions arise on who owns these data, the patients, the physicians or even insurance providers. Another feature of medical data mining is the fear of lawsuits directed against physicians and other health-care providers. Controversies and confusions that relate to these issues exist that complicate aggregation and data analysis other than for individual patient care. The call for a national framework for the secondary use of health information is largely a recommendation that society, health care providers and government resolve these issues so that the patients can benefit from techniques such as data mining [4].

Despite the difficulties associated with the domain, there has been may efforts to study the potential of data mining in medicine. Most have examined targeted treatment procedures: cesarean delivery rate (270,774 women) [8], coronary artery bypass graft (CABG) surgery volume (267,089 procedures) [9], routine chemistry panel testing (438,180 people) [10], and patient care: cancer risk for non-aspirin NSAID users (172,057 individuals) [11], preoperative beta-blocker use and mortality and morbidity following CABG surgery (629,877 patients) [12], and incidence and mortality rate of acute (adult) respiratory distress syndrome (ARDS) (2,501,147 screened discharges) [13], to name a few. These studies have several factors in common: large sample size, clinical information source, and they support or build upon pre established hypotheses or defined research paradigms that use specific procedure or disease data. In the next few sections, we examine some of the latest studies done.
3.1 HealthMiner

To investigate the potential of searching large databases for novel insights, the HealthMiner tool is used on a 667,000 inpatient and outpatient digital records from the academic health center at the University of Virginia and its partner Virginia Commonwealth University Health System. These centers have established or are developing clinical data repositories which are large, usually relational databases that receive a variety of clinical and administrative data from primary electronic sources. The 667,000 deidentified patient data was mined using unsupervised techniques mentioned in earlier sections from IBM’s HealthMiner suite which consists of Association Analysis using CliniMiner, Predictive Analysis using decision rule induction methods and pattern discovery using THOTH. These methods are related in that they are all unsupervised “rule discovery” techniques but differ in the different outcomes expected of the rules they discover.

The Pattern Discovery/THOTH (which is an Egyptian word) begins by seeking to enumerate all patterns that occur in the dataset k times. Then the patterns are clustered based on distances computed from comparison of the lists of individual patients that match the patterns. The clusters are formed such that all patients with the same patterns are identified and the clusters also identify the relationships between the parameters shared by all the patients. From these clusters, the tool then constructs IF-THEN association rules as mentioned in previous sections, with scores for rule according to the support and confidence in the training data set.

The predictive analysis tool in HealthMiner learns or generates decision rules from the medical data such as “Diastolic Blood >100 AND Overweight IMPLIES High Risk of Heart Attack”. Since the method is unsupervised, which means that the output class label for these rules are not predefined by the user, each of the variables are considered as a class label and relationship of all other variables to the output are computed. 100,000 records were used for testing the rules and the rest for training. The rules formed by predictive analysis must predict the outcome of unknown data with a possibility which is significantly greater than chance.

CliniMiner has been revised for clinical data and contains security features to maintain patient privacy. Also, laboratory data values can be automatically converted to low, normal, and high ranges, while times and dates are converted to universal decimal year time (e.g. 2003.4752827) to facilitate time-stamping of clinical events and time series analysis, all which are known to be the difficult steps of data pre processing in the medical domain. In order to have the maximum potential to discover patterns in the dataset, no tuples were removed and hence the infrastructure that compiles with all Health Insurance Portability and Accountability (HIPAA) regulations was developed to protect personal health information. CliniMiner is used to find association rules in the dataset using the “Zeta Theory” approach which uses a probabilistic approach to find the rules. Since the other two tools were in their infancy stage of development, the CliniMiner tool was used for data cleaning for the other two methods.

The results from mining displays some interesting medical knowledge. Figure 4 shows the output from the three HealthMiner algorithms in tabular form. A number of already well-published medical correlations were found as rules from the dataset. The resultant rules are supported with measures similar to support and confidence with terms such as Predictive value, Sensitivity (which is similar to support), Specificity (which is similar to confidence), Accuracy
Predictive value represents the percentage that is correct when the rule is true. Prevalence indicates the percentage of diseased patients in the population.

### A. Representative FANO triplet data output

<table>
<thead>
<tr>
<th>Info.</th>
<th>Complexity</th>
<th>Saw</th>
<th>Expected</th>
<th>Event 1</th>
<th>Event 2</th>
<th>Event 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.4</td>
<td>3</td>
<td>1565</td>
<td>51.54</td>
<td>BUNFirst(Urea+Nitrogen:_Blood_serum:mg/dl:=&gt;0.7)</td>
<td>CREATFirst(CREATINE:_Blood_serum:mg/dl:=&gt;0.06)</td>
<td>Renal failure: =&gt;&gt;0.85</td>
</tr>
<tr>
<td>2.91</td>
<td>3</td>
<td>871</td>
<td>46.6</td>
<td>CREATFirst(CREATINE:_Blood_serum:mg/dl:=&gt;0.06)</td>
<td>Diabetes_Complicated: =&gt;&gt;0.76</td>
<td>Renal failure: =&gt;&gt;0.85</td>
</tr>
<tr>
<td>2.84</td>
<td>3</td>
<td>774</td>
<td>44.61</td>
<td>BUNFirst(Urea+Nitrogen:_Blood_serum:mg/dl:=&gt;0.7)</td>
<td>Diabetes_Complicated: =&gt;&gt;0.76</td>
<td>Renal failure: =&gt;&gt;0.85</td>
</tr>
</tbody>
</table>

### B. Representative Predictive Analysis output

Cardiac arrhythmias

<table>
<thead>
<tr>
<th>[Congestive heart failure &amp; age at diagnosis &gt; 7.500]</th>
<th>OR [Res_Diagnosis &amp; Res_Metabolic &lt; 6.500]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictive value</td>
<td>68.04%</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>52.46%</td>
</tr>
<tr>
<td>Specificity</td>
<td>95.78%</td>
</tr>
<tr>
<td>Accuracy</td>
<td>89.44%</td>
</tr>
<tr>
<td>Prevalence</td>
<td>14.62%</td>
</tr>
</tbody>
</table>

### C. Representative Pattern Discovery output

<table>
<thead>
<tr>
<th>% Cluster 30</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.830986</td>
</tr>
<tr>
<td>Gender =&gt; Male</td>
</tr>
<tr>
<td>0.741784</td>
</tr>
<tr>
<td>Gender =&gt; Male</td>
</tr>
</tbody>
</table>

**Figure 4: Representative output from the three HealthMiner Algorithms**

A subset of the rules that were previously medically unknown were correlations among alcohol abuse + drug abuse + AIDS, alcohol abuse + depression + drug abuse and fluid and electrolyte disorder + AIDS + other neurological conditions. Some of these are shown in the table below.

Three correlations were found among the rules that is previously unknown in the medical domain. A strong correlation (expected =21.07, saw=76) was observed between paralysis (=>−0.83), peptic ulcer disease (=>−0.82), and renal failure (=>−0.85). The correlation between paralysis (=>−0.83), peptic ulcer disease (=>−0.82), and rheumatoid arthritis (=>−0.87) was strong (expected=18.61, saw=48). A strong correlation (expected=55.42, saw=166) was observed between paralysis (=>−0.83), peptic ulcer disease (=>−0.82), and psychoses (=>−0.62). These correlations, though it may not make much sense to someone who does not have much medical background, could, once domain expert does analysis and further research to support the relations, could become the foundations for solutions for many medical conditions which at the moment have no remedy.

Similar results were achieved from the Predictive Analysis and Pattern Discovery methods. In the predictive analysis method, of the 120 rules examined, 73 (61%) were established, 18 (15%) were less well known, and 29 (24%) were unknown in the published biomedical domain. Again, three examples of previously unknown medical correlations are: The correlation between prescription famotidine, prescription midazolam >2.5, prescription magnesium, and any antibiotic >1.500 was strong (predictive value: 73%, sensitivity: 52.25%, prevalence:14.16%). Prescription of both omeprazole and magnesium was associated with liver disease (predictive value: 65.41%, sensitivity: 7.34%, prevalence: 5.69%). A strong correlation
between the prescription of albuterol and magnesium was associated with temporary and permanent tracheostomies (predictive value: 67.31%, sensitivity: 19.71%, prevalence: 11.82%).

For the Pattern Discovery technique, three medically known correlations that were mined included valvular disease + warfarin + cardiac arrhythmias; cardiac arrhythmias + valvular disease + echocardiogram + congestive heart failure, and congestive heart failure + valvular disease + hypertension. As for rules unknown in the medical domain that were generated, a strong correlation (fraction of consequent given antecedent: 0.71429) was found to exist between diabetes uncomplicated, physical therapy, head CT scans, and fluid and electrolyte disorder and the correlation between deficiency anemias, omeprazole, and hypertension was found to be strong (fraction of consequent given antecedent: 0.72).

It was also found that unsupervised data mining of large repositories is feasible. Both rules that were already well established in the medical domain as well as those that were lesser know and unknown were mined from the dataset. In future, filtering tools which will eliminate the previously established rules in the medical domain could be used to find only new knowledge. For filtering the data mining results through comparison with existing medical literature, tools such as Collexis is suggested. Given a set of documents, Collexis constructs a concept fingerprint of each document which is then stored in a catalog. The software then reads the collection of fingerprints and creates an associative concept space(ACS) which is stored in the database. The API browser visualizes the ACS models and is used to input a seed terms to find all related concepts or input concepts and find a related path through them(hypothesis testing) or retrieve reference that support related concepts. This is indeed a very useful tool that can be used to support rules, eliminate rules and provide a window into the research of lesser known rules.

3.2 Medical Data Mining on the internet: Research on a cancer Information System

The previous section was an application on patient records stored in a database. In this application, mining on the internet is considered. Data mining on the internet presents the problem of unstructured data. However, the dataset chosen for the study is from NCI’s CancerLit collection which has a structure and each document has a major topic or theme, making it easier that other internet data to mine patterns. The application presented here varies significantly from the previous one, in that in this example, data mining is used to organize and retrieve data in an efficient manner than mine patterns that exist in the underlying dataset.

The system is developed at the University of Arizona Artificial Intelligence lab. The algorithms developed have been implemented into several prototypes, including which focuses on medical information developed in cooperation with the National Cancer Institute(NCI) and the University of Illinois at Urbana-Champaign. Several government agencies such as National Institutes of Health(NIH), National Science foundation(NSF) and National Cancer Institute(NCI) have developed websites where we can mine extensive amounts of information. The techniques developed at the University of Arizona Artificial Intelligence lab address important issues in mining data from sources similar to these, such as high-dimensional data, over fitting, missing or incomplete data, visualization techniques as well as privacy. The automatic indexing feature in the concept space tool developed and the part-of-speech tagger represent attempts to reduce the high dimensionality of the data. The indexing system is shown in Figure 5. Over fitting in the neural networks model which is employed for clustering is reduced by appropriately controlling
several important parameters. The neural networks have been developed to cope to missing, incomplete or noisy data. For better visualization of the final results, two tools have been developed, the graphical concept space and the dynamic SOM. Privacy is not an issue since the CancerLit collection is available to public and does not contain any personal sensitive data.

Figure 5: Automatic indexing session where user entered “Breast cancer” Concept Space shown suggests terms related to Concept Space

The NCI is responsible for managing an extensive collection of cancer-related information. The information management includes being able to share the information in a timely, efficient and intuitive manner. NCI has instituted a series of small information sharing initiatives which are publicly available online. These include CancerNet, which provides information about cancer, screening, prevention and care as well as some summaries of clinical trials. CancerNet for Patients and the Public includes access to physician data query(PDQ) and related information on treatments, detection and prevention as well as directory of counselors. CancerNet for Health Professionals includes access to PDQ, related information on screening, treatment, prevention and genetics. There also exists CancerNet for Basic Researchers which includes access to a co-operative breast cancer tissue research database, an AIDS malignancy
bank database, a cooperative family registry for breast cancer studies and the Journal of the National Cancer Institute. Also, most importantly, for the techniques developed by the center, the CancerLit which is a comprehensive archival file of more than one million records describing 30 years of cancer research.

The NCI’s cancer information center has deployed and implemented several channels that will improve access to information available. PDQ is one such example. Another is the incorporation of Mosaic for timely dissemination of the information. A current challenge that exists is the diversity of users, with respect to their levels of subject expertise and understanding of the information sources. To enable a fast dissemination of information and to overcome other challenges faced, the University of Arizona Artificial Intelligence lab has proposed an architecture for the medical information system that will support a variety to use data mining tools that can address the scalability, modularity, manageability, diversity, efficiency, timeliness and cost-effective requirements. The architecture is shown in the figure below. The data mining techniques used on the system will focus on categorization and clustering and specific neural networks such as Hopfield neural networks and self organizing maps.

Concept Space is the main idea which are ideal for refining a broad search topic to a more refined one and for discovering relationships between document descriptors. This idea will be linked to other pre-existing indexing and summarization sources. A concept space is an automatically generated thesaurus that is based on terms extracted from documents. Users explore the concept space using neural networks. The concept spaces are linked via commonly held vocabulary and document descriptors, which serve as bridges between them. This allows users to explore a variety of concept spaces using mining tools that they are comfortable with.

The CancerLit concept space consist of names from the author field of documents, medical subject heading terms that are used in the keyword field of documents and descriptor terms. A threshold value is used to limit the descriptor terms to those that commonly occur. The significance of each word in different documents differs. Hence term frequency and inverse document frequency are used to represent the significance of the term. Term frequency represents how frequently the term occurs in the entire document while inverse term specificity allows terms to have different weights, relative to their importance. For example, a word or term in the title is given more weight than a word found in the experiments section of a document. Clustering analysis is then used to convert the word distances into a dissimilarity matrix to represent the relationship between the words. The retrieval of information is performed using a specialized neural network known as the Hopfield net. Knowledge and information can be stored in a single layered, interconnected neurons and weighted synapses and can be retrieved based on Hopfield network’s parallel relaxation and convergence methods. Each terms identified previously is treated as a neuron and the difference between terms is calculated and stored as the weights between the neurons. Given a user-supplied input sentence, the Hopfield algorithm activates their neighbors which terms strongly associated with them, combines the weights from all associated neighbors and repeats this process until the sentence has been discovered. The terms that are less linked to the terms in the user input sentence are given gradually decreasing weights until they fade off.

The Arizona Noun Phraser research is investigating the potential of combining traditional keyword and syntax with a semantic approach to improve the quality of the retrieved information. Clustering algorithms are again used to compare the dissimilarity between words and phrases as well as to identify relationships.

This example shows a different approach to data mining. The mined “data” is not knowledge but rather a faster way to associate terms in a document and retrieve relevant documents according to user query.

3.3 Rough Set Data Mining of Diabetes Mellitus Data

A major role of data mining in the medical domain is for creation of IF-THEN induction rules or decision trees in the area of classification and prediction. In this application, rough set theory is used to identify the most important attributes and to induce decision rules from them. Three methods are considered for the identification of the most important attributes, In first reduction is used by considering the elimination of every variable and the set induced by it. The accuracy is calculated for the approximated classes and the attributes that pass a minimum given threshold is retained. In the second method, attribute significance is calculated by considering the
relative decrease of the positive region after its removal and in the third case, inspired by the wrapper approach, classification accuracy is used for ranking attributes. The least significant attributes are then removed. As a final step, once the important attributes are identified, decision rules are formed according to k-nearest neighbors (which similar to the k-means algorithm described in the data mining section.

This method has been applied to a dataset from Second Department of Children’s Diseases, University Medical School in Bialystok, Poland. Diabetes Mellitus, Type 1 data on 107 patients, aged 5-22 suffering from insulin dependent diabetes for 2-13 years is used as the sample. Results from the mining concluded that the most important aspect are the following: age of disease diagnosis, criteria of metabolic balance and disease duration that influence the incidence of microalbuminuria in children suffering from diabetes of Type 1.

3.4 Heart Disease Prediction with Association Rules

Association rule mining is another important technique used in the medical domain. In the application presented, a variation of the Apriori algorithm, presented under the Data Mining section, is used to search constraints to reduce the number of unnecessary rules, search for rules on the training set and finally test their validity on the testing set. The medical significance of the evaluated rules is tested with support, confidence and lift. Search constraints and test set validity used by this application significantly reduces the number of rules produced with the remaining ones having a high predictive accuracy.

The dataset used contains 655 patient records with 113 attributes, combining numeric, categorical and image data. Diagnostic procedures performed by a clinician are also included. Risk factors are observed as age, race, gender and smoking habits attributes. The other attributes include weight, heart rate blood pressure and information on other conditions experienced by the patient, such as diabetes. The data set also has an important set of measurements that estimate the degree of disease in certain regions of heart. The rules are formed such that some predict the absence of heart disease while some predict the presence of heart disease. It is noted that an intuitive final screening should be done so that there is no overlap between these two sets of rules. Figures 7 and 8 show the rules derived from the dataset.

\[ e = 1 \]

\[ \text{L[1.0,2] SMOKE=n CHOL(200,250) \Rightarrow LCX(0,50) s=0.03, c=1.1, l=1.62} \]

\[ \text{HTA=n SMOKE=n CHOL(1.200) \Rightarrow LCX(0,50) s=0.07, c=1.1, l=1.62} \]

\[ \text{AGE[40,60] S[1.0,2] CHOL(200) \Rightarrow LM(0,30) s=0.05, c=1, l=1.07} \]

\[ \text{SEX=F HTA=n CHOL(20) \Rightarrow RCA(0,50) s=0.02, c=1, l=1.76} \]

\[ e \geq 0.9 \]

\[ \text{AP[-1.0,2] CHOL(250,500) \Rightarrow LM(0,30) s=0.06, c=0.99, l=1.66} \]

\[ \text{AGE[60,100] S[1-1.0,2] DIAB=n \Rightarrow LM(0,30) s=0.22, c=0.95, l=1.01} \]

\[ \text{ALI[-1.0,2] SEX=F FHCAD=n \Rightarrow LM(0,30) s=0.15, c=0.98, l=1.06} \]

\[ \text{L[1.0,2] SMOKE=n PCARSUR=n \Rightarrow LM(0,30) s=0.06, c=0.99, l=1.66} \]

\[ \text{AGE[40,60] AL[-1.0,2] DIAB=n \Rightarrow LAD(0,50) s=0.03, c=0.91, l=1.86} \]

\[ \text{AGE[40,60] IL[-1.0,2] SMOKE=n \Rightarrow RC(0,50) s=0.10, c=0.91, l=1.6} \]

\[ \text{AGE[40,60] DIAB=n \Rightarrow LM(0,30) s=0.23, c=0.95, l=1.02} \]

\[ l \geq 1.5 \]

\[ \text{AGE[0,40] AL[-1.0,2] \Rightarrow LAD(0,50) s=0.03, c=0.91, l=1.86} \]

\[ \text{SEX=F HTA=n CHOL(20) \Rightarrow RCA(0,50) s=0.02, c=1.00, l=1.76} \]

\[ \text{AGE[40,60] L[-1.0,2] CHOL(200,250) \Rightarrow LCX(0,50) s=0.2, c=0.95, l=1.5} \]

Two arteries in the consequence:

\[ \text{AGE[0,40] AL[-1.0,2] \Rightarrow LM(0,30) LAD(0,50) s=0.03, c=0.91, l=1.96} \]

\[ \text{SEX=F CHOL(200) \Rightarrow LM(0,30) RCA(0,50) s=0.03, c=0.83, l=1.45} \]

\[ \text{IL[-1.0,2] SMOKE=n \Rightarrow LM(0,30) LCX(0,50) s=0.18, c=0.77, l=1.31} \]

Two items (simple):

\[ \text{CLAUDI=n \Rightarrow LM(0,30) s=0.71, c=0.94, l=1.01} \]

\[ \text{AGE[40,60] \Rightarrow LM(0,30) s=0.37, c=0.95, l=1.01} \]

\[ \text{AS[-1.0,2] \Rightarrow LM(0,30) s=0.73, c=0.94, l=1.01} \]

\[ \text{CHOL(250,500) \Rightarrow LM(0,30) s=0.07, c=0.99, l=1.06} \]

Figure 7: Association rules predicting the absence of heart disease
4. Conclusion

The exponential growth of information, data and knowledge in the medical field has made the implementation and use of efficient computer based analysis a necessity to enhance the traditional manual methods of data analysis. Data mining is one of the important steps in the knowledge discovery process. It consists of data cleaning, data integration, data mining and then the final presentation of useful knowledge to the user. Data warehouse is a new database model that has been developed to better support data mining techniques. Data mining typically consists of descriptive and predictive data mining. Concept hierarchies and association rule mining are common topics under predictive data mining while classification, prediction and clustering are common subtopics of predictive data mining.

The application of information mining techniques to the medical domain are helpful in extracting medical knowledge for diagnosis, decision-making, screening, monitoring, therapy support and patient management. However, there are certain challenges in this domain that has to be overcome before data can be mined successfully. Not only are medical datasets large, complex, heterogeneous, hierarchical, time-varying and of varying quality, but there exists a substantial medical knowledge base that demands a robust collaboration between the data miner and health care professionals if useful data is to be mined. Despite the challenges, researchers have tried to incorporate data mining in the medical domain and some applications for the usage was presented. 667,000 clinic data from a health center in an academic institution was mined using IBM’s HealthMiner and the results were presented. Several correlations which are unknown in the medical industry were present in the results and with a domain expert’s help these correlations might be used to extract important information for the benefit of patients. The second application was different in that data mining was not used to find underlying patterns but rather to retrieve documents from the data repository at the National Cancer Institute. Two more applications were presented, and the former used classification and prediction while the latter used association rule mining techniques.
References
[17] Shalvi S, DeClaris N An Unsupervised Neural Network Approach to Medical Data Mining Techniques, Medical Informatics and Computational Intelligence Research
[23] Jaroslaw Stepaniuk Rough set data mining of diabetes mellitus data,
[29] Ying Lu, Jiawei Han, Cancer classification using gene expression data, Information Systems, v.28 n.4, p.243-268, June 2003 [doi>10.1016/S0306-4379(02)00072-8]
[32] Ordonez C Association Rule Discover with Train and Test Data Approach for Heart Disease Prediction
[33] Huang MJ, Chen MY, Lee SC Integrating data mining with care-based reasoning for chronic diseases prognosis and diagnosis
[34]Chang CL, Chen CH Applying decision tree and neural network to increase quality of dermatological diagnosis