21

Data Mining in Biomedicine and Science

Objectives:

- Data mining can be used to help predict future patient behavior and to improve treatment programs. Data mining has been used in a number of nonmedical domains.
- To turn information into knowledge, health care organizations are implementing data mining technologies to help control costs and improve the efficacy of patient care.
- The data mining algorithms such as rough sets and prediction algorithms significantly reduce patients' risk and diagnosis costs. The proposed algorithms use features extracted from data sets of different origins.
- Sequential pattern analysis and similarity search techniques have been developed in data mining, which has become a powerful tool and contributes substantially to DNA analysis.
- A series of unsupervised neural networks approach to medical data mining were designed and actual data were used to evaluate their performance in identifying natural clusters of patient population.
- The process of diagnosis can be enhanced if the physician can always be up to date with any new information – the kind of information procured through data mining. Data mining is the process of discovering nonobvious and potentially useful patterns in large data repositories such as warehouses.
- An architecture that uses the assistance of data mining to provide decision support for fever diagnosis is proposed. The proposed system also uses concepts of artificial intelligence to carry out the process of diagnosis. The main difference between a usual expert system is that the data mining process enhances the learning process.
- Scientific instruments can easily generate terabytes and even petabytes (a million gigabytes) of information. Scientific data is frequently in the form of images; there is also time series and sequence data such as DNA sequences, which need special algorithms to be dealt with.

S. Sumathi and S.N. Sivanandam: *Data Mining in Biomedicine and Science*, Studies in Computational Intelligence (SCI) **29**, 499–543 (2006)
www.springerlink.com

- Databases that used to store millions of bits of useless information can be mined for insights that can greatly profit the miners. Recently, scientific instruments and business systems have been gathering extra information that was apparently useless simply because it was so easy to do.
- Algorithm development and m(ADaM) incorporates algorithms for detecting a variety of geophysical phenomena to address the needs of the earth science community. Its flexible architecture design has made it possible for ADaM to handle the multiple formats, scales, resolutions, and large granule sizes typical of spatial data for many different science problems.

Abstract. The explosive growth in data collection in business and scientific fields has literally forced upon us the need to analyze and mine useful knowledge from it. Data mining refers to the entire process of extracting useful and novel patterns/models form large data sets. Due to the huge size of data and amount of computation involved in data mining, high-performance computing is an essential component for any successful large-scale data mining applications.

The past decade has seen an explosive growth in biomedical research, ranging from the development of new pharmaceuticals and advances in cancer therapies to the identification and study of the human genome by discovering large-scale sequencing patterns and gene functions.

In this section, the applications in biomedicine and DNA data analysis and an unsupervised neural network approach to medical data mining techniques and data mining-assisted decision support for fever diagnosis – case studies are presented. In addition, the applications of data mining in science with case studies are illustrated.

An architecture that uses the assistance of data mining to provide decision support for fever diagnosis is proposed. Today's advanced scientific instruments can easily generate terabytes and even petabytes (a million gigabytes) of information. Although data mining can be invaluable tool in analyzing this data, it faces the additional challenge that scientific data frequently is not in a convenient flat file format. Scientific data is frequently in the form of images, which are relatively easily examined by humans, but which present a myriad of problems for data mining programs. There is also time series and sequence data such as DNA sequences, which need special algorithms to be dealt with effectively. Finally there are categorical values such as protein sequences. The problem with such data is that many algorithms rely on feature vectors allowed by numerical data, so these algorithms cannot be used on categorical data sets. Despite these extra difficulties, scientific data mining has still been making rapid progress.

Data mining is a concept that is taking off in the commercial sector as a means of finding useful information out of gigabytes of data. While products for the commercial environment are starting to become available, tools for a scientific environment are much rarer (or even nonexistent). Yet scientists have long had to search through reams of printouts and rooms full of tapes to find the gems that make up scientific discovery.

This section explores some of the ad hoc methods generally used for data mining in the scientific community, including such things as scientific visualization, and outline how some of the more recently developed products used in the commercial environment can be adapted to scientific data mining.

The Information Technology and Systems Center (ITSC) at the University of Alabama in Huntsville developed the Algorithm Development and Mining (ADaM)

system under a research grant from NASA to investigate new methods of processing large volumes of Earth Observing System (EOS) remote-sensing data sets. This system provides knowledge discovery and data mining capabilities for data values as well as for metadata and catalogs the information discovered. ADaM incorporates algorithms for detecting a variety of geophysical phenomena to address the needs of the earth science community. This data mining system has been used for other research studies dealing with topics such as texture classification, image processing, and statistical analysis of earth science data sets. This section provides a detailed description of the ADaM system architecture, design, components, client interface, and the processing environment. It also describes the future directions that ITSC intends to pursue with ADaM.

21.1 Applications in Medicine

With the widespread use of medical information systems that include databases, which have recently featured explosive growth in their sizes, physicians and medical researchers are faced with a problem of making use of the stored data. The traditional manual data analysis has become insufficient, and methods for efficient computer-assisted analysis indispensable, in particular those of data mining and other related techniques of knowledge discovery in databases and intelligent data analysis.

21.1.1 Health Care

The effective use of information and technology is crucial for the health care organizations to stay competitive in today's complex, evolving environment. The challenges faced when trying to make sense of large, diverse, and often complex data source of considerable. In an effort to turn information into knowledge, health care organizations are implementing data mining technologies to help control costs and improve the efficacy of patient care. Data mining can be used to help predict future patient behavior and to improve treatment programs. By identifying high-risk patients, clinicians can better manage the care of patients today so that they do not become the problems of tomorrow.

21.1.2 Data Mining in Clinical Domains

As clinical data becomes larger, either through the proliferation of features to be collected and stored, or the number of records in databases, or both, a concern has been growing as to how to "get the data back out." Even more urgently, researchers, clinicians, and managers have become concerned with getting knowledge out of these large, complex databases. Simple database queries fail to address this concern of several reasons: a query might not retrieve the information desired because of user bias, lack of skill or experience, or limitations of the query software or database platform. In addition, the data in large databases often represents extremely complex concepts that may escape

even the most experienced content expert working with a highly competent database developer. As a result, various industries have looked to data mining as a general approach for automatically discovering knowledge "hidden" in these databases. Data mining has been used in a number of nonmedical domains, such as fraud detection and marketing, and is now increasingly being used in a variety of health care database environments, including insurance claims, electronic medical records, epidemiological surveillance, and drug utilization.

21.1.3 Data Mining In Medical Diagnosis Problem

The interest in systems for autonomous decisions in medical and engineering applications is growing, as data is becoming more easily available. Though the two areas – medicine and engineering – appear to be remote in terms of the underlying processes, both face many common challenges. One of the problems of interest of both areas is autonomous prediction. Here one instance of the prediction problem, i.e., the diagnosis problem in medicine is considered.

An SPN (solitary pulmonary nodule) is a lung abnormality that may be cancerous or benign. Over 160,000 people in the US only are affected by lung cancer, and over 90% of them die. It is important that SPNs are diagnosed early and accurately. The clinical diagnosis of SPN using information from noninvasive tests is 40–60% accurate. This figure implies that many patients suspected of malignancy have to undergo biopsy that involves considerable risks (including death) and costs to them. The data mining algorithms such as rough sets (it is one unique theory in data mining) and prediction algorithms as in A. Kusiak *et al.*, 2000, significantly reduce patients' risk and diagnosis costs. In a typical SPN disease occurrence scenario, a nodule is detected on a patient's chest radiograph. As this SPN may be either benign or malignant, further testing is required to determine is exact nature. The diagnosis is perceived to depend on many features, such as the SPN diameter, border character, presence of calcification, patient's age, smoking history, and so on. Multiple medical disciplines are involved collecting a large volume of clinical data at different times and locations, with varying accuracy and consistency. Therefore, an approach that fuses information from different sources and intelligently processes large volumes of data is needed. The proposed algorithms use features extracted from data sets of different origins. The research shows that the number of features (results of noninvasive tests, patient's data, etc.) necessary to diagnose an SPN is smaller than that used in current medical practice. At the same time the decisions made are 100% accurate.

21.2 Data Mining for Biomedical and DNA Data Analysis

Since a great deal of biomedical research has focused on DNA data analysis, we study this application here. Recent research in DNA analysis has led to the

discovery of genetic causes for many diseases and disabilities, as well as the discovery of new medicine and approaches for disease diagnosis, prevention, and treatment.

An important focus in genome research is the study of DNA sequences since such sequences form the foundation of the genetic codes of all living organisms. All DNA sequences comprise four basic building blocks (called *nucleotides*): adenine(A), cytosine(C), guanine(G), and thymine(T). These four nucleotides are combined to form long sequences or chains that resemble a twisted ladder.

Human beings have around 100,000 genes. A gene usually comprises hundreds of individual nucleotides arranged in a particular order. There are almost an unlimited number of ways by which the nucleotides can be ordered and sequenced to form distinct genes. It is challenging to identify particular gene sequence patterns that play roles in various diseases. Since many interesting sequential pattern analysis and similarity search techniques have been developed in data mining, data mining has become a powerful tool and contributes substantially to DNA analysis in the following ways.

21.2.1 Semantic Integration of Heterogeneous, Distributed Genome Databases

Due to the highly distributed, uncontrolled generation and the use of a wide variety of DNA data, the semantic integration of such heterogeneous and widely distributed genome databases becomes an important task for systematic and coordinated analysis of DNA databases. This has promoted the development of integrated data warehouses and distributed federated databases to store and manage the primary and derived genetic data. Data cleaning and data integration methods developed in data mining will help the integration of genetic data and the construction of data warehouses for genetic data analysis.

21.2.2 Similarity Search and Comparison Among DNA Sequences

We have studied similarity search methods in time-series data mining. One of the most important search problems in genetic analysis is similarity search and comparison among DNA sequences. Gene sequences isolated from diseased and healthy tissues can be compared to identify critical differences between the two classes of genes. This can be done by first retrieving the gene sequences from the two tissue classes, and then finding and comparing the frequently occurring patterns of each class. Usually, sequences occurring more frequently in the diseased samples than in the healthy samples might indicate the genetic factors of the disease; on the other hand, those occurring only more frequently in the healthy samples might indicate mechanisms that protect the body from the disease. Notice that although genetic analysis requires similarity search, the technique needed here is quite different from that used for time-series data. For example, data transformation methods such as scaling, normalization,

and window stitching, which are popularly used in the analysis of time-series data, are ineffective for genetic data since such data are nonnumeric data and the precise interconnections between different kinds of nucleotides play an important role in their function. On the other hand, the analysis of frequent sequential patterns is important in the analysis of similarity and dissimilarity in genetic sequences.

21.2.3 Association Analysis: Identification of Co-occurring Gene Sequences

Currently, many studies are focusing on the comparison of one gene to another. However, most diseases are not triggered by a single gene but by a combination of genes acting together. Association analysis methods can be used to help determine the kinds of genes that are likely to co-occur in target samples. Such analysis would facilitate the discovery of groups of genes and the study of interactions and relationships between them.

21.2.4 Path Analysis: Linking Genes to Different Stages of Disease Development

While a group of genes may contribute to a disease process, different genes may become active at different stages of the disease. If the sequence of genetic activities across the different stages of disease development can be identified, it may be possible to develop pharmaceutical interventions that target the different stages separately, therefore achieving more effective treatment of the disease. Such path analysis is expected to play an important role in genetic studies.

21.2.5 Visualization Tools and Genetic Data Analysis

Complex structures and sequencing patterns of genes are most effectively presented in graphs, trees, cuboids, and chains by various kinds of visualization tools. Such visually appealing structures and patterns facilitate pattern understanding, knowledge discovery, and interactive data exploration. Visualization therefore plays an important role in biomedical data mining.

21.3 An Unsupervised Neural Network Approach to Medical Data Mining Techniques: Case Study

In this section we present the application of unsupervised neural networks with data visualization approach on a set of conventional pathology data. Inherent difficulties in the utilization of such data were overcome by utilizing three data subspaces identified as Drugs, Topography and Morphology. This case

study is taken from D.Shalvi et al., Medical Informatics and Computational Intelligence Research Lab, University of Maryland, Baltimore, Maryland.

A series of unsupervised neural networks were designed and actual data were used to evaluate their performance in identifying natural clusters of patient population. Included is a method to examine and validate the underlying reasons for clustering. Preliminary examinations of identified clusters by qualified pathologists have shown promising results, which supports the conclusion that the suggested methodology yields discoveries and medical interpretations that can eliminate or serve as alternatives to special purpose of epidemiological studies.

21.3.1 Knowledge Extraction Through Data Mining

Data collection is often undertaken to monitor individual cases such as a patient in a hospital. This type of data is distinguishable from data sets collected for the purposes of studying a population; for example, determining the television shows that are the most popular. Individual data sets may be gathered and studied collectively for purposes other than those for which the sets were originally created; in such fashion new knowledge may be obtained while simultaneously eliminating one of the largest costs in developing knowledge, data collection. This approach is especially appropriate for medical data, which often exists in vast quantities in an unstructured format. Applying data mining techniques can facilitate systematic analysis.

Data mining is the process of sifting through and analyzing rich sets of domain specific data and then extracting the information and knowledge in the form of new relationships, patterns, or clusters for decision making purposes. Thus data mining is a form of knowledge discovery essential for solving problems in a specific domain. Conventionally, data is gathered to test an existing hypothesis (a top-down search). Alternatively, the existing data is mined and allowed to form natural clusters (a bottom-up finding). Cluster detection may be employed through statistical techniques such as Bayes' theorem or nonstatistical techniques such as unsupervised neural networks, which form clusters on the data set without knowing what the output clusters should model.

In this section Kohonen self-organizing maps (SOMs) are used to cluster a specific medical data set containing information concerning the patient's drugs, topographies (body locations), and morphologies (physiological abnormalities); these categories can be identified as the three input subspaces. Data mining techniques are used to collapse the subspaces into a form suitable for network classification. The goal is to acquire medical knowledge, which may lead to tool formation, automating an assist to medical decisions regarding populations.

21.3.2 Traditional Difficulties in Handling Medical Data

Medical data typically requires a large amount of preprocessing in order to be useful. There is numeric and textual data interspersed. Frequently different

symbols are used with the same meaning; "male" may be denoted as "M," "m," 0, or a variety of other formats. One medication or condition may be commonly referred to by a variety of names. There is often a redundancy of data; age may appear in several places. Erroneous data is very common; medical terms are frequently misspelled. Finally, medical data is frequently sparse; when a structure is imposed on medical data much of the structure remains empty for a large portion of the population due to the breadth required of any structure.

A robust data preprocessing system is required in order to draw any kind of knowledge from even medium-sized medical data sets. The data must not only cleaned of errors and redundancy but organized in a fashion that makes sense for the problem; in this today's context, the data must be organized so that the benefits of using unsupervised neural networks may be maximized.

21.3.3 An Illustrative Case Study

For the purposes of this study a confidential real-world medical data set was used. This data set is medical of the many in the medical field in terms of the data difficulties mentioned above; redundant and erroneous data is frequent. The data makes use of and introduces main-specific knowledge organized as three hierarchical trees, identified as Drugs, Topography, and Morphology. Each tree is several layers deep providing sample opportunity for exploiting the trees' structure at various levels of depth. The most significant portion of the Morphology tree is displayed in Fig. 21.1.

After the requisite data preprocessing we can use unsupervised neural networks to organize the data into clusters of patients containing similar features. Data mining is employed in both before and after the neural network; before, to exploit the substructure of the provided trees, and after, to group the finegrain output clusters into larger clusters that perhaps contain more meaningful data. Finally, the original data is mined to explain the output clustering and extract the knowledge of the neural network in terms that the pathologist can understand and qualify. The goal is to discover hidden knowledge of the relationship between drugs, Topography, and Morphology.

21.3.4 Organizing Medical Data

Standard techniques were employed to clean erroneous and redundant data; for example, "GI Bleeding," "Gi Bleed," and "Gastrointestinal Bleeding" were all mapped to "Gastrointestinal Bleeding." If "Albuterol" appeared more than once for a patient, multiple entries were discarded.

The data set used contains only data at the leaves; each leaf node is present in only one or two tuples on average. This poses two problems: any conclusions formed would be statistically insignificant and the level of computation required for such an analysis would be exceedingly high due to the existence of roughly ten thousand different leaf nodes. To alleviate these problems the

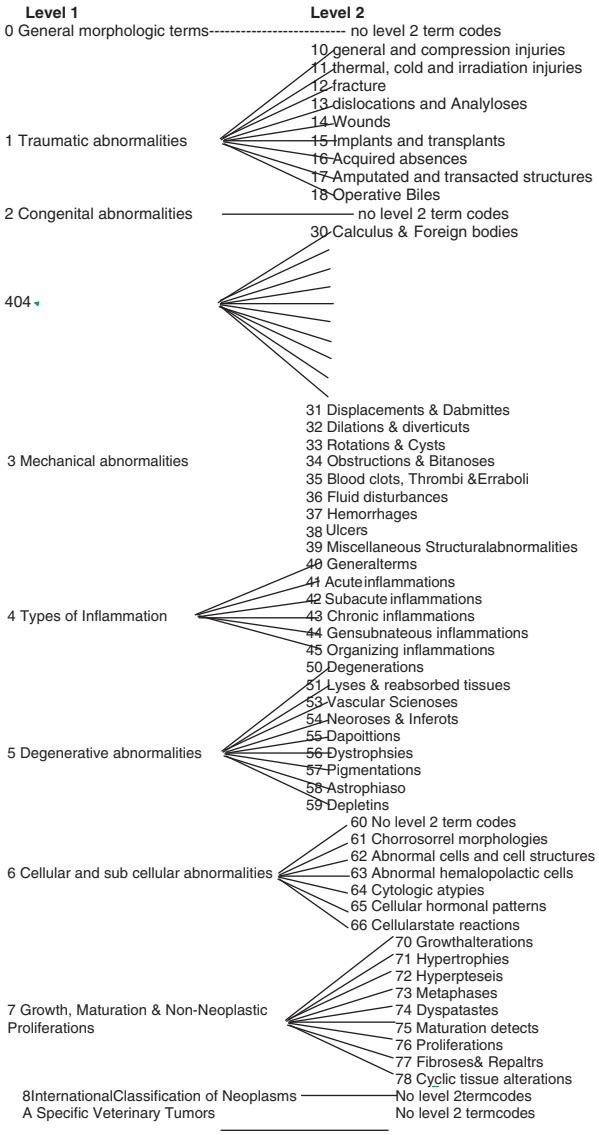

Fig. 21.1. Morphology Tree – Condensed

data was processed at the root level of each tree. At the root level each tree collapses into much fewer nodes; fourteen root-level drugs, sixteen root-level topographies, and ten root-level morphologies. By constraining all data to the root level the degree of differentiation has been greatly reduced from thousands to $40\,(14 + 16 + 10)$. The trade-off in cost is a great reduction in precision, while a benefit is the possibility of detecting trends within the data at the general level.

As the trees were collapsed to the root level the per tuple data was converted to bipolar format. For every tuple each of the 40 root level nodes was assigned a value of either 1 or -1 depending on whether any data existed for the leaves of that root node. The node is assigned a value of 1 if at least one data value is present at the leaves. The node is assigned a value of -1 if no data is at the leaves. In other words, only existence is preserved; quantity is lost. The end result for each tuple is a 40-dimensional bipolar array. The original data is at most 18-dimensional, containing one to six drugs, one to six topographies, and one to six morphologies. Each dimension may contain one of thousands of values, and there is no structure to the data due to variable dimensionality and null and repeated values. By contrast, every tuple of the transformed data contains 40 dimensions, each of which may take one of only two values, 1 or -1. The transformed data is much more consistent and lends itself to computationally intensive analysis such as neural networks.

21.3.5 Building the Neural Network Tool

The Kohonen self-organizing map (SOM) was chosen to organize the data in order to make use of a spatially ordered 2-dimensional map of arbitrary granularity. This architecture proved useful when applying data mining and data visualization techniques to the output. An $n \times n$ SOM was implemented for $n = 5, 7, 10$, and 20. The Input layer consists of 40 input nodes, corresponding to the root layers of the three trees: 14 Drug nodes, 16 Topography nodes, and 10 Morphology nodes.

The following parameters were used to train a 10×10 network for trial 0. The training period was set to the default value of 30 epochs, which yields 62430 training iterations for 2081 input tuples. The learning coefficient α is initialized to 0.06. After approximately 7.5 epochs α is halved to 0.03. After another 7.5 epochs it is halved again to 0.015. For the final set of 7.5 epochs it is halved again to become 0.0075. Thus, the network is trained relatively quickly at the beginning; as learning progresses training is reduced to finetuning. After 30 epochs no learning occurs.

Other network parameters are varied over time as well. The frequency estimation parameter β is initialized to 0.0005 and successively halved to 0.0001 for the final set of 7.5 epochs. The conscience parameter γ is initialized to 1.0 and successively halved to 0.125 for the final set of 7.5 epochs. The neighborhood shape is a constant square of variable width. The neighborhood

width is reduced over time from 7 nodes to 1 node; this determines the portion of the network that learns after every iteration.

After the network is trained it is used for one final pass through the input data set in which the weights are not adjusted. This network testing provides the final classification of each input data tuple into a single node in the 10×10 grid. The output is taken from the coordinate layer as an (x, y) pair. SOM trial 1 also uses n = 10 and is similar to trial 0 save for random initial conditions. For trial 2 n = 20; correspondingly, the neighborhoods are larger throughout training. In trial 3 n = 5 and in trial 4, n = 7; these trials contain smaller neighborhoods throughout training. For all trials, the learning parameters mentioned above are very similar.

21.3.6 Applying Data Mining and Data Visualization Techniques

The output of the SOM is a population distribution of tuples with spatial significance, which may be decomposed into three subspace distributions representing the decision boundaries determined by the network. The population distribution may be aggregated into larger groups to provide clusters not distinguishable in a high-resolution map. Such clusters can then be analyzed according to the subspace decomposition to determine significant patterns in the data.

Figure 21.2 displays the output data population distribution for the trial 0 network. This grid displays the number of tuples that were classified into each Kohonen layer node (square) during testing. Square (1,1) contains 180 tuples, by far the largest number of any square. Square (2,1) contains only one tuple and square (1,2) contains none. As SOMs tend to shrink or expand the output map as needed to fit the data these valleys suggest that the cluster at (1,1) is very sharp and well defined.

This hypothesis can be confirmed by looking at the raw data, which upon examination reveals that every one of the tuples in square (1,1) contains root level data only for Drug 6, Topography 6 and Morphology 5. The tuple at square (2,1) contains these three root level nodes as well as Drug 7, a difference slight enough for the network to distinguish the tuple by classifying it one square away from (1,1). All of the 34 tuples in square (3,1) contain Drug 6, Topography D and Morphology 5 but only 29 of the 34 tuples contain Topography 6. Clearly the difference between square (3,1) and square (1,1) is greater than that between square (2,1) and square (1,1).

The initial network organization of 100 squares is aggregated to form clusters that are spatially large; the outline of these areas was determined very subjectively. It is represented by any (x, y) coordinate pair, which has a population greater than 10 (an arbitrary threshold) combined to form spatially large clusters.

Cluster A comprises only one square at 1; this is 1% of the output space. However, this cluster of population of 180; or almost 9% of all the tuples! As this tends to distribute the data over the entire 2-dimensional output space,

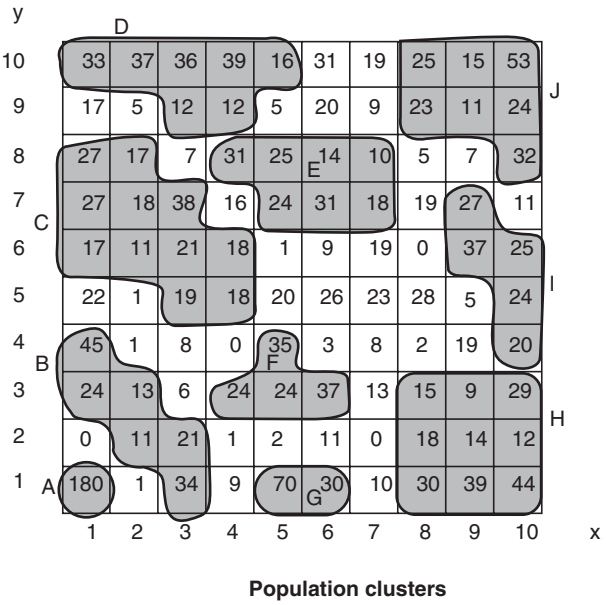

Fig. 21.2. Population of Clusters

this indicates that cluster A is well defined. Population and spatial considerations and on domain-specific knowledge were used to create cluster map in Fig. 21.2. Upon examination of the raw within these clusters one finds similarities between the tuples, which are indicative of medical relationships or dependencies. Numerous hypotheses can be made regarding these relationships, many of which were not a *priori* known.

The SOM groups together tuples in each square regarding to their similarity. The only level at which the SOM can detect similarities between tuples is at the root level of each of the three subspace trees since this was the level of differentiation presented to the SOM's input. Consequently we should expect that tuples in the same are often have the same root level drugs, topographies morphologies. Using data mining principles we can down from the output map into the raw data to over the source of the SOM's clustering.

For instance, square (3,10) belongs to cluster D contains 36 tuples. Drilling down into the data it is covered that of the 36 tuples:
36 tuples (100%) contain Drug 2 (Industrial products) 3 tuples (8%) contain Topography 5 (Digestive Tract) 13 tuples (36%) contain Topog. 6 (Digestive Organs) 3 tuples (8%) contain Topography 7 (Urinary Tract) 12 tuples (33%) contain Topography A (Nervous System & Special Sense Organs)
1 tuple (2%) contain Topog. B (Endocrine System) 5 tuples (13%) contain Topography C (Hematopoietic & Reticuloendothelial Systems) 2 tuples (5%) contain Topog. D (Topog. Regions) 36 tuples (100%) contain Morphology 8 (International Classification of Neoplasms)

Clearly the presence of Drug 2 and Morphology 8 defined are (3,10) since all of the tuples belonging to this are root-level nodes. For most squares, if it is covered that there is one or at most two morphologies each can be found in all or roughly 95% of the tuples. If arbitrary threshold level is dropped to 66% of the tuples there is almost always one or two morphologies that will surpass the threshold. If the threshold level is at 100% the following morphology cluster map may be constructed; this appears as Fig. 21.3.

The number inside each square in the morphology cluster map represents the root-level code of the morphology tree. For example, there are a large group of squares in the bottom of the graph that contain a 5; these square form a cluster for which 100% (the chosen threshold) of the tuples inside the cluster contain Morphology 5, Degenerative Abnormalities. There are at least six distinct clusters in this graph. Clearly the morphology clusters are arranged in a spatially meaningful way by the SOM; one cluster flows into the next.

A comparison between Fig. 21.2 and Fig. 21.3 shows that there appears to be some relation between the Population clusters and the Morphology clusters. Population clusters A (Morph. 5), C(4), F(5), G(5), and J(3) are entirely enclosed by their respective Population clusters. Clusters D and H are almost entirely enclosed, while clusters B, E, and I span two or more clusters, indicating that the Population clusters are not well formed.

The threshold level of 100% was chosen because in most squares, one or two morphologies are found in 100% of the square's tuples. If the thresh-

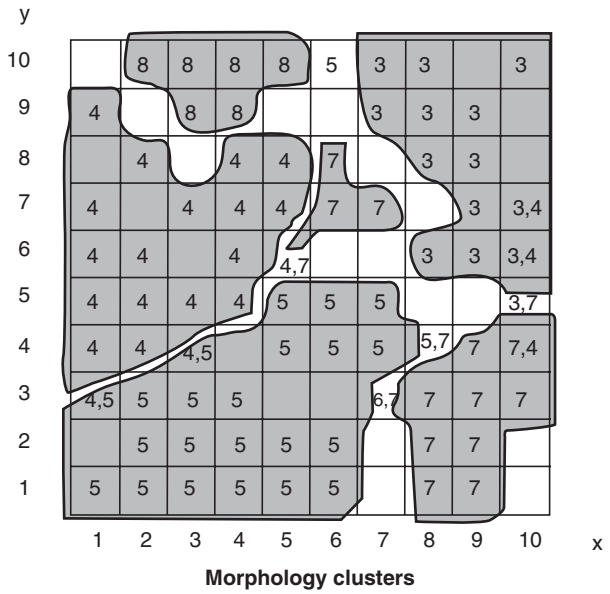

Fig. 21.3. Morphology Clusters

old is dropped to 66%, the clusters somewhat expand in size, and the overlap between the Morphology and Population clusters increases. In particular, cluster H is now completely enclosed by Morphology 7, translating to a cluster in which 66% of the population contains Morphology 7.

A similar drill-down analysis is done for the Drug and Topography subspaces, yielding maps that can be used to explain many of the population clusters. Additionally, trials 1–4 yielded extremely similar results, with identical clusters identifiable to those of trial 0, indicating that the clustering of the data is independent of network size and training parameters. It is now possible to introduce several heuristics to rate the worthiness of the population clusters.

The clusters delineated in the previous section were formed on the basis of population and spatial distribution alone. The worthiness of these clusters can be measured by the following parameters, which make use of square tuple population, cluster square count, and overlap with the three tree subspaces. Subsequently, the population distribution is clustered again with a heavier emphasis placed on tree subspace overlap.

We can estimate how well formed the Population clusters are by looking at the total number of tuples in each cluster compared to the average number of tuples expected for each square of that cluster in a random distribution. Table 21.1 displays this analysis.

Table 21.1. Cluster assessment worksheet

Cluster	A	E	F	G	Av.
Population	180	153	121	106	165
Square Count	1	7	4	2	5.9
Tuples/Square	180	21.9	30.3	53	43.4
Cluster Intensity	**8.6**	**1.1**	**1.5**	**2.5**	**2.1**
Drug Inclusion	1	.43	.25	1	.59
Topo Inclusion	1	.86	1	.5	.69
Morp Inclusion	1	.43	1	1	.85
Total Inclusion	3.0	1.71	2.25	2.5	2.1
Scaled Square Count	1.5	2.4	1.95	1.65	2.2
Relative Inclusion	**4.5**	**4.1**	**4.4**	**4.1**	**4.6**
Cluster Assessment	**38.8**	**4.5**	**6.6**	**10.3**	**9.5**

Tuples/Square reflects the average number of tuples per square for that particular cluster. A completely random distribution of the data would produce 20.81 tuples per square. A well-defined cluster should exhibit more; the cluster Intensity value grossly reflects this; it is simply Tuples/Square divided by 20.81. Cluster E, which has an intensity of 1.1, is not a very well-defined cluster since a cluster chosen from a random collection of squares would on average produce the same intensity. By contrast cluster A, which has an intensity of 8.6, is extremely well defined. Note that the Cluster Intensity parameter

does not incorporate the population of neighboring (border) squares. A very small population in the border squares, as is the case with cluster A, should translate to a more well-defined cluster; this factor is not accounted for in the definition and would be a useful parameter modification.

Most clusters exhibit an intensity slightly greater than 1, as would be expected. However, a cluster with an intensity of 1.5 (cluster F) is much more well-defined than a cluster with an intensity of 1.1 (Cluster E); it contains 40% more tuples per square! Clusters A and G overshadow this significant difference.

Cluster assessment becomes more accurate when the Drug, Topography and Morphology cluster maps are included in the analysis. The Inclusion parameters represent the percentage of each population cluster that is encompassed in the most dominant Drug, Topography, and Morphology cluster for that Population cluster, as defined by the 100% threshold. For example, cluster F is composed of 4 squares. One of the 4 squares (25%) is part of Drug cluster A, all 4 squares (100%) are part of Topography cluster 6 and all 4 squares (100%) are part of Morphology cluster 5. The total Inclusion parameter is the sum of the 3 inclusion proportions; it is a relative measure of how well each Population clusters overlaps with the three subspace map clusters. The Scaled Square count parameter resets the Square Count scale range, originally from 1 to 11, to a new range, 1.5 to 3.0. Relative Inclusion is simply Scaled Square Count multiplied by Total Inclusion. Square Count is scaled so that the new range has the same numerical impact ($\sim$1.5 to 3.0) as total inclusion on the derived parameter. In this fashion the dependence on cluster size is reduced.

In terms of Total Inclusion clusters A, B, G and J are notably sharp. This is impressive for large clusters such as B and J. This expectation is realized in the Relative inclusion parameter for which clusters B (5.3) and J(6.2) have significantly higher values than all other clusters. Cluster C (4.6) also has a fairly large Relative Inclusion primarily due to its large size of 11 squares.

Finally, the cluster Assessment parameter is an overall assessment of the quality of the cluster, which combines Cluster Intensity (relative population) and Relative Inclusion (overlap with subspace clusters). As would be expected clusters A and G rate extremely well. Clusters B (6.3) and J (8.0) also rate fairly well, while cluster E rates very poorly. This can be attributed to cluster E's low intensity and inclusion.

These derived parameters are of course very subjective and only crudely attempt to rate a cluster. The sharpness of cluster boundaries is not included in the estimate and there is too great of a reliance on cluster size. Nevertheless, the above numbers are still very useful in rating the worthiness of heuristically determined clusters.

The following clusters were completely enclosed within two or more subspace root nodes:

- A: CNS Drugs; Digestive Organs; Degen. Abnormal.

- B: CNS Drugs; Digestive Organs
- F: Digestive Organs; Degenerative Abnormalities
- G: General Terms, Antihistamines, and Antiinfective Drugs; Degenerative Abnormalities.

These clusters are of particular interest since they relate root level nodes from different trees. Clusters A and F indicate a strong relationship between Digestive Organs and Degenerative Abnormalities; this relationship may simply be symptomatic of necrosis of the liver. A look at raw data indicates that cluster G contains many cases and lipoid degenerations (forms of degenerative Abnormalities) and the patient was often Tetracycline or Isoniazid (penicillins, or antiactive Drugs). This extracted relationship is a data set that has been discovered through data mining.

Knowing that the origin Population cluster map efficient we can recluster the network output map information gleaned from the subspace inclusion parameters. In Figure 21.4 a new cluster map is shown. Each node contains a code representing the root-level nodes for that square at 100%. For example, square (1,1) contains code 665, indicating that 100% of the tuples in that square contain Drug 6, Topography 6, and Morphology 5.

The cluster map was regenerated using clusters the subspace maps as well as spatial configuration of the original SOM feature space to realize nonactive clusters and population (to create larger clusters). Conflicting goals of the cluster formation are the cluster size and dimensionality of cluster definition, given the number of the root level nodes that are contained by the tuples

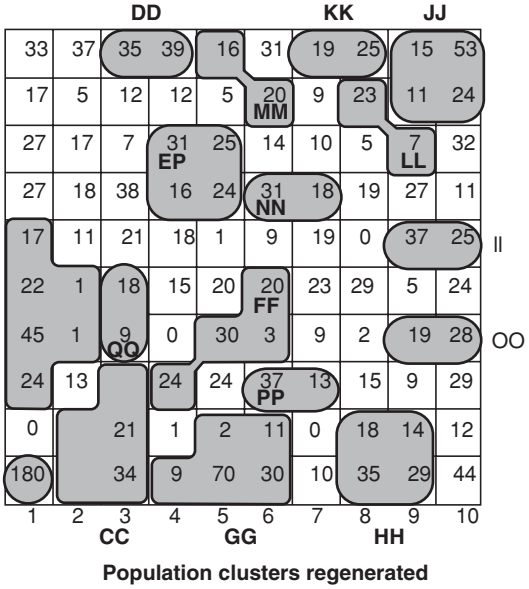

Fig. 21.4. Population clusters regenerated

in the cluster. The increased specificity and enable detection of fine clusters perhaps indicative of known relations on the data that would likely be covered with broader clustered feature space.

All of the multiple square clusters contain at least root level nodes from different trees. Many of the training single squares contain two root level nodes from different trees as well; for example, all 29 tuples in square (10,3) contains Topo. 8 and Morp. 7 as indicated by the code "x87" within the square; the "x" denotes that no single drug was contained by 100% of the tuples. In summary, the root-level node frequency data allows construction of root-level node clusters, individually by the subspace and as an intersection of subspaces. This allows generation more meaningful clusters using spatial, population, and subspace correspondence consideration. Many of the clusters identified using the above method not only comprise multiple squares but often are defined by multiple root-level nodes across the subspace trees. These clusters are a potential gold mine of data nuggets.

21.4 Data Mining – Assisted Decision Support for Fever Diagnosis – Case Study

In this section, we explain the dangers of inaccurate of delayed diagnosis of fever and show that fever diagnosis is an ideal candidate for decision support. An architecture that uses the assistance of data mining to provide decision support for fever diagnosis is proposed. The proposed system also uses concepts of artificial intelligence to carry out the process of diagnosis. We describe the various components of the architecture and the interactions between them. The architecture contains a data warehouse, which is incrementally refreshed with case details. The patterns obtained by mining the warehouse are used to formulate rules and in turn form questions in simple English. An interactive GUI, which poses questions to the physician. Each successive question is chosen based on the responses received previously. We conclude by saying that the results of such a system cannot be expected to occur in any definite time frame information can be discovered only if it is present. This case study is taken from S. Jayakumar et al. S.V. College of Engg, Pennalur, India, 2002.

Apart from being one of the most common "symptoms" encountered by physicians while treating patients, a fever is also a symptom that can have an extremely wide array of causes. Thus, a physician when confronted with a feverish patient, has to ask a number of questions and possibly make a large number of eliminations in order to arrive at the right diagnosis. The causes of fever may vary from neoplastic disease to sunstroke, and may have geographical, age-related, and racial variations. Also there are certain cases, wherein it is absolutely vital that a fever is diagnosed correctly at a very early stage. A classic example would be that of malaria. There are four species of the protozoan Plasmodium, which produce malaria in its various forms. Of these, the malaria caused by *Plasmodium falciparum* is the most dangerous, because,

if left untreated, it can cause cerebral malaria, which is fatal. Meningitis is another form of fever that can turn fatal if not treated early.

The factors mentioned above, among others, make diagnosis of fever an ideal candidate for decision support. The process of diagnosis can be enhanced if the physician can always be up to date with any new information – the kind of information procured through data mining. Data mining is the process of discovering nonobvious and potentially useful patterns in large data repositories such as warehouses.

21.4.1 Architecture for Fever Diagnosis

Here we give a brief overview of the proposed architecture. We explain the basic functions carried out by the various components of the architecture and also the interactions between these components. The detailed functioning of each of the components is explained in the sections that follow along with relevant examples.

The main components of the architecture are the Medical Data Definition component (MDD), Physician system interface (PSI), Diagnostic Question Banque (DQBanque), Pattern Extractor and Rule constructor.

The pattern extractor consists of both a data warehouse and a data mining tool set. The data warehouse consists of exhaustive case histories of patients. The patterns that are generated by mining the warehouse are used by the rule constructor in order to construct new rules. These rules are then sent to the DQBanque. The DQBanque performs the task of converting these rules into questions in simple English, while also acting as a storage area for both the rules and the questions. During the process of diagnosis, the DQBanque sends appropriate questions to the PSI, which is a GUI used by the physician. The responses received by the PSI are sent back to the DQBanque, and based on these responses the appropriate follow-up questions are selected and sent to the PSI. The proposed system also contains the MDD that contains the various fever-related terms and their explanations. This component also contains information regarding the form in which data are stored in the warehouse, the facilitating data cleaning.

21.4.2 Medical Data Definition Component

The Medical Data Definition component defines the data and the metadata related to the topic of fever diagnosis. This includes specifics like symptoms, type of fever, diet classification, frequency of symptoms, temperature, duration of symptoms, blood pressure, strain of pathogen, etc. It also includes descriptions of various patient-related information in the form of attributes. This information includes name, age, sex, height, weight, race, and location among others. To be stored in the data warehouse, it is necessary that the data be taken through a data preparation process, or in other words the data must be cleaned. Data cleaning is the process of resolving inconsistencies in

format and naming. The MDD helps in specifying the categories under which the data that is got via the PSI will be stored in the data warehouse after cleaning. It has a two-way connection to the pattern generator.

21.4.3 Physician–System Interface

This is basically a GUI that assists the physician in carrying out the process of diagnosis. It serves a dual purpose:

- Displays the appropriate questions, which it acquires from the DQBanque
- Accepts the responses and sends them to the Pattern Extractor and the DQBanque

The questions displayed by the PSI are framed by the DQBanque (in simple English) based on the rules present in it. These questions predominantly require yes/no type responses or at the most, one-phrase answers. The responses thus obtained are sent to both the Pattern Extractor and the DQBanque. The Pattern Extractor will convert the data into the appropriate form (with the help of the MDD) and feed it into the data warehouse. The DQBanque analyzes the responses and determines the next appropriate question to be posed to the physician.

21.4.4 Diagnostic Question Banque

This component serves as a repository for both diagnoses-related rules and the simple English questions derived from those rules. The rules are obtained from the rule constructor as and when they are formed. The DQBanque performs the function of converting these rules into questions, which are in simple English, along with storing these questions. Even after the conversion process, the DQBanque retains the rules in their original form as well. When the diagnosis takes place, at first the DQBanque send over standard patient related and preliminary disease-related questions to the PSI. The responses are sent back to the DQBanque in real time. After a certain stage, the next question to be asked, is decided by the DQBanque on the basis of the responses received so far. Often a particular response would necessitate the asking of a set of questions rather than just one. These questions would be placed in a stack, as would the sets of questions necessitated by each one of these questions. Thus the DQBanque performs the task of traversing all these stacks to whatever extent necessary in order to form a diagnosis.

From the above information we can see that the DQBanque has a one-way connection with the Rule Constructor and a two-way connection with the PSI.

The functioning of the DQBanque is now discussed in the context of diagnosis of bacterial meningitis. The following is a typical set of questions that would be displayed on the PSI during the course of diagnosis of this disease:

What is the name of the patient?

Andrew Miller
What is Andrews' sex?
Male
What is Andrew's age?
25
Country of Residence?
United States
State/Province?
New York
Race
Caucasian

The above questions are examples of standard queries (put to all patients), the responses to which are stored in the hospital records. The individual responses may or may not be fed into the data warehouse.

Temperature?
98.6
Pulse?
100
Respirations?
20
Blood Pressure?
110/70
Recent occurrence of seizures?
Yes

The information collected so far suggests the possibility of both bacterial meningitis and herpes encephalitis headaches?

Yes

If the response had been "No," the next question would have possibly been:

Presence of temporal lobe lesions?

A "Yes" response for this question would strongly suggest the presence of herpes encephalitis, and the line of questioning would proceed appropriately. Considering that the response was "yes" to the query about headaches, the questioning continues as follows.

Stiff neck?
Yes
Unusual change in mental status?
Yes

At this point all the classic symptoms of bacterial meningitis have been established.

Thus the typical working of the DQBanque was illustrated in the context of bacterial meningitis.

21.4.5 Pattern Extractor

This component is composed of two parts: the data warehouse and the data mining toolset. The data warehouse basically consists of a data cube and a two-dimensional table. The data cube contains an exhaustive list of case histories with all the possible relevant information such as case number, age, sex, race, geographical location, existing ailments and conditions, hereditary ailments, and current symptoms among others. The 2-D table would contain only information stating the various symptoms and conditions corresponding to each individual fever-causing disease. The data cube is augmented with fresh case details in real time as and when diagnosis takes place. The responses to the questions posed by the PSI (which are either yes/no responses or single phrases) are transferred to the pattern extractor and held in temporary storage. After the necessary data cleaning is done (with the help of the MDD), the data is fed into the data cube. The data cleaning process is done with the assistance of the MDD to resolve any inconsistencies in storage format and naming. Thus we see that the data cube is incrementally refreshed.

The data mining tool set implements a rich set of mining algorithms based on association rules, classification, and time-series analysis. These algorithms are periodically used on the data present in the data cube, in order to generate any kind of "interesting" patterns. The aimed end result of these patterns is basically to find a set of symptoms, conditions, and other factors such as race, age, gender, and location associated with each individual fever-causing disease. The information thus generated is stored in a dynamically expandable 2-D table. A typical row in the table would contain the name of the disease in the first column, and the various associated information in the other columns. So, to put it succinctly, the results of mining the data cube are stored in the 2-D table. The rule constructor receives the information present in the 2-D table and formulates new rules.

So, we see that the pattern extractor has one-way connections to both the PSI and the rule constructor. It also has a two-way connection with the MDD.

When the patient's age is taken as input, he/she is automatically classified as a neonate, a child, or an adult. The data cube present extractor contains an attribute called age group, which stores this information. The data cube also contains attributes specifying the organism, which caused the disease. For example, it is known that H. influenzae or N. meningitis generally causes the occurrence of bacterial meningitis in children. This is the kind of knowledge that can potentially be gained from the data mining process. One can also discover nonobvious racial, geographical, or diet-related patterns in the occurrence of fever. The new knowledge thus found can greatly quicken the process of diagnosis and possibly make it more accurate.

21.4.6 Rule Constructor

The Rule Constructor has a one-way connection to the pattern extractor and another one to the DQBanque. This component gleans the data stored in the

2-D table (present in the pattern extractor) and formulates rules based on this data. Any rule thus formed is first compared to the database of rules present in the DQBanque. If the rule already exists, it is discarded. Else, it is added on to the database of rules in the DQBanque. The following rules illustrate the basic format of rules formed by the Rule Constructor.

IF seizure = yes AND

Headache = yes AND

Stiffneck = yes AND

Mentalstatechange = yes

THEN Bacterial/Meningitis.

IF seizure = yes AND

Nonseasonalfocalneurologicdefect = yes

AND

Temporallobeleisions = yes

THEN Herpesencaphalitis

When a "yes" response is received for the query "seizures?" the DQBanque is searched for rules having an "If" condition "seizures = yes." When such a rule is found, the other conditions of the rule are traversed one by one, and the corresponding questions are chosen from the DQBanque and sent to the PSI. When a certain condition in the rule is not satisfied, that rule is discarded, and the search starts for another rule, which has the condition "seizures-yes." This is basically how questions are chosen and sent to the PSI. Although this explanation basically pertains to a process taking place in the DQBanque, we chose to give it here, as we wanted all rule-related discussions to appear together. These rules have scope for evolving further as the mining process takes place. It would generally be a slow and almost static evolution.

21.5 Data Mining and Science

Although data mining can be invaluable tool in analyzing this data, it faces the additional challenge that scientific data frequently is not in a convenient flat file format. Scientific data is frequently in the form of images, which are relatively easily examined by humans, but which present a myriad of problems for data mining programs. There is also time-series and sequence data such as DNA sequences, which need special algorithms to be dealt with effectively.

An example of mining scientific data was the cataloging of a sky survey. The Second Palomar Observatory Sky Survey took six years to collect three terabytes of image data containing an estimated two billion sky objects. The

three thousand photographic images were scanned into 16-bit pixel resolution digital images at $23,040 \times 23,040$ pixels per image. The problem was generating a survey catalog of all these sky objects from this information. Additionally, the attributes and class of each object are needed to be determined and recorded in the catalog. To solve this problem, the Sky Image Cataloging and Analysis Tool (SKI-CAT) system was developed.

The majority of objects in each image were faint, making determination of their class by visual inspection or classical computational approaches in astronomy impossible. SKI-CAT used decision tree-learning algorithms to accurately predict sky object classes. This accuracy was verified by comparison with a set of high-resolution charged-couple device images. SKI-CAT was 94% accurate at predicting the class of sky objects, which increased the number of reliably classified objects by 300%. These results have already helped astronomers discover sixteen new high red-shift quasars. Such quasars are difficult to find and provide clues about the early history of the universe.

The Magellan spacecraft orbited the planet Venus for over five years and used synthetic aperture radar to penetrate the gas and cloud cover to map the surface of the planet. The result is that we have a unique high-resolution map of the entire planet. In fact, we have more of Venus mapped at the 75-m pixel resolution than we do of the Earth because so much of the Earth is covered by water. This data set is valuable because of its completeness and because Venus is the most similar to Earth in size. It is hoped that learning about the geological evolution of Venus will produce valuable lessons about the Earth.

The immense size of this data set prevents planetary geologists from personally examining all the images. To assist geologists in analyzing the Venus map, the Jet Propulsion Laboratory developed the Adaptive Recognition Tool (JARtool). The system seeks to automate the search for small volcanoes by training the system via examples. The geologists would label a small sample of the images and the system would then use these to train itself to recognize small volcanoes. The system would then attempt to locate and measure the planet's estimated one million small volcanoes. It used classification learning to distinguish true detections of volcanoes from false alarms. It performed as well as scientists in identifying common types of small volcanoes, but rarely detected those scientists are not sure about.

The geoscientific data mining system Quakefinder automatically detects and measures tectonic activity in the earth's crust using satellite data. It was used to map the direction and magnitude of group displacements due to the 1992 Landers earthquake in Southern California over a spatial region of several hundred square kilometers at a resolution of 10 m to a subpixel precision of 1 m. Quakefinder is implemented on a 256-node Gray T3D parallel supercomputer so that the gathered data can rapidly produce scientific results. Besides automatically measuring known faults, it also allows automatic knowledge discovery by indicating novel unexplained tectonic activity away from the primary faults never before observed. Future work will focus on the

measurement of continuous processes over many images, instead of simply measuring abrupt behavior seen during earthquakes.

In summary, although data mining is still limited in its functionality, its potential is nearly unlimited. Already business, science, and security have derived benefits from its development. Databases that used to store millions of bits of useless information can be mined for insights that can greatly profit the miners. Recently, scientific instruments and business systems have been gathering extra information that was apparently useless simply because it was so easy to do. The creation of data mining makes this excess information useful.

21.6 Knowledge Discovery in Science as Opposed to Business-Case Study

The relatively new discipline of data mining is most often applied to extraction of useful knowledge from business data. However, it is also useful in some scientific applications where this more empirical approach complements traditional data analysis. The example of machine learning from air quality data illustrates this alternative. This case study is taken from Brian J Read, CLRC Rutherford Appleton Lab, Oxon, UK

Data mining is the essential ingredient in the more general process of knowledge discovery in databases (KDD). The idea is that by automatically sifting through large quantities of data it should be possible to extract nuggets of knowledge.

Data mining has become fashionable, not just in computer science (journals and conferences), but particularly in business IT. (An example is its promotion by television advertising.) The emergence is due to the growth in data warehouses and the realization that this mass of operational data has the potential to be exploited as an extension of *business intelligence.*

21.6.1 Why is Data Mining Different?

Data mining is more than just conventional data analysis. It uses traditional analysis tools (like statistics and graphics) plus those associated with artificial intelligence (such as rule induction and neural nets). It is all of these, but different. It is a distinctive approach or attitude to data analysis. The emphasis is not so much on extracting facts, but on generating hypotheses. The aim is more to yield questions rather than answers. Insights gained by data mining can then be verified by conventional analysis.

21.6.2 The Data Management Context

"Information technology" was originally "data processing." Computing in the past gave prominence to the processing algorithms – data were subservient.

Typically, a program processed input data tapes (such as master and detail records) in batches to output a new data tape that incorporated the transaction. The structure of the data on the tapes reflected the requirements of the specific algorithm. It was the era of Jackson Structured Programming. The concept of database broke away from this algorithm-centric view. Data assumed an existence independent of any programs. The data could be structured to reflect semantics of relationships in the real world. One had successively hierarchical, network, relational, and object data models in commercial database management systems, each motivated by the desire to model better the structure of actual entities and their relationships.

A database is extensional, storing many facts. Some information is intentional; that is, it manifests as rules. Some limited success was achieved with deductive databases that stored and manipulated rules, as for example in Prolog-based systems. This encouraged *Expert Systems.* However, it was hard to achieve solid success. The main difficulty was the knowledge elicitation bottleneck: how to convert the thought processes of domain experts into formal rules in computer.

Data mining offers a solution: automatic rule extraction. By searching through large amounts of data, one hopes to find sufficient instances of an association between data value occurrences to suggest a statistically significant rule. However, a domain expert is still needed to guide and evaluate the process and to apply the results.

21.6.3 Business Data Analysis

Popular commercial applications of data mining technology are, for example, in direct mail targeting, credit scoring, churn prediction, stock trading, fraud detection, and customer segmentation. It is closely allied to data warehousing in which large (gigabytes) corporate databases are constructed for decision support applications. Rather than relational database with SQL, these are often multidimensional structures used for the so-called *on-line analytical processing* (OLAP). Data mining is a step further from the directed questioning and reporting of OLAP in that the relevant results cannot be specified in advance.

21.6.4 Scientific Data Analysis

Rules generated by data mining are empirical – they are not physical laws. In most research in the sciences, one compares recorded data with a theory that is founded on an analytical expression of physical laws. The success or otherwise of the comparison is a test of the hypothesis of how nature works expressed as a mathematical formula. This might be something fundamental like an inverse square law. Alternatively, fitting a mathematical model to the data might determine physical parameters (such as a refractive index).

On the other hand, where there are no general theories, data mining techniques are valuable, especially where one has large quantities of data containing noisy patterns. This approach hopes to obtain a theoretical generalization automatically from the data by means of indication, deriving empirical models, and learning from examples. The resultant theory, while may not be not fundamental, can yield a good understanding of the physical process and can have great practical utility.

21.6.5 Scientific Applications

In a growing number of domains, the empirical or black box approach of data mining is good science. Three typical examples are:

Sequence Analysis in Bio Informatics

Genetic data such as the nucleotide sequences in genomic DNA are digital. However, experimental data are inherently noisy, making the search for patterns and the matching of subsequences difficult. Machine-learning algorithms such as artificial neural nets and hidden Markov chains are a very attractive way to tackle this computationally demanding problem.

Classification of Astronomical Objects

The thousands of photographic plates that comprise a large survey of the night sky contain around a billion faint objects. Having measured the attributes of each object, the problem is to classify each object as a particular type of star or galaxy. Given the number of features to consider, as well as the huge number of objects, decision tree-learning algorithms have been found accurate and reliable for this task.

Medical Decision Support

Patient records collected for diagnosis and prognosis include symptoms, bodily measurement, and laboratory test results. Machine-learning methods have been applied to a variety of medical domains to improve decision making. Examples are the induction of rules for early diagnosis of rheumatic diseases and neural nets to recognize the clustered microcalcifications in digitized mammograms that can lead to cancer.

The common technique is the use of data instances or cases to generate an empirical algorithm that makes sense to the scientist and that can be put to practical use for recognition or prediction.

21.6.6 Example of Predicting Air Quality

To illustrate the data mining approach, both advantages and disadvantages, this section describes its application to a prediction of urban air pollution.

Motivation

One needs an understanding of the behavior of air pollution in order to predict it and then to guide any action to ameliorate it. Calculations with dynamical models are based on the relevant physics and chemistry.

An interesting research and development project pursuing this approach is DECAIR (Development of an earth observation data converter with application to air quality forecast- http://www-air.inria.fr/decair/). This concerns a generic system for exploiting urban quality models by incorporating land use and could cover data from remote-sensing satellite images.

To help with the design and validation of such models, a complementary approach is described here. It examines on air quality empirically. Data mining and, in particular, machine-learning techniques are employed with two main objectives:

1. To improve our understanding of the relevant factors and their relationships, including the possible discovery of nonobvious features in the data that may suggest better formulations of the physical models;
2. To induce models solely from the data so that dynamical simulations might be compared to them and that they may also have utility, offering (shortterm) predictive power.

Source Data

The investigation uses urban air quality measurements form the City of Cambridge (UK) [Cambridge City Council Air Quality Monitor: http://www.io-ltd.co.uk/ccc.html]. These are especially useful since contemporary weather data from the same location are also available. The objectives are, for example, to look for and interpret possible correlations between each pollutant (NO, NO_2, NO_x, CO, O_3, and PM_{10} particulates) and

a. the other pollutants;
b. the weather (wind strength and direction, temperature, relative humidity and radiance); looking in particular for lags – that is, one attribute seeming to affect another with a delay of perhaps hours or of days.

Data Preparation

Before trying to apply machine learning and constructing a model, there are three quite important stages of data preparation. The data need to be *cleaned*, *explored*, and *transformed*. In typical applications, this can be most of the overall effort involved.

a) Cleaning

Though not elaborated here, but commonly a major part of the KDD process is data cleaning. In this case, one is concerned with imposing consistent formats for dates and times, allowing for missing data, finding duplicated data, and weeding out bad data – the latter are not always obvious. The treatment of missing or erroneous data needs application dependent judgment.

b) Exploration

Another major preliminary stage is a thorough examination of the data to acquire familiarity and understanding. One starts with basis statistics – means, distributions, ranges, etc – aiming to acquire a feeling for data quality. Other techniques such as sorting, database queries, and especially exploratory graphics help one gain confidence with the data.

c) Transformation

The third preparation step is data set sampling, summarization, transformation, and simplification. Working with only a sample of the full data, or applying a level of aggregation, may well yield insight and result that are quicker (if not even discernible at all) than with the complete data source. In addition, transforming the data by defining new variables to work with can be a crucial step. Thus one might, for instance, calculate ratios of observations, normalize them into bins, bands, or classes.

Initial Analysis

The initial analysis concentrated on the daily averages for the weather measurements and daily maxima of the pollutants. This simplifies the problem, the results providing a guide for a later full analysis. In addition, the peak values were further expressed as bands (e.g., "low," "medium," and "high"). For example, ozone (O_3) values were encoded as

LOW	<50 ppb
MEDIUM	50–90 ppb
HIGH	>90 ppb

The bands relate to standards or targets set by the UK Expert Panel on Air Quality Standards (EPAQS) that the public can appreciate. (For ozone, the recommended limit is 50 ppb as an 8-hour running average).

The data exploration and analysis is guided by domain knowledge and enhanced by it, Examination of the Cambridge air pollution data confirmed initial expectations:

- There is a daily cycle with peaks in the afternoon.
- Sundays have low pollution
- An east wind (from industrial Europe) increases ozone levels.
- Sunlight on nitrogen dioxide (NO_2) produces ozone.
- Particulates (PM_{10}) come from vehicle exhausts.

Cambridge has little industry and within an urban environment traffic is the dominant pollution agent. Its effect depends on the local topography, so mesoscale dynamical models have restricted value.

Modeling

The two principal machine-learning techniques used in this application are neural networks and the induction of decision trees. Expressing their predictions as band values make the results of such models easier to understand.

a) Decision Trees

Applying the C5.0 algorithm to the data to generate a simple decision tree, one gets for ozone bands:

AirTemp = <28.3 → LOW
AirTemp > 28.3
RelHum = <58.1 → HIGH
RelHum > 58.1 → MEDIUM

This suggests how the ozone concentration depends mainly on the air temperature and relative humidity. The same tree, expressed as a rule set is:

Rules for HIGH:

Rule #1 for HIGH:

If AirTemp >28.3

And RelHum = <58.1

Then → HIGH

Rules for LOW:

Rule #1 for LOW:

If AirTemp = <28.3

Then → LOW

Rules for MEDIUM:

Rule #1 for MEDIUM:

If AirTemp > 28.3

And RelHum > 58.1

Then → MEDIUM

Default: → LOW

In fact, the support for these rules is modest. The handicap is that there are too few instances of HIGH ozone days into the data. Reliable predictions would need something more elaborate, but this illustrates the idea.

b) Neural Networks

Alternatively, the daily data can be fitted with an artificial neural network to model the ozone band value. A first attempt yields:

Neural Network "O3band" architecture
Input Layer: 5 neurons
Hidden Layer #1: 4 neurons
Output Layer: 4 neurons
Predicted Accuracy: 96%
Relative Importance of Inputs
Air Temp: 0.29
RelHum: 0.06
Rad: 0.04
Wet: 0.02
Windspeed: 0.004

Again, this shows that air temperature is the dominant predictor. However, given the limited quantity of data summarized to daily values, it is not worth trying to refine the model network.

Software

The air quality data were analyzed using the data mining software package *Clementine* (http://www.spss.com/software/clementine/and http://www.isl.co.uk/-originally from Integral Solutions Ltd. and now from SPSS Inc.) While this provides standard machine-learning algorithms to generate models, its great virtue is the powerful visual environment it offers for data exploration. This case of data exploration and modeling is crucial is in allowing the domain expert to attack the problem and find applicable results.

In summary, work so far supports the common experience in data mining that most of the efforts is in data preparation and exploration. The data must be cleaned to allow for missing and bad measurements. Detailed examination leads to transforming the data into more effective forms. The modeling process is very iterative, using statistics and visualization to guide strategy. The temporal dimension with its lagged correlations adds significantly to the search space for the most relevant parameters.

Investigation that is more extensive is needed to establish under what circumstances data mining might be as effective as dynamical modeling. (For instance urban air quality varies greatly from street depending on buildings and traffic.) A feature of data mining is that it can "short circuit" the postinterpretation of the output of numerical simulations by directly predicting the probability of exceeding pollution thresholds. A drawback is the need for large datasets in order to provide enough high-pollution episodes for reliable rule induction. More generally, data mining analysis is useful to provide a reference model in the validation of physically based simulation calculations.

21.7 Data Mining in a Scientific Environment

The advent of the computer has brought with it the ability to generate and store huge amounts of data. For example, it is not unusual for power users to have the equivalent of three or four encyclopedias worth of data online. When you add the data generated by government and other organizations, such as the recently completed census or the data collected every time you make a purchase at any modern supermarket, the volume of data available is almost incomprehensible. The problem is in how to turn this data into usable information.

However, this is not a new phenomenon. Scientists, especially experimentalists, have long had to tackle this problem. While Isaac Newton may have formulated his theory of gravity when an apple fell on his head, it was still followed by hundreds, if not thousands, of experiments demonstrating, validating, and/or refining the original equation. Taking more recent examples, the volume of data generated by space probes and particle physics, dwarfs anything previously contemplated. Looking closer to home, scientists at ANSTO often analyze data generated over their entire working career of twenty or thirty years.

Over the centuries, various methods have been developed to deal with this volume of data, many of which were seen as major steps forward for mathematics at the time. Some of these methods include fast Fourier transforms, multivariate regression analyses, as well as a whole range of statistical methods. More recently, visualization has been widely adopted by scientists as a means of studying the ever-growing masses of data.

21.7.1 What is Data Mining?

With the current trends in centralization of an organization's data in large databases, particularly in a commercial environment, the process of extracting useful information has become more formalized and the term **Data Mining** has been coined for it. In one of the first papers on commercial data mining, Evangelos Simoudis of IBM defined it as:

The process of extracting previously unknown, comprehensible and actionable information from large databases and using it to make crucial business decisions (Simoudis 1996).

This definition has a definite business flavor and much of IBM's development of data mining has been in this direction. In practice, data mining is a process that can take on different approaches depending on the type of data involved and the objectives desired. As this is still very much an evolving discipline, much work is being undertaken to determine standard processes for the varied environments. Further, as the context in which the data is gathered is often an important component, this must be factored into any analysis.

Data mining consists of three components: the captured data, which must be integrated into organization-wide views, often in a *Data Warehouse*; the

mining of this warehouse; and the organization and presentation of this *mined* information to enable understanding.

The data capture is a fairly standard process of gathering, organizing, and cleaning up; for example, removing duplicates, deriving missing values where possible, establishing derived attributes and validation of the data. Much of the following processes assume the validity and integrity of the data within the warehouse. These processes are no different to any data gathering and management exercise.

The *Data Mining* process itself is the extraction of valid and previously unknown information, as given in the definition above. There are two approaches: *verification driven*, whose aim is to validate a hypothesis postulated by a user, or *discovery driven*, which is the automatic discovery of information by the use of appropriate tools.

The data mining process is not a simple function, as it often involves a variety of feedback loops since while applying a particular technique, the user may determine that the selected data is of poor quality or that the applied techniques did not produce the results of the expected quality. In such cases, the user has to repeat and refine earlier steps, possibly even restarting the entire process from the beginning. This is best illustrated in Figure 21.5 from Simoudis article.

The *final* step is the presentation of the information in a format suitable for interpretation. This may be anything from a simple tabulation to video production and presentation. The techniques applied depend on the type of data and the audience as, for example, what may be suitable for a knowledgeable group would not be reasonable for a naive audience.

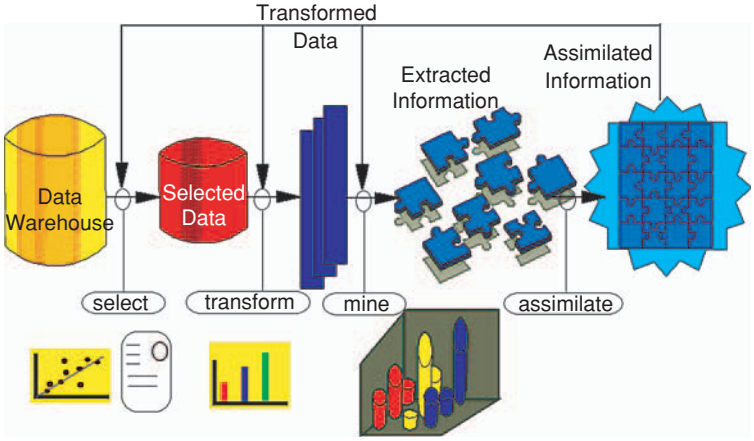

Fig. 21.5. The Data Mining Process

Verification Driven Data Mining

Currently, the most common use of data mining is verification driven and is primarily aimed at confirmation of an idea. Generally, the mechanism is to propose some association or pattern and then to study the data to find support, or otherwise, for the proposal.

There are a number of standard techniques used in verification driven mining; these include the most basic form of query and reporting, presenting the output in graphical, tabular, and textual forms, through to multidimensional analysis and on to statistical analysis.

Discovery-driven Data Mining

The discovery-driven approach depends on a much more sophisticated and structured search of the data for associations, patterns, rules, or functions, and then having the analyst review them for value. The current techniques for performing discovery-driven mining consist of four different approaches: predictive modeling, including neural nets, link-analysis technique that attempts to establish links between records, database segmentation that partitions the data into collections of related records, and finally deviation detection that identifies point that do not fit in a segment.

21.7.2 Traditional Uses of Data Mining

Within the business world, data mining is being seen as a method of tapping into the value of the data with an organization and providing a competitive advantage. An example of this is the analysis of purchase histories, drawn from credit card transactions, preferred customer schemes, frequent shopper schemes, and any other purchasing data that includes customer information. Using a method called *neural segmentation*, a number of different types of purchase patterns can be identified and then customer groupings can be associated with this data.

For instance, such analysis of shopping has identified two groups of people who purchase baking items, the first being older, retired couples, and the second, young couples with large families. The next step may be to look at product linkage; for example, there may be a group of people who purchase men's suits, women's high-fashion shoes, men's ties and expensive chocolates. They do not buy baby clothes, houseware, and greeting cards. This indicates that a store may be able to bring in more customers for a sale of suits if they have chocolates for half-price, or better yet, give away the chocolates.

These procedures can be used further for the analysis of any activity that generates large volumes of data, from specific surveys through to the collection of operational data, such as stock movements, or point-of-sale information. An example of this is *Market Basket Analysis*, which refers to the discovery of patterns within items purchased as is illustrated by such correlations between

the purchase of paint and paint brushes or paint thinner. These associations can then be used to determine shelf locations and promotional sales planning.

Such analysis is the main force driving the introduction of data mining within large organizations and, thus, the current interest in such research. It is invariably related to the interrogation of large volumes of data, using high-performance systems and massive amounts of storage. However, there is still the need to apply some common sense to the results as spurious patterns and associations may be found. It is quite possible for an association to be found between the purchase of paint and cat food, which may be caused by other factors that were not part of the original analysis.

Most commonly, data mining is a single step in the entire process of *Decision Support*, and fits into the general process: *Data Warehouse - Data Mining - Decision Support.*

21.7.3 Data Mining in a Scientific Environment

While IBM may be driving data mining in the commercial marketplace, the origins are in fact in scientific computing with considerable work being done at UCLA and the University of Helsinki. Some of the original work was on geophysical databases in an attempt to process some of the large volumes of data they have available.

What is not considered in much of the work on data mining is that most, if not all, of this work is just as applicable to the scientific environment. One of the critical issues with data mining is a *credibility check* being performed by someone who is aware of the field. Most scientists, and in particular experimentalists, have a great respect for their data, being well aware of the dangers of using inapplicable methods for analysis. An excellent example of this is given in Clifford Stoll's new book *Silicon Snake Oil*, in which he describes a study by an astronomer, Professor Li Fang, into the periodic motions of the earth's axis. This study involved the analysis of thousands of years of astronomical measurements. Dr Li had performed all the measurements by hand and Clifford was attempting to show him how easy it would have been with a computer. On presenting his results, Dr Li replied:

When I compare the computer's results to my own, I see that an error has crept in. I suspect it is from the computer's assumption that our data is perfectly sampled throughout history. Such is not the case, especially during the Sung dynasty. And so, it may be that we need to analyze the data in a slightly different manner.

Having a computer, I had naturally cast the problem as simple data analysis.... The real challenge was understanding the data and finding a good way to use it (Stoll 1995).

The underlying principles of the science method, being the cycle *observation-hypothesis-experiment* fits well with the processes of data mining with discovery driven mining working well for the *observation-hypothesis* step and the verification driven mining for the *hypothesis-experiment* step.

As scientists have been working with this principle for centuries, and as most mathematics has been intended to support such scientific endeavors, many, if not all, of the methods are already being used by them. In many cases, the only change is in the terminology, not in the practice.

The final stage in any data mining is the presentation of results and this has both a very long history and is an area of rapid change in scientific work. This stems from simple graphs that scientists have long studied through to the latest techniques in visualization being demonstrated on high-performance graphics workstations.

21.7.4 Examples of Scientific Data Mining

One example of the scientific analysis of such data found in farming and the environment, is optimization of crop yield while minimizing the resources supplied. To minimize the resources, it is necessary to identify what factors affect the crop yield, out of such items as chemical fertilizers and additives; for example, phosphate, the moisture content and type of the soil.

One analysis looked at over 64 separate items measured over a number of years to extract the items that were significant. Initially the analysis was discovery-driven mining to attempt to find what parameters were significant, either by themselves or in conjunction with others. Using such statistical methods as multivariate regression analysis, the parameters that are significant and their relative influence was determined. From this, an equation was developed, which was then further verified through verification driven mining against new datasets.

Of more general interest, global climate change studies, a *hot* research area, are primarily a verification-driven mining exercise. Climate data has been collected for many centuries and is being extended into the more distant past through such activities as analysis of ice core samples from the Antarctic and, at the same time, a number of different predictive models have been proposed for future climatic conditions. The sample data is then used to verify these models by seeing if they accurately predict past conditions, which can then be compared with the sample data. From this, the models are then further refined and used for another round of verification driven mining.

21.7.5 Concluding Remarks

Data mining is a new term and formalism for a process that has been undertaken by scientists for generations. The massive increase in the volume of data collected or generated for analysis with the use of computers has made it an essential tool. However, despite the more *formal* approach, data mining is something that scientists perform on an *ad hoc* basis and can easily adapt to. Many of the methods used for the analysis of the data were originally developed to process scientific data and are used unchanged.

21.8 Flexible Earth Science Data Mining System Architecture

Designing a data mining system for Earth Science applications is complex and challenging. The issues that need to be addressed in the design are (1) variability of data sets, (2) operations for extracting information, and (3) providing the capability to the user to write complex mining plans. Earth Science data sets not only come in different formats, types, and structures, there are also many different states of processing such as raw data, calibrated data, validated data, derived data, or interpreted data. The mining system architecture must be designed to be flexible to handle these variations in data sets. The operations required in the mining system vary for different application areas within Earth Science. Operations could range from general purpose operations such as image processing techniques or statistical analysis to highly specialized, data set-specific science algorithms. The mining system architecture should be flexible in its ability to process new data sets and incorporate new operations without too much effort. The design of the architecture should also allow other users to build new clients to utilize such a system. The Information Technology and Systems Center at the University of Alabama in Huntsville originally developed the Algorithm Development and Mining (ADaM) system under a research grant from NASA Headquarters Research Announcement (NRA) to investigate new methods of processing large volumes of Earth Observing System (EOS) remote-sensing data sets. ADaM is designed to handle the complexity of mining Earth Science data. It can process heterogeneous data sets and allows users to add research problem-specific science algorithms to the system.

This section discusses issues that had to be considered in designing a flexible system architecture. It describes the ADaM system and its user interface as an example of a flexible design. This section also describes the research directions that are evolving from this innovative architecture.

21.8.1 DESIGN ISSUES

As stated in the introduction, several issues had to be considered while designing a flexible mining system for Earth Science. These are:

Data Handling Capabilities

Earth Science data introduces complexity in designing, building, and utilizing a data mining system, because these data sets can be quite varied. They can be point data collected by a meteorological instrument, swath or grid data collected by satellites, or volume scan data collected by weather radar. The formats of these data sets also vary from simple binary or ASCII files to more complex structures such as Hierarchical Data Format for the Earth

Observing System (HDF-EOS). The spatial and the temporal resolutions of these data sets depend upon the measuring instrument and the platform. The spatial resolution could vary from hundreds of kilometers to a few meters. The temporal range of a data file could vary from 15 minutes to a day or longer. Temporal resolution could vary from instantaneous measurements to accumulation of data over some period. To utilize mining techniques over the broad range of data sets, the mining system had to be designed to handle these types of data set variations.

Addition of New Algorithms

In certain circumstances, a known scientific algorithm can be utilized to extract the information needed from data sets. Detecting Mesoscale Convective Systems (MCS) from SSM/I data utilizing the Devlin algorithm is one such example. The data mining system had to be designed to be flexible enough to allow not only data set specific algorithms but also other new algorithms to be added to it without affecting the other operations.

Allow Scientists to Select and Sequence Different Operations

The mining system also needed the capability to allow scientists to create their own mining plans. A mining plan is a sequence of specified steps, where each step is a processing operation. The scientist should be able to piece together different operations/algorithms to reach their goal.

21.8.2 ADaM System Features

The ADaM system was designed using the latest object Oriented techniques to achieve a high degree of portability, accessibility, and modularity. The implementation in standard C++ allows the system to run on multiple operating systems, including IRIX, Linux, and Microsoft Windows NT. One of the design goals was to have ADaM work at both data archive centers or on a user's desktop workstation.

Overview of the Architecture

The ADaM data mining system has been designed to extract content based metadata from large Earth Science data archives. It can detect phenomena or events that are of interest to scientists and then store this information in a way that facilitates the data search and order process. Some mining results are stored in Event/Relationship Search System (E/RSS), an ITSC-developed spatial data search engine used to find coincidences between mining-generated phenomena, climatological events, and static information such as country and river basin boundaries. The data mining engine also provides

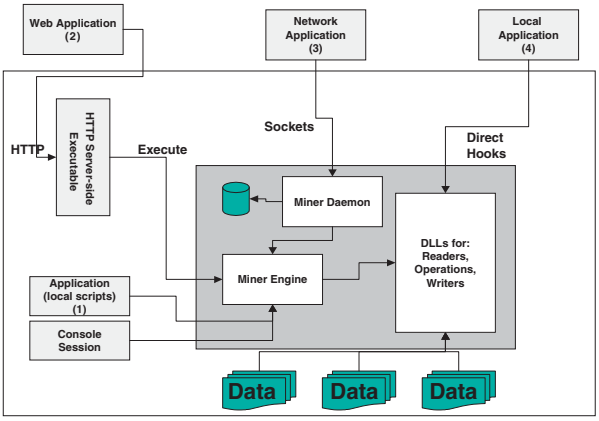

Fig. 21.6. Multiple Process Flows Utilizing the ADaM Data Mining Architecture

other data ordering-related capabilities such as subsetting and custom data product generation through specialized client applications. Custom processing may include gridding, resampling, filtering, format conversion, or other analysis depending on the needs of the customers. For example, ADaM can generate a monthly total rain accumulation image from radar reflectivity data. Both the E/RSS and custom processing client are Web applications, so the clients are capable of running in almost any environment. Figure 21.6 depicts a generalized view of how the ADaM data mining architecture has been utilized. This architecture allows the clients to communicate to the system in a variety of ways such as: (1) The miner engine can be driven directly via local scripts or an interactive console session, (2) A Web application can guide the user in creating mining plans, which execute the mining engine, (3) A network application can submit mining plans via the miner daemon, and (4) The system may also be used as a library with the application directly linking to the individual operations needed.

Processing Flow

The ADaM system architecture is based on a processing stream, in that mining is broken down into a series of steps with results from each step passed to the next one in line. Figure 21.7 illustrates both ADaM's data processing stream, as well as the three basic types of modules: input, processing, and output. The use of data input filters, specialized for a variety of data types, has been instrumental in simplifying the development of the processing and output operations. The selected input filter translates the data into a common internal structure so that the processing operations can all be written for a single data representation.

This allows the addition of new operations to the system without having to address input data format problems. Similarly, the addition of a new input

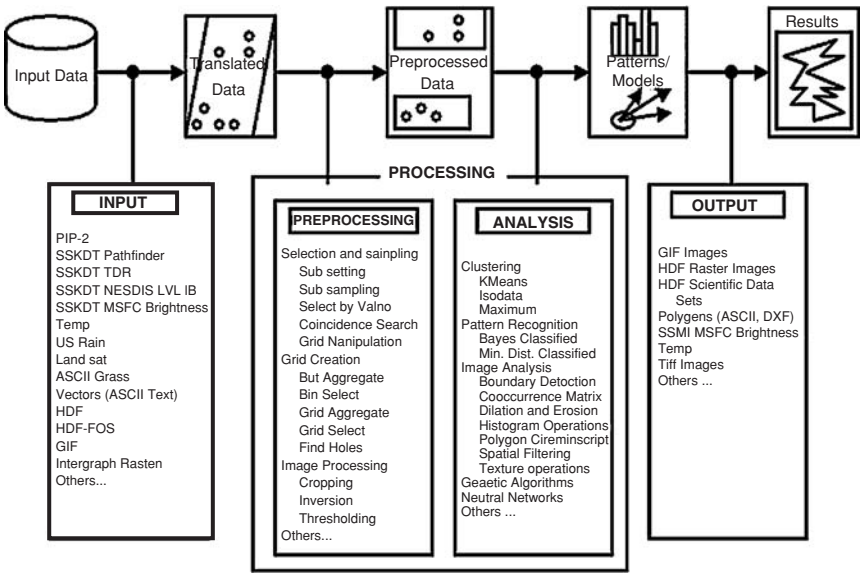

Fig. 21.7. Schematic diagram depicting the stream of a mining system

filter provides access to the entire suite of processing operations for the data type in question. This design feature allows ADaM to handle heterogeneous Earth Science data sets. The mining system currently allows over 120 different operations to be performed on the input data stream. These operations vary from specialized atmospheric science data set specific algorithms to generalized image processing techniques. The last step in the mining process is the selection of the input modules, the output filters effectively insulate the processing operations from having to support all the possible output formats. Since the input data has been converted to ADaM's internal format, the output modules allow the user the option to select either the input format or a different format for the final data product.

Components

In order to allow for the distributed use of the data mining functionality, the ADaM system was designed as a client-server architecture, which supports remote client applications communicating with the data mining server. This allows the server system to be co-located with archived data stores while being driven by either remote or local clients. In support of this architecture, the ADaM data mining system is composed of the mining engine and mining daemon, both located on the server. The daemon supports a specific protocol of messages and listens on a configured port for instructions from client applications.

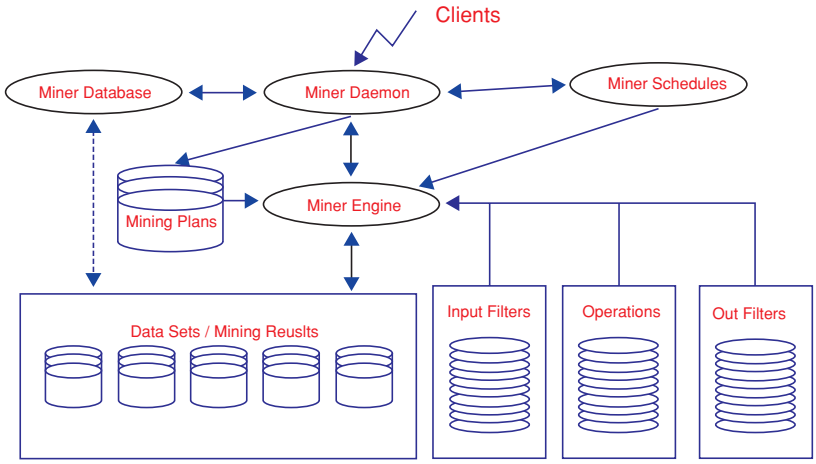

Fig. 21.8. AdaM Data Mining Server Components

Through instructions from the client, the daemon is responsible for managing user access information, file management operations, and job scheduling and management. The daemon ultimately sends the correct information and directions to the engine in the form of a "mining plan" for actual processing. A software interface layer was created providing tools to assist client application developers in communicating with the mining daemon across network sockets. Figure 21.8 depicts the connections between the components of the ADaM client-server architecture. Each component performs a specific well-defined task, and therefore the components themselves may be replaced or updated provided that the new components conform to the same interfaces.

Mining Engine:

The mining engine is the software component that manages the processing of data through a series of specified operations. The input, processing, and output modules are dynamically loaded as needed at execution time, and this allows for the addition of newly developed modules without the need to rebuild the engine. The mining engine interprets a mining plan script that provides the details about each specified operation and the order that they should be executed. Other communication with the mining engine is managed through the mining daemon process.

Mining Daemon:

The mining daemon is the gateway to the mining engine. All network communications with the mining system are handled by the daemon through a message handling protocol. Upon installation the daemon is configured to listen on a specific port for any socket communications. The daemon is capable

of handling a fairly rich set of messages that allows it to perform file management duties, command the mining engine, and provide user security screening. The daemon can also determine at run time, which processing modules are available on the server.

Mining Database:

The database component is used to store information that is required for the smooth operation of the system and the interaction of its components. This information includes the names, locations, and related metadata for input data sets available on the server. It also includes information about users, jobs, mining results, and other related information. A relational database is currently used for this task. Access to the database is provided by the daemon.

Mining Scheduler:

The scheduler component examines the list of jobs to be executed on the server and determines which job or jobs to execute at any given time. The scheduling policy used can be unique to each server. The scheduler invokes the mining engine for each job and monitors its progress, updating the job status in the database whenever it changes.

Operations and Data set Input/Output Filters:

Each of the operations and data set filters is implemented as a shared library. The libraries are loaded dynamically by the mining engine, which means that new modules may be added to the system without recompiling or relinking. Each of the operations and filters is completely independent of all the others. All operations and filters either produce or operate on a common format representing scientific data. This design feature allows science specific algorithms to be incorporated into the system with relative ease.

Mining Plan:

The mining plan script conveys the processing instructions to the mining engine. The plan contains the number and sequence of processing steps as well as the detailed parameters (tokens) describing how to perform each step, such as where to find the input data, where to store the output and configuration parameters for all the various operations. Mining plans may be created using the mining plan editor. Since mining plans are text files, they may also be created using any text editor. It is easy for applications to write mining plans. The mining plan begins with a number indicating the number of operations in the plan. The remainder of the plan is a series of token/value pairs where the tokens and values are delimited by newlines.

21.8.3 ADaM Plan Builder Client

In order to allow users to build complex mining plans, ITSC has designed an easy-to-use and functional user interface called the ADaM Plan Builder. This user interface is a client that communicates directly to the mining engine. It makes it easier for the user to select the right operation for the task and to provide values for the parameters for that operation. The individual ADaM operation documentation is written in XML™. Since this standardizes the documentation, the Plan Builder written in Java parses and utilizes the information contained in those XML files. Thus, the Plan Builder Interface utilizes these XML files to provide the user options on the operations available, what parameters each operation requires, the meaning of each parameter, default values for those parameters and finally a sample mining plan. Through the Plan Builder, users can select sample operation steps and modify the values for the parameters according to their needs. The ADaM Plan Builder allows the user to chain together complex mining plans for scientific research.

The Plan Builder also allows the users to edit and modify the mining database. The user can feed the metadata information about the data files to be mined into the database via this client. The database then automatically selects the correct files for mining based on the time range given. The architecture of the Plan Builder is shown in Figure 21.9 and a screen capture of the interface is shown in Figure 21.10.

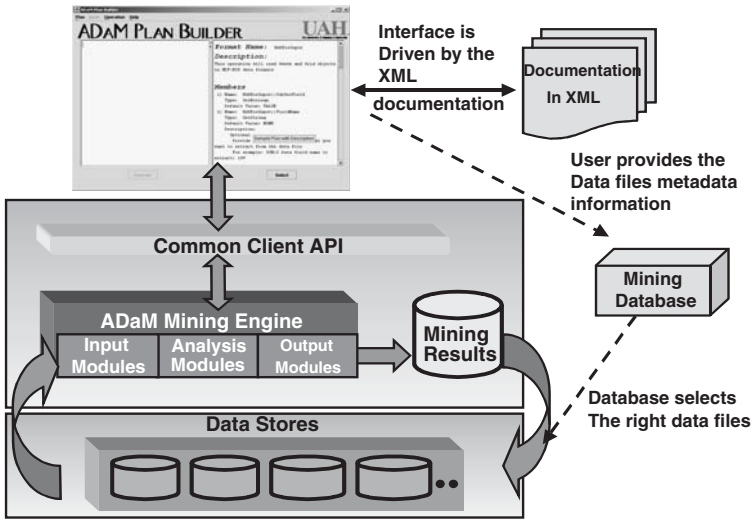

Fig. 21.9. Screen capture of the ADaM Plan Builder Interface

21.8.4 Research Directions

ADaM is currently undergoing a metamorphosis in the sense of becoming uncoupled from an environment that is dependent on centralized processing on a single server platform with the availability of local data. The following sections describe some of the efforts underway to migrate ADaM into a highly distributed environment that will provide broader access to the system and distributed heterogeneous scientific data sets, while addressing improved scalability and flexibility.

Distributed Mining

ITSC is currently investigating and prototyping emerging distributed component technologies. To address the use of distributed mining services and access to distributed data sets, the use of distributed mining services opens the system to greater possibilities of extensibility, performance, scalability, and reliability by distributing the processing burden and lessening the possibility of centralized points of system failure. Current research efforts have been successful with the development of an Earth Science Markup Language [http://esml.itsc.uah.edu] that will make great strides towards realizing generic access to heterogeneous data sets. The integration of ESML technology with planned distributed mining components is expected to result in a virtual processing environment that capitalizes on improved networking bandwidth and underutilized distributed processors.

Grid Mining

Another approach to distributed mining is also being prototyped in the form of grid mining. ITSC researchers, in collaboration with NASA/Ames researchers, have been successful with implementing and testing the ADaM system on the NASA Information Power Grid. The Grid approach employs a sophisticated infrastructure of message passing, scheduling, and security in an effort to utilize large capacity processing and data centers for scientific research. This approach to distributed data mining promises to be of particular benefit to scientific researchers in need of massive processing and data resources.

Mining Onboard Space Craft

ITSC is also investigating and developing an innovative processing system capable of handling the unique constraints and characteristics of the on-board satellite data and information environment. The EnVironmEnt for On-Board Processing (EVE) system will serve as a proof of concept of advanced information systems technology for remote-sensing platforms. EVE's on-board, real-time processing will provide capabilities focused on the areas of autonomous data mining, classification, and feature extraction. These will contribute to

Earth Science research applications, including natural hazard detection and prediction, fusion of multisensor measurements, intelligent sensor control, and the generation of customized data products for direct distribution to users. EVE is being engineered to provide high-performance data processing in a real-time operational environment. A ground-based test bed is being created to provide testing of EVE and associated Earth Science applications in a heterogeneous embedded hardware and software environment.

Concluding Remarks

The ability of the user to create complex mining plans by chaining together different operations is also possible because of the flexibility of the architecture. ITSC plans to utilize its experience in designing ADaM to meet the scientific mining requirements of the next generation of scientists in several other domains. Research efforts are focused toward distributed mining across the Web; mining large volumes of data on the information power grid; and finally designing a system that would be used onboard aircraft or spacecraft to extract features or phenomena as soon as they are sensed by the instrument.

21.9 Summary

An architecture was proposed that uses assistance of data mining to provide decision support for diagnoses of fever. Although we have dealt exclusively with the diagnoses of fever in this section the basic architecture can be implemented for the diagnoses of any other physical condition. The system we proposed is meant to make the process of fever diagnosis less time consuming and possibly more accurate. It must be emphasized that the system would serve only as an accessory and not as a substitute for the human physician. The main difference between a usual expert system is that the data mining process enhances the learning process. It must also be noted that the diagnosing power is never complete. Its knowledge could become better than its current knowledge in several years or months or weeks or days or even in an hour.

Research to expand the types and magnitude of data that data mining systems can effectively mine is well underway. The needs of business, security, and science will provide incentive to invest time and money into such development. Perhaps someday data mining will advance faster than the growth of databases and allow the mining of nearly infinite databases, such as mining the entire World Wide Web.

Is data mining as useful in science as in commerce? Certainly, data mining in science has much in common with that for business data. One difference, though, is that there is a lot of existing scientific theory and knowledge. Hence, there is less chance of knowledge emerging purely from data. However, empirical results can be valuable in science (especially where it borders on

engineering) as in suggesting causality relationships or for modeling complex phenomena.

Another difference is that in commerce, rules are soft sociological or cultural and assume consistent behavior. For example, the plausible myth that *"30% of people who buy babies' nappies also buy beer"* is hardly fundamental, but one might profitably apply it as a selling tactic (until perhaps the fashion changes from beer to lager).

On the other hand, scientific rules or laws are, in principle, testable objectively. Any results from data mining techniques must sit within the existing domain knowledge. Hence, the involvement of a domain expert is crucial to the data mining process. Naïve data mining often yields "obvious" results. The challenge is to incorporate rules known a priori into the empirical induction, remembering that the whole KDD process is exploratory and iterative.

ADaM has proven to be an effective and valuable tool to mine Earth Science spatial data. Its flexible architecture design has made it possible for ADaM to handle the multiple formats, scales, resolutions, and large granule sizes typical of spatial data for many different science problems. The design permits the easy addition of new algorithms, especially domain-specific science algorithms. Thus this section discussed the various case studies involved in science and biomedicine with respective illustrations.

As a final point, the biggest of all, the Internet, is becoming more and more important, and while there is useful information, extracting that from the terabytes being added daily is an enormous task. The techniques of data mining are applicable here more than any other domain. However, to make use of it takes time, effort and, above all, people with a knowledge of the field, to differentiate the true solutions from the infeasible.

21.10 Review Questions

1. How does data mining help in clinical and medical diagnosis problems>
2. How does data mining contribute substantially to DNA analysis?
3. With a case study explain the approach of an unsupervised neural network to medical data mining techniques.
4. Explain in detail how data mining is used for fever diagnosis and its application in science.
5. Define data management context as applicable to mining.
6. Explain with example, how data mining is used in scientific applications.
7. What are the traditional uses of data mining?
8. Give the examples of scientific data mining.
9. Explain how data mining is used in scientific environment with an example.
10. Explain the AdaM system features with its architecture and its components.
11. Write a short note on AdaM plan builder client.