8

Data Mining in Hospital Information System

Jing-song Li, Hai-yan Yu and Xiao-guang Zhang
Zhejiang University,
China

1. Introduction

Data mining aims at discovering novel, interesting and useful knowledge from databases. Conventionally, the data is analyzed manually. Many hidden and potentially useful relationships may not be recognized by the analyst. Nowadays, many organizations including modern hospitals are capable of generating and collecting a huge amount of data. This explosive growth of data requires an automated way to extract useful knowledge. Thus, medical domain is a major area for applying data mining. Through data mining, we can extract interesting knowledge and regularities. The discovered knowledge can then be applied in the corresponding field to increase the working efficiency and improve the quality of decision making.

This chapter introduces the applications of data mining in HIS (Hospital Information System). It will be presented in three aspects: data mining fundamentals (part 1 and part 2), tools for knowledge discovery (part 3 and part 4), and advanced data mining techniques (part 5 and part 6). In order to help readers understand more intuitively and intensively, some case studies will be given in advanced data mining techniques.

2. Part 1 Overview of data mining process

Nowadays, data stored in medical databases are growing in an increasingly rapid way. Analyzing that data is crucial for medical decision making and management. It has been widely recognized that medical data analysis can lead to an enhancement of health care by improving the performance of patient management tasks. There are two main aspects that define the need for medical data analysis.

1. Support of specific knowledge-based problem solving activities through the analysis of patients' raw data collected in monitoring.
2. Discovery of new knowledge that can be extracted through the analysis of representative collections of example cases, described by symbolic or numeric descriptors.

For these purposes, the increase in database size makes traditional manual data analysis to be insufficient. To fill this gap, new research fields such as knowledge discovery in databases (KDD) have rapidly grown in recent years. KDD is concerned with the efficient computer-aided acquisition of useful knowledge from large sets of data. The main step in the knowledge discovery process, called data mining, deals with the problem of finding interesting regularities and patterns in data.

A simple data mining process model mainly includes 6 steps:

1. Assembling the data

Data mining requires access to data. The data may be represented as volumes of records in several database files or the data may contain only a few hundred records in a single file. A common misconception is that in order to build an effective model a data mining algorithm must be presented with thousands or millions of instances. In fact, most data mining tools work best with a few hundred or a few thousand pertinent records. Therefore once a problem has been defined, a first step in the data mining process is to extract or assemble a relevant subset of data for processing. Many times this first step requires a great amount of human time and effort. As in healthcare industry, we need domain experts such as doctors, nurses, hospital managers and so on to work closely with the data mining expert to develop analyses that are relevant to clinical decision making. There are three common ways to access data for data mining:

1. Data can be accessed from a data warehouse.
2. Data can be accessed from a database.
3. Data can be accessed from a flat file or spreadsheet.

Because medical data are collected on human subjects, there is an enormous ethical and legal tradition designed to prevent the abuse of patients' information and misuse of their data. In data assembling process, we should pay more attention to the five major points:

- Data ownership
- Fear of lawsuits
- Privacy and security of human data
- Expected benefits
- Administrative issues

2. The data warehouse

A common scenario for data assembly shows data originating in one or more operational database. Operational databases are transaction-based and frequently designed using the relational database model. An operational database fixed on the relational model will contain several normalized tables. The tables have been normalized to reduce redundancy and promote quick access to individual records. For example, a specific customer might have data appearing in several relational tables where each table views the customer from a different perspective. But medical data is almost the most heterogeneous data which contains images like SPECT, signals like ECG, clinical information like temperature, cholesterol levels, urinalysis data, etc. as well as the physician's interpretation written in unstructured texts. Sometimes the relational database model can't describe the heterogeneous data with tables and we can use post-relational database model.

The data warehouse is a historical database designed for decision support rather than transaction processing. Thus only data useful for decision support is extracted from the operational environment and entered into the warehouse database. Data transfer from the operational database to the warehouse is an ongoing process usually accomplished on a daily basis after the close of the regular business day. Before each data item enters the warehouse, the item is time-stamped, transformed as necessary, and checked for errors. The transfer process can be complex, especially when several operational databases are involved. Once entered, the records in the data warehouse become read-only and are subject to change only under special conditions.

A data warehouse stores all data relating to the same subject (such as a customer) in the same table. This distinguishes the data warehouse from an operational database, which stores information so as to optimize transaction processing. Because the data warehouse is

subject-oriented rather than transaction-oriented, the data will contain redundancies. It is the redundancy stored in a data warehouse that is used by data mining algorithms to develop patterns representing discovered knowledge.

3. Relational database and flat files

If a data warehouse does not exist, you can make use of a database query language to write one or more queries to create a table suitable for data mining. Whether data is being extracted for mining from the data warehouse or the data extraction is via a query language, you will probably need a utility program to convert extracted data to the format required by the chosen data mining tool. Finally, if a database structure to store the data has not been designed, and the amount of collected data is minimal, the data will likely be stored in a flat file or spreadsheet.

4. Mining the data

Prior to giving the data to a data mining tool, preprocessing of the data is necessary. Preprocessing the data includes multiple steps to assure the highest possible data quality, thus efforts are made to detect and remove errors, resolve data redundancies, and taking into account of the patient privacy, to remove patient identifiers. Data are analyzed using both statistical and data mining methods to produce information; output formats will vary depending upon the method used. Predictive modeling efforts are iterative, thus statistical and data mining results are repeated with different permutations until the best results (metrics) are obtained.

Patients and health care consumers are increasingly concerned about the privacy of their personal health information. All data mining should carefully attempt to create completely anonymous data before analyses are begun.

Anticipating that data will be 100% complete and error free is unrealistic when working with patient data which collected in complex health care systems. Cleaning the data is proved a nontrivial and tedious task. Data error identification is both an automated and a manual process, and required an iterative procedure that drew upon expertise from the clinical experts as well as statistical experts and the data warehouse engineer. Errors that detected out-of-range values (for example, a systolic blood pressure of 700) are identified by the clinical experts and eliminated from the research data sets. Errors where a variable included inconsistently recorded text require an iterative extraction and programming solution; clinical experts reviewe the text extraction and provide guidelines for converting data for consistency, coding, or deleting the variable if data conversion is not possible.

Medical data is often very high dimensional. Depending upon the use, some data dimensions might be more relevant than others. In processing medical data, choosing the optimal subset of features is such important, not only to reduce the processing cost but also to improve the usefulness of the model built from the selected data. So before the step of mining, we have several choices to make.

1. Should learning be supervised or unsupervised?
2. Which instances in the assembled data will be used for building the model and which instances will test the model?
3. Which attributes will be selected from the list of available attributes?
4. Data mining tools require the user to specify one or more learning parameters. What parameter settings should be used to build a model to best represent the data?

5. Interpreting the results

Result interpretation requires us to examine the output of our data mining tool to determine

if what has been discovered is both useful and interesting. If the results are less than optimal we can repeat the data mining step using new attributes and/ or instances. Alternatively, we may decide to return to the data warehouse and repeat the data extraction process.

As most medical datasets are large and complex, only those models that are validated by experts are retained in the knowledge base for system testing and verification. There are several techniques to help us make decisions about whether a specific model is useful (Evaluating Performance):

- Evaluating supervised learner models
- Two-class error analysis
- Evaluating numeric output
- Comparing models by measuring lift
- Unsupervised model evaluation

For example, if we use several fuzzy modeling methods to process medical data. When interpreting the results, concerning only the accuracy values might be misleading and not revealing other important information, as demonstrated by Cios and Moore. To double check seven other performance measures including sensitivity (a.k.a. recall in information retrieval community), specificity, precision, class weighted accuracy, F-measure, geometric mean of accuracies, and area under the receiver operating characteristics (ROC) curve were also needed to be computed for the top rank result obtained for each dataset.

Medical data mining using some fuzzy modeling methods without or with the use of some feature selection method. Belacel and Boulassel developed a supervised fuzzy classification procedure, called PROAFTN, and applied it to assist diagnosis of three clinical entities namely acute leukaemia, astrocytic, and bladder tumors. By dividing the Wisconsin breast cancer data (the version with 10 features) into 2/ 3 for training and 1/ 3 for testing, test accuracy of 97.9% was reported. Seker et al. used the fuzzy KNN classifier to provide a certainty degree for prognostic decision and assessment of the markers, and they reported that the fuzzy KNN classifier produced a more reliable prognostic marker model than the logistic regression and multilayer feedforward backpropagation models. Ruiz-Gomez et al. showed the capabilities of two fuzzy modeling approaches, ANFIS and another one that performs least squares identification and automatic rule generation by minimizing an error index, for the prediction of future cases of acquired immune deficiency syndrome.

6. Result application

Our ultimate goal is to apply what has been discovered to new situations. Data mining methods offer solutions to help manage data and information overload and build knowledge for information systems and decision support in nursing and health care. For instance, we can build nursing knowledge by discovering important linkages between clinical data, nursing interventions, and patient outcomes.

Applying data mining techniques enhances the creation of untapped useful knowledge from large medical datasets. The increasing use of these techniques can be observed in healthcare applications that support decision making, e.g., in patient and treatment outcomes; in healthcare delivery quality; in the development of clinical guidelines and the allocation of medical resources; and in the identification of drug therapeutic or adverse effect associations. Recent studies using data mining techniques to investigate cancer have focused on feature extraction from diagnostic images to detect and classify, for example, breast cancers.

3. Part 2 Techniques of data mining

Health care now collects data in gigabytes per hour volume. Data mining can help with data reduction, exploration, and hypothesis formulation to find new patterns and information in data that surpass human information processing limitations. There is a proliferation of reports and articles that apply data mining and KDD to a wide variety of health care problems and clinical domains and includes diverse projects related to cardiology, cancer, diabetes, finding medication errors, and many others.

Over the past two decades, it is clear that we have been able to develop systems that collect massive amounts of data in health care, but now what do we do with it? Data mining methods use powerful computer software tools and large clinical databases, sometimes in the form of data repositories and data warehouses, to detect patterns in data. Within data mining methodologies, one may select from an extensive array of techniques that include, among many others, classification, clustering, and association rules.

3.1 Classification

Classification maps data into predefined groups or classes. It is often referred to as supervised learning because the classes are determined before examining the data. Classification algorithms require that the classes be defined based on data attribute values. They often describe these classes by looking at the characteristics of data already known to belong to the classes. Pattern recognition is a type of classification where an input pattern is classified into one of several classes based on its similarity to these predefined classes.

One of the applications of classification in health care is the automatic categorization of medical images. Categorization of medical images means selecting the appropriate class for a given image out of a set of pre-defined categories. This is an important step for data mining and content-based image retrieval (CBIR).

There are several areas of application for CBIR systems. For instance, biomedical informatics compiles large image databases. In particular, medical imagery is increasingly acquired, transferred, and stored digitally. In large hospitals, several terabytes of data need to be managed each year. However, picture archiving and communication systems (PACS) still provide access to the image data by alphanumerical description and textual meta information. This also holds for digital systems compliant with the Digital Imaging and Communications in Medicine (DICOM) protocol. Therefore, integrating CBIR into medicine is expected to significantly improve the quality of patient care.

Another application is constructing predictive model from severe trauma patient's data. In management of severe trauma patients, trauma surgeons need to decide which patients are eligible for damage control. Such decision may be supported by utilizing models that predict the patient's outcome. To induce the predictive models, classification trees derived from a commonly-known ID3 recursive partitioning algorithm can be used. The basic idea of ID3 is to partition the patients into ever smaller groups until creating the groups with all patients corresponding to the same class (e.g. survives, does not survive). To avoid overfitting, a simple pruning criterion is used to stop the induction when the sample size for a node falls under the prescribed number of examples or when a sufficient proportion of a subgroup has the same output.

From the expert's perspective, classification tree is a reasonable model for outcome prediction. It is based on the important representatives from two of the most important groups of factors, which affect the outcome, coagulopathy and acidosis. The two mentioned

features, together with body temperature, are the three that best determine the patient's outcome.

3.2 Clustering

Clustering is similar to classification except that the groups are not predefined, but rather defined by the data alone. Clustering is alternatively referred to as unsupervised learning or segmentation. It can be thought of as partitioning or segmenting the data into groups that might or might not be disjointed. The clustering is usually accomplished by determining the similarity among the data on predefined attributes. The most similar data are grouped into clusters.

Cluster analysis is a clustering method for gathering observation points into clusters or groups to make (1) each observation point in the group similar, that is, cluster elements are of the same nature or close to certain characteristics; (2) observation points in clusters differ; that is, clusters are different from one another. Cluster analysis can be divided into hierarchical clustering and partitioning clustering. Anderberg (1973) believed that it would be objective and economical to take hierarchical clustering's result as the initial cluster and then adjust the clusters with partitioning clustering. The first step of cluster analysis is to measure the similarity, followed by deciding upon cluster methods, deciding cluster manner of cluster method, deciding number of clusters and explanations for the cluster. Ward's method of hierarchical clustering is the initial result. K-means in partitioning clustering adjusts the clusters.

A special type of clustering is called segmentation. With segmentation a database is partitioned into disjointed groupings of similar tuples called segments. Segmentation is often viewed as being identical to clustering. In other circles segmentation is viewed as a specific type of clustering applied to a database itself.

Clustering can be used in designing a triage system. Triage helps to classify patients at emergency departments to make the most effective use of resources distributed. What is more important is that accuracy in carrying out triage matters greatly in terms of medical quality, patient satisfaction and life security. The study is made on medical management and nursing, with the knowledge of the administrative head at the Emergency Department, in the hope to effectively improve consistency of triage with the combination of data mining theories and practice. The purposes are as follows:

1. Based on information management, the information system is applied in triage of the Emergency Department to generate patients' data.
2. Exploration of correlation between triage and abnormal diagnosis; cluster analysis conducted on variables with clinical meanings.
3. Establishing triage abnormal diagnosis clusters with hierarchical clustering (Ward's method) and partitioning clustering (K-means algorithm); obtaining correlation law of abnormal diagnosis with decision trees.
4. Improving consistency of triage with data mining; offering quantified and scientific rules for triage decision-making in the hope of serving as a foundation for future researchers and clinical examination.

3.3 Association rules

Link analysis, alternatively referred to as affinity analysis or association, refers to the data mining task of uncovering relationships among data. The best example of this type of

application is to determine association rules. An association rule is a model that identifies specific types of data associations. These associations are often used in the retail sales community to identify items that are frequently purchased together. Associations are also used in many other applications such as predicting the failure of telecommunication switches.

Users of association rules must be cautioned that these are not causal relationships. They do not represent any relationship inherent in the actual data (as is true with functional dependencies) or in the real world. There probably is no relationship between bread and pretzels that causes them to be purchased together. And there is no guarantee that this association will apply in the future. However, association rules can be used to assist retail store management in effective advertising, marketing, and inventory control.

The discovery of new knowledge by mining medical databases is crucial in order to make an effective use of stored data, enhancing patient management tasks. One of the main objectives of data mining methods is to provide a clear and understandable description of patterns held in data. One of the best studied models for pattern discovery in the field of data mining is that of association rules. Association rules in relational databases relate the presence of values of some attributes with values of some other attributes in the same tuple. The rule [A = a]) ⇨ [B = b] tells us that whenever the attribute A takes value a in a tuple, the attribute B takes value b in the same tuple. The accuracy and importance of association rules are usually estimated by means of two probability measures called confidence and support respectively. Discovery of association rules is one of the main techniques that can be used both by physicians and managers to obtain knowledge from large medical databases.

Medical databases are used to store a big amount of quantitative attributes. But in common conversation and reasoning, humans employ rules relating imprecise terms rather than precise values. For instance, a physician will find more appropriate to describe his/ her knowledge by means of rules like "if fever is high and cough is moderate then disease is X" than by using rules like "if fever is 38.78C and cough is 5 over 10 then disease is X". It seems clear that rules relating precise values are less informative and most of the time they seem strange to humans. So nowadays, some people apply semantics to improve the association rules mining from a database containing precise values. We can reach that goal by

1. Finding a suitable representation for the imprecise terms that the users consider to be appropriate, in the domain of each quantitative attribute,
2. Generalizing the probabilistic measures of confidence and support of association rules in the presence of imprecision,
3. Improving the semantics of the measures. The confidence/ support framework has been shown not to be appropriate in general, though it is a good basis for the definition of new measures,
4. Designing an algorithm to perform the mining task.

4. Part 3 A KDD Process Model

The terms knowledge discovery in database (KDD) and data mining are distinct.

KDD refers to overall process of discovering useful knowledge from data. It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge. It also includes the choice of encoding schemes, preprocessing, sampling, and projections of the data prior to the data mining step.

Data mining refers to the application of algorithms for extracting patterns from data without the additional steps of the KDD process.

The KDD process is often to be nontrivial; however, we take the larger view that KDD is an all-encompassing concept. KDD is a process that involves many different steps. The input to this process is the data, and the output is the useful information desired by the users. However, the objective may be unclear or inexact. The process itself is interactive and may require much elapsed time. To ensure the usefulness and accuracy of the results of the process, interaction throughout the process with both domain experts and technical experts might be needed.

Data mining is the step in the process of knowledge discovery in databases, that inputs predominantly cleaned, transformed data, searches the data using algorithms, and outputs patterns and relationships to the interpretation/ evaluation step of the KDD process. The definition clearly implies that what data mining (in this view) discovers is hypotheses about patterns and relationships. Those patterns and relationships are then subject to interpretation and evaluation before they can be called knowledge. Fig. 3.1 illustrates the overall KDD process.

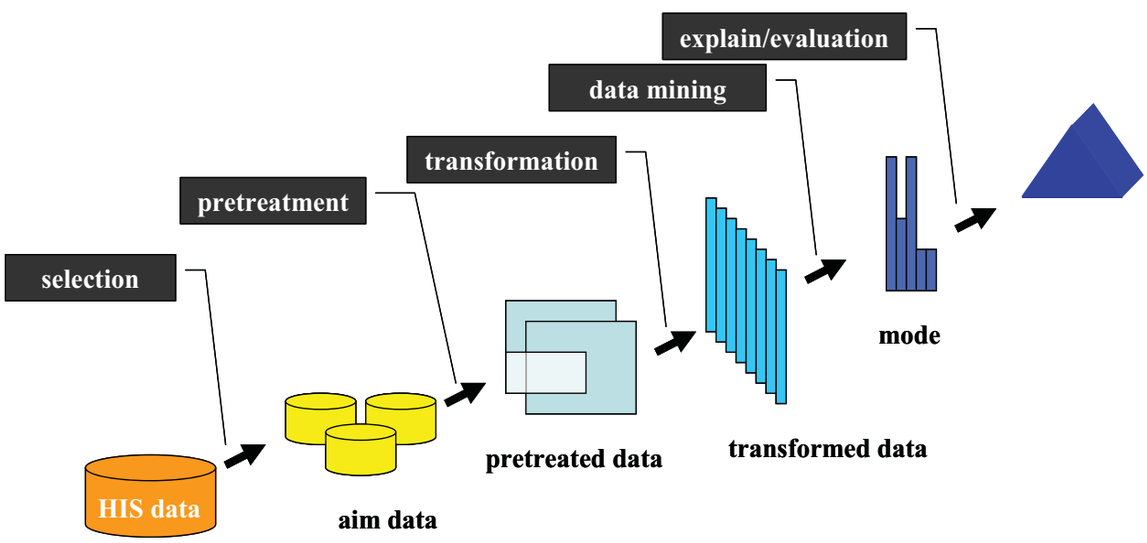


Fig. 3.1 KDD process

The KDD process consists of the following five steps:

1. Select a target data set: The data needed for the data mining process may be obtained from many different and heterogeneous data sources. This first step obtains the data from various databases, files, and nonelectronic sources. With the help of one or more human experts and knowledge discovery tools, we choose an initial set of data to be analyzed.
2. Data preprocessing: The data to be used by the process may have incorrect or missing data. There may be anomalous data from multiple sources involving different data types and metrics. There may be many different activities performed at this time. We use available resources to deal with noisy data. We decide what to do about missing data values and how to account for time-sequence information.
3. Data transformation: Attributes and instances are added and/ or eliminated from the target data. Data from different sources must be converted into a common format for processing. Some data may be encoded or transformed into more usable formats. Data reduction may be used to reduce the number of possible data values being considered.

4. Data mining: A best model for representing the data is created by applying one or more data mining algorithms. Based on the data mining task being performed, this step applies algorithms to the transformed data to generate the desired results.
5. Interpretation/ evaluation: We examine the output from step 4 to determine if what has been discovered is both useful and interesting. Decisions are made about whether to repeat previous steps using new attributes and/ or instances. How the data mining results are presented to the users is extremely important because the usefulness of the results is dependent on it. Various visualization and GUI strategies are used at this last step.

Another important step not contained in the KDD process is goal identification. The focus of this step is on understanding the domain being considered for knowledge discovery. We write a clear statement about what is to be accomplished. Hypotheses offering likely or desired likely or desired outcomes can be stated. A main objective of goal identification is to clearly define what is to be accomplished. This step is in many ways the most difficult, as decisions about resource allocations as well as measures of success need to be determined. Whenever possible, broad goals should be stated in the form of specific objectives. Here is a partial list of things to consider at this stage:

- A clear problem statement is provided as well as a list of criteria to measure success and failure. One or more hypotheses offering likely or desired outcomes may be established.
- The choice of a data mining tool or set of tools is made. The choice of a tool depends on several factors, including the level of explanation required and whether learning is supervised, unsupervised, or a combination of both techniques.
- An estimated project cost is determined. A plan for human resource management is offered.
- A project completion/ product delivery data is given.
- Legal issues that may arise from applying the results of the discovery process are taken into account.
- A plan for maintenance of a working system is provided as appropriate. As new data becomes available, a main consideration is a methodology for updating a working model.

Our list is by no means exhaustive. As with any software project, more complex problems require additional planning. Of major importance is the location, availability, and condition of resource data.

The data mining process itself is complex. As we will see in later chapters, there are many different data mining applications and algorithms. These algorithms must be carefully applied to be effective. Discovered patterns must be correctly interpreted and properly evaluated to ensure that the resulting information is meaningful and accurate.

5. Part 4. Warehouse and OLAP

5.1 Data warehousing

The term data warehouse was first used by William Inmon in the early 1980s. He defined data warehouse to be a set of data that supports DSS and is "subject-oriented, integrated, time-variant, and nonvolatile." With data warehousing, corporate-wide data (current and historical) are merged into a single repository. Traditional databases contain operational data that represent the day-to-day needs of a company. Traditional business data processing

(such as billing, inventory control, payroll, and manufacturing support) support online transaction processing and batch reporting applications. A data warehouse, however, contains informational data, which are used to support other functions such as planning and forecasting. Although much of the content is similar between the operational and informational data, much is different. As a matter of fact, the operational data are transformed into the informational data.

The basic components of a data warehousing system include data migration, the warehouse, and access tools. The data are extracted from operational systems, but must be reformatted, cleansed, integrated, and summarized before being placed in the warehouse. Much of the operational data are not needed in the warehouse and are removed during this conversion process. This migration process is similar to that needed for data mining applications except that data mining applications need not necessarily be performed on summarized or business-wide data. The applications to access a warehouse include traditional querying, OLAP, and data mining. Since the warehouse is stored as a database, it can be accessed by traditional query language.

The data transformation process required to convert operational data to informational involves many functions including:

- Unwanted data must be removed.
- Converting heterogeneous sources into one common schema. This problem is the same as that found when accessing data from multiple heterogeneous sources. Each operational database may contain the same data with different attribute names. In addition, there may be multiple data types for the same attribute.
- As the operational data is probably a snapshot of the data, multiple snapshots may need to be merged to create the historical view.
- Summarizing data is performed to provide a higher level view of the data. This summarization may be done at multiple granularities and for different dimensions.
- New derived data may be added to better facilitate decision support functions.
- Handling missing and erroneous data must be performed. This could entail replacing them with predicted or simply removing these entries.
- When designing a data warehouse, we must think of the uniqueness of medical data carefully. Below we comment on some unique features of medical data.
- Because of the sheer volume and heterogeneity of medical databases, it is unlikely that any current data mining tool can succeed with raw data. The tools may require extracting a sample from the database, in the hope that results obtained in this manner are representative for the entire database. Dimensionality reduction can be achieved in two ways. By sampling in the patient-record space, where some records are selected, often randomly, and used afterwards for data mining; or sampling in the feature space, where only some features of each data record are selected.
- Medical databases are constantly updated by, say, adding new SPECT images (for an existing or new patient), or by replacement of the existing images (say, a SPECT had to be repeated because of technical problems). This requires methods that are able to incrementally update the knowledge learned so far.
- The medical information collected in a database is often incomplete, e.g. some tests were not performed at a given visit, or imprecise, e.g. ‘‘the patient is weak or diaphoretic.’’

- It is very difficult for a medical data collection technique to entirely eliminate noise. Thus, data mining methods should be made less sensitive to noise, or care must be taken that the amount of noise in future data is approximately the same as that in the current data.
- In any large database, we encounter a problem of missing values. A missing value may have been accidentally not entered, or purposely not obtained for technical, economic, or ethical reasons. One approach to address this problem is to substitute missing values with most likely values; another approach is to replace the missing value with all possible values for that attribute. Still another approach is intermediate: specify a likely range of values, instead of only one most likely. The difficulty is how to specify the range in an unbiased manner.

The missing value problem is widely encountered in medical databases, since most medical data are collected as a byproduct of patient-care activities, rather than for organized research protocols, where exhaustive data collection can be enforced. In the emerging federal paradigm of minimal risk investigations, there is preference for data mining solely from byproduct data. Thus, in a large medical database, almost every patient-record is lacking values for some feature, and almost every feature is lacking values for some patient-record.

- The medical data set may contain redundant, insignificant, or inconsistent data objects and/ or attributes. We speak about inconsistent data when the same data item is categorized as belonging to more than one mutually exclusive category. For example, a serum potassium value incompatible with life obtained from a patient who seemed reasonably healthy at the time the serum was drawn. A common explanation is that the specimen was excessively shaken during transport to the laboratory, but one cannot assume this explanation without additional investigation and data, which may be impractical in a data mining investigation.
- Often we want to find natural groupings (clusters) in large dimensional medical data. Objects are clustered together if they are similar to one another (according to some measures), and at the same time are dissimilar from objects in other clusters. A major concern is how to incorporate medical domain knowledge into the mechanisms of clustering. Without that focus and at least partial human supervision, one can easily end up with clustering problems that are computationally infeasible, or results that do not make sense.
- In medicine, we are interested in creating understandable to human descriptions of medical concepts, or models. Machine learning, conceptual clustering, genetic algorithms, and fuzzy sets are the principal methods used for achieving this goal, since they can create a model in terms of intuitively transparent if . . . then . . . rules. On the other hand, unintuitive black box methods, like artificial neural networks, may be of less interest.

5.2 OLAP

Online analytic processing (OLAP) systems are targeted to provide more complex query results than traditional OLTP or database systems. Unlike database queries, however, OLAP applications usually involve analysis of the actual data. They can be thought of as an extension of some of the basic aggregation functions available in SQL. This extra analysis of the data as well as the more imprecise nature of the OLAP queries is what really

differentiate OLAP applications from traditional database and OLTP applications. OLAP tools may also be used in DSS systems.

OLAP is performed on data warehouse or data marts. The primary goal of OLAP is to support ad hoc querying needed to support DSS. The multidimensional view of data is fundamental to OLAP applications. OLAP is an application view, not a data structure or schema. The complex nature of OLAP applications requires a multidimensional review of the data. The type of data accessed is often (although not a requirement) a data warehouse.

OLAP tools can be classified as ROLAP or MOLAP. With MOLAP (multidimensional OLAP), data are modeled, viewed, and physically stored in a multidimensional database (MDD). MOLAP tools are implemented by specialized DBMS and software systems capable of supporting the multidimensional data directly. With MOLAP, data are stored as an n-dimensional array (assuming there are n dimensions), so the cube view is stored directly. Although MOLAP has extremely high storage requirements, indices are used to speed up processing. With ROLAP (relational OLAP), however, data are stored in a relational database, and a ROLAP server (middleware) creates the multidimensional view for the user. As one would think, the ROLAP tools tend to be less complex, but also less efficient. MDD systems may presummarize along all dimensions. A third approach, hybrid OLAP (HOLAP), combines the best features of ROLAP and MOLAP. Queries are stated in multidimensional terms. Data that are not updated frequently will be stored as MDD, whereas data that are updated frequently will be stored as RDB.

There are several types of OLAP operations supported by OLAP tools:

- A simple query may look at a single cell within the cube.
- Slice: Look at a subcube to get more specific information. This is performed by selecting on one dimension. This is looking at a portion of the cube.
- Dice: Look at a subcube by selecting on two or more dimensions. This can be performed by a slice on one dimension and the rotating the cube to select on a second dimension. A dice is made because the view in slice in rotated from all cells for one product to all cells for one location.
- Roll up (dimension reduction, aggregation): Roll up allows the user to ask questions that move up an aggregation hierarchy. Instead of looking at one single fact, we look at all the facts. Thus, we could, for example, look at the overall total sales for the company.
- Drill down: These functions allow a user to get more detailed fact information by navigating lower in the aggregation hierarchy. We could perhaps look at quantities sold within a specific area of each of the cities.
- Visualization: Visualization allows the OLAP users to actually "see" the results of an operation.

To assist with roll up and drill down operations, frequently used aggregations can be precomputed and stored in the warehouse. There have been several different definitions for a dice. In fact, the term slice and dice is sometimes viewed together as indicating that the cube is subdivided by selecting on multiple dimensions.

6. Part 5 Embedded real-time KDD process

6.1 A modified KDD process

As referred in Part 3, traditional knowledge discovery process includes five steps: selection, pretreatment, transformation, data mining, and interpretation and evaluation. It is not a

simple linear course, but an iterative one including recurrence between every two steps. Refine and deepen the knowledge continuously, and finally make it easier to understand.

Medical data is huge and disorganized generally because of residing in different information systems. This makes data mining of medical data different from others. Data mining in the HIS database is an important work for hospital management. Hospital managers can utilize the knowledge mined sufficiently into decision-making for the hospital with the final purpose to provide the hospital better development. A modified KDD process was proposed for medical data mining using Intersystems BI tool DeepSee (Fig. 5.1).

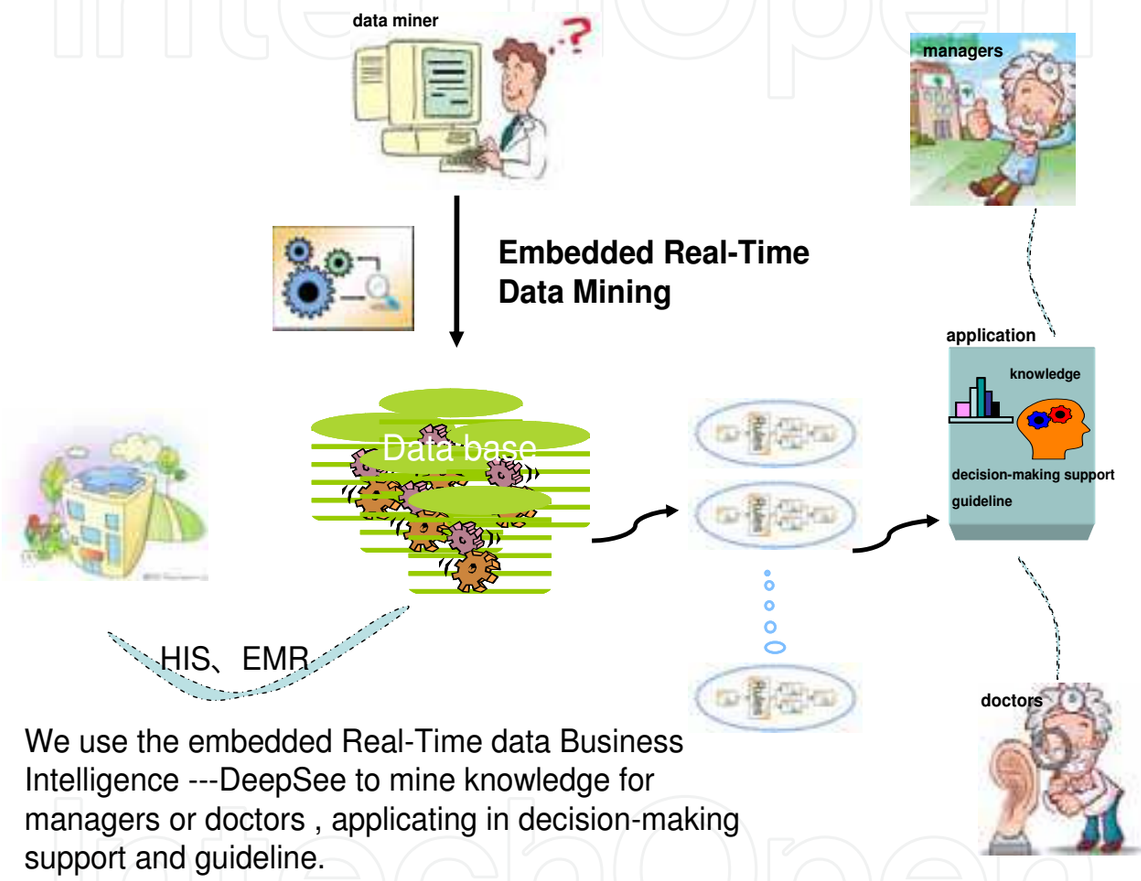


Fig. 5.1 Embedded real-time process mining

6.2 Embedded real-time process mining

DeepSee is innovative software that enables data miners to embedded real-time business intelligence capabilities into the existing and future transactional applications. Embedded real-time business intelligence is different from the traditional data mining process. That's because it's focused on providing every user with useful, timely information related to making operational decision. InterSystems DeepSee is used for supporting the decision-making process in hospital management. With DeepSee, hospital managers can look at what's happening in the hospital while it's actually taking place and they can make the changes needed to improve the medical process.

Fig. 5.2 Market Evolution of DeepSee

1. **Business Intelligence** is the art of putting the data collected by applications to good use-analyzing it to provide information that helps managers make better business decisions. Traditionally, such analysis has been performed by small groups of "data experts" working with specialized tools, looking at data gathered into a data warehouse. Because loading data into a warehouse often takes considerable time, the information gleaned from traditional business intelligence is usually historical in nature.
2. **Real-time Business Intelligence** takes the data warehouse out of the picture. It allows timely analysis of data stored within transactional applications. Because the data is "fresh", real-time business intelligence helps users make better operational decisions.
3. **Embedded Real-time Business Intelligence** means that the capability to turn operational data into immediately useful in formation is included as a feature of the transactional application. Users don't have to be data analysis experts or use separate tools to gain insight from their data.

With DeepSee, business intelligence is:

- Fast- Utilizing InterSystems' breakthrough transitional bit indexing map technology that provides excellent retrieval performance for complex queries plus top-tier update performance for high-volume transaction processing, information is accessible in real time.
- Easy- Using DeepSee, application developers rapidly build interactive dashboards containing graphs, charts, filters, images, links, etc.
- Cost-effective- DeepSee removes the costly requirement to create and maintain a data warehouse because, unlike traditional BI, it accesses your current transactional data.

Managers can monitor every branch department to compare the results of local hospital activities at any point during the medical process. Depending on the information delivered by operational BI, decisions can be made in real time to implement activities that are proving successful in other locations- a promotion for high interest bearing account, for

example- and to immediately cut off promotions that aren't working effectively in certain locations. The emphasis of data mining is on an implementation of a general approach to rule based decision-making.

Four steps to embedded BI with DeepSee

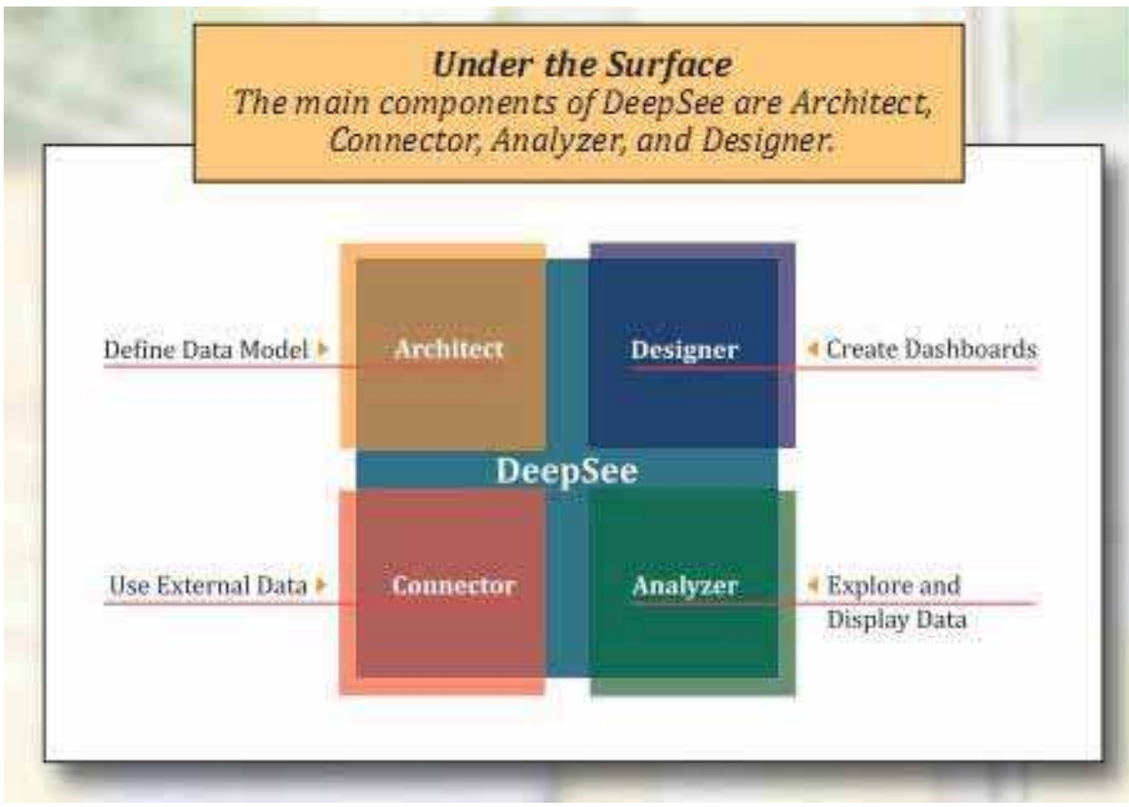


Fig. 5.3 The main four components of DeepSee

Step 1. Determine Key Performance Indicators

Better than anyone, your users know how to do their jobs. And they can tell you what knowledge they need to be able to do their jobs better. Through discussions with end users, you can figure out what key performance indicators they want to be able to analyze in real time. A key performance indicator might be a particular bit of raw data that is collected by your application, or it might be a measurement that is calculated from disparate solutions. The best way to determine meaningful, useful performance indicators is to talk to your users.

Step 2. Define a data model

Your data model is a definition of how to organize the raw data that aggregates into various key performance indicators. If a performance indicator must be calculated from raw data, the data model will define how that is done. The data model is also where you can give data, dimensions, and key performance indicators names that will be intuitive and meaningful for end users.

The dimensions of a data model determine how many ways a performance can be analyzed, and thus what raw data needs to be included in your model. In order to speed analysis time,

some of those dimensions may have indices defined within your model. DeepSee works with transactional data, so information will be organized and indexed by your data model in real time.

The task of defining a data model is accomplished using the DeepSee Architect.

Step 2a. (if necessary): Incorporate "foreign" data

If any of the raw data needed in your data model comes from applications or repositories that are not powered InterSystems' technology, that data must be incorporated using Ensemble and the DeepSee Connector. The Connector provides a "snapshot" of the external data, which may be transformed (through the use of configurable business rules) to fit into your data model. The snapshot can be a one-time import, or occur on a scheduled basis. Incremental updates are supported.

Step 3. Build components

The DeepSee Analyzer enables point-and-click or drag-and-drop creation of pivot tables, graphs, and charts that use data models defined by the DeepSee Architect. These components are dynamic, allowing users to drill all the way down to the underlying detail data.

Step 4. Design a dashboard

With the DeepSee Designer, you will create dashboards that include the graphs, charts, and pivot tables you built with the DeepSee Analyzer, as well as links, combo-boxes, lists, and other user interface components. Dashboards can be tailored to specific topics, functions, or individuals. You can control how much flexibility users have when exploring data for example, pre-defined filters can exclude sensitive data from users who have no need to see it.

The dashboards you create with the Designer are Web pages that can easily be embedded within the user interface of your application. Users do not have to be data analysis experts to reap the benefits of real-time business intelligence. They merely need a working knowledge of your application.

7. Part 6 Two case studies of data mining in HIS

7.1 KDD based on DeepSee

DeepSee contains four main components:

A: Connector

- For non-Caché data, the Connector can extract data from external sources so that it can be modeled using the Architect.
- The connector is not used if the data is already in Caché (or Ensemble).

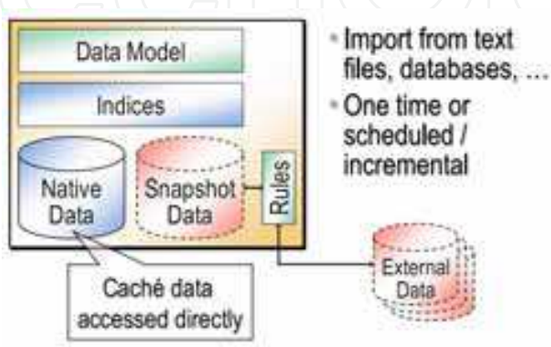


Fig. 6.1 DeepSee Connector

B: Architect

- Defines the data models to be used by the Analyzer.
- A data model defines the dimensions and measures by which data can be analyzed.
- High-performance bitmap indices are created for optimal performance.
- Models are based on current transactional data-no data warehouse is required.

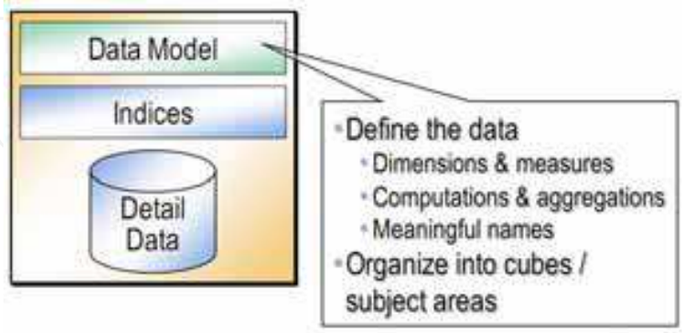


Fig. 6.2 DeepSee Architect

C: Analyzer

- Point-and-click/ drag-and-drop creation of pivot tables and graphs.
- Uses data models defined by the Architect.
- Dynamic drill down all the way to the underlying detail data.

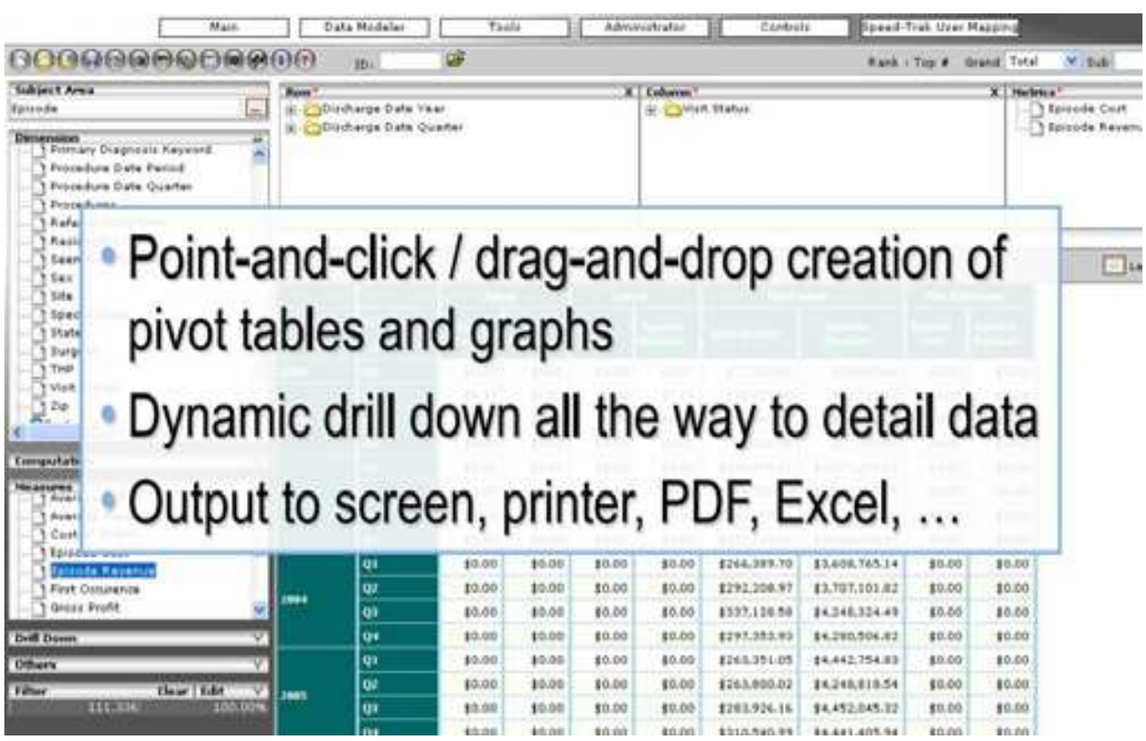


Fig. 6.3 DeepSee Analyzer

D: Designer

- Create dashboards that use the pivot tables and graphs built with the Analyzer.
- Dashboards are Web pages that can be embedded into applications.
- Dashboards can also include interactive UI controls like combo-boxes, lists, radio buttons, links, etc.

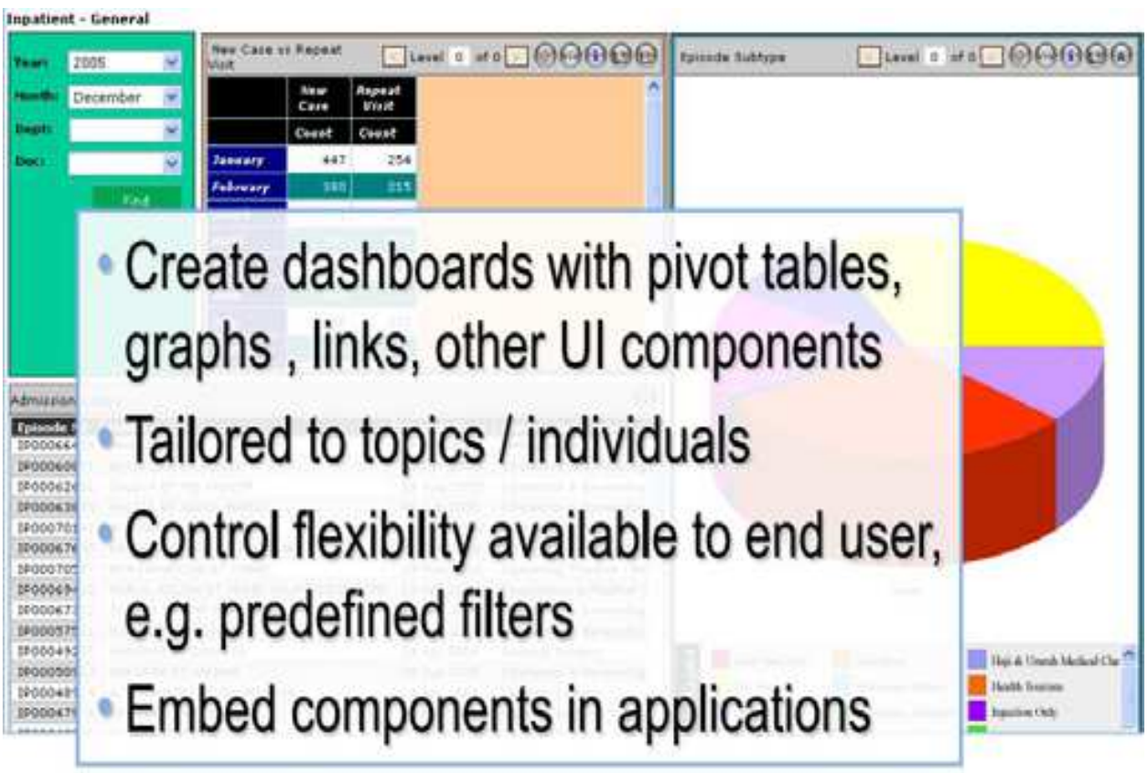


Fig. 6.4 DeepSee Designer

In the following part, we will introduce two KDD examples using DeepSee.

7.2 KDD of Inpatient Fees

Health care organizations are increasing expenditures on information technology as a means to improve safety, quality, and access. Decision-making is the core of hospital management, and goes widely throughout the whole hospital behavior. It is pivotal for hospitals to make appropriate decision, which is closely related to hospital development. Managers suffer pressure to make decision, when the knowledge from Hospital Information System is not so high-quality, comprehensive or reliable. Therefore it is important to integrate data from different Hospital Information Systems, carry out data mining and subsequently discover the knowledge in the mining result to promote the hospitals' competition.

HIS database involves fee information related to the whole medical behavior, such as examinations, tests, treatments, prescriptions, nursing and supplies et al. We can relevantly construct models based on various themes for data mining. The database contains information as follow: 1. patient information, 2. diagnostic information (diseases category and diagnose information), 3. medical information (surgeries, radio chemotherapies,

nursing, prescriptions, examinations, and treatment orders, treatment departments and corresponding managers) 4. fee information (treatment fee details, beds and medical consumables), 5. insurance patient data, 6. time information (starting time and duration of the whole medical behavior). All the information can be discovered to provide knowledge for decision-making support.

7.2.1 Modeling of inpatient fee

For every theme, the dimensions, relevant indexes and data source each instance should be defined. A model is built based on inpatient fee to analyze HIS data (in the duration from 2001 to 2007) of a hospital in Zhejiang Province in China. The dimensions and data source defined, accordingly, are in Table 1 as follow:

Theme	**Inpatient fees**
Dimensionalities	Date dimensionalities (duration, day, week, month, quarter, year)
Related indexes	Fee category, medical insurance category, department, net payment
Dataset and the sources (summary)	Inpatient master records: PAT_VISIT Inpatient master index: PAT_MASTER_INDEX Inpatient settle master: INP_SETTLE_MASTER Inpatient settle details: INP_SETTLE_DETAIL
Data details	Charge class: FEECLASSNAME.FEECLASSNAME Discharge time: RCPTNO.visitfk.DISCHARGEDATETIME Admission time: RCPTNO.visitfk.ADMISSIONDATETIME Admission department: RCPTNO.visitfk.DEPTADMISSIONTO.DEPTNAME Discharge department: RCPTNO.visitfk.DEPTDISCHARGEFROM.DEPTNAME Admission form: RCPTNO.visitfk.PATIENTCLASS Discharge form: RCPTNO.visitfk.DISCHARGEDISPOSITION Insurance type: RCPTNO.visitfk.INSURANCETYPE Payment: PAYMENTS

Table 6.1 Theme and dimensions

Then a data model is built in the HIS database according to the data source. It is object model below (Fig. 6.5) to explain the relationship between the data:

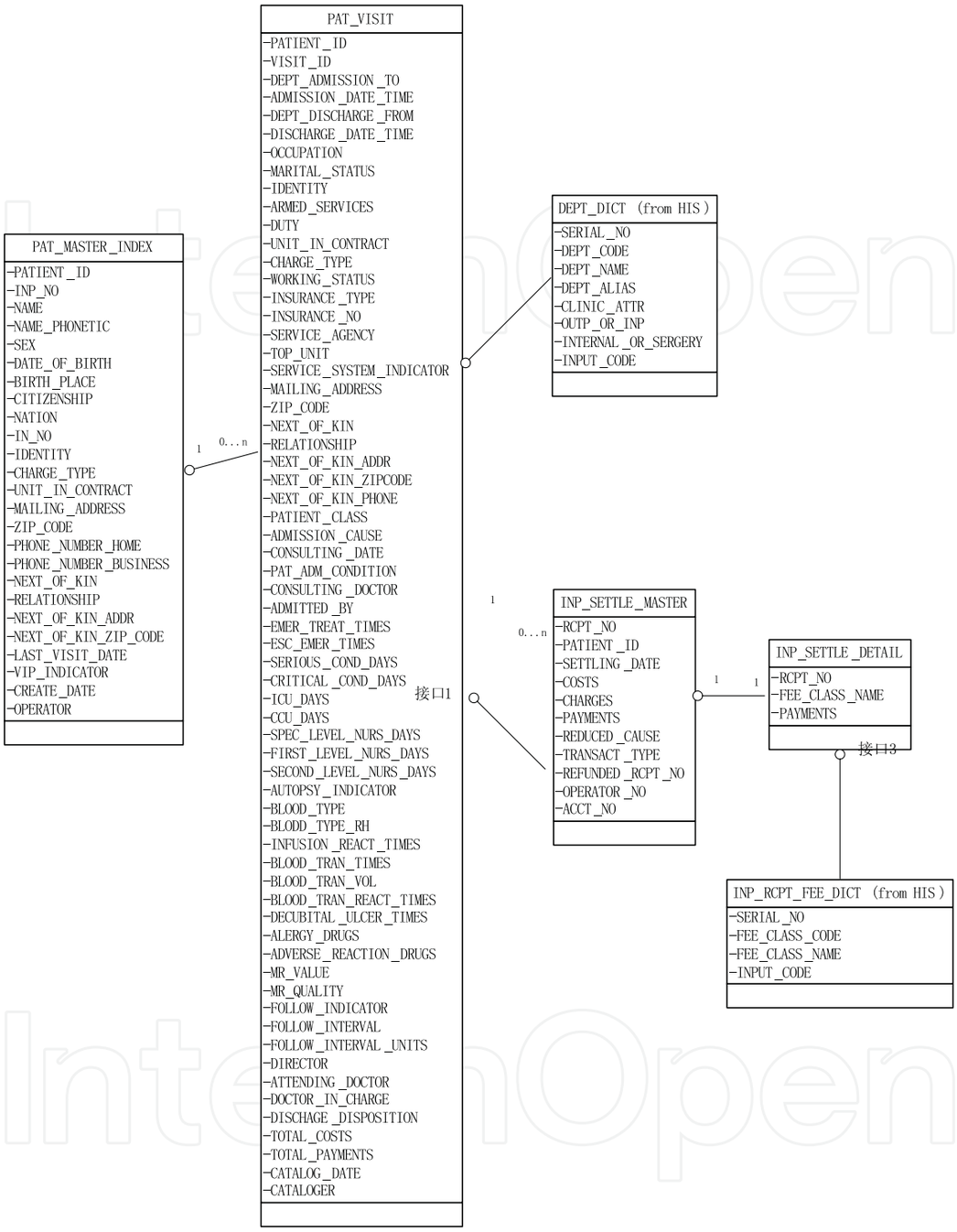


Fig. 6.5 Object model of inpatient fees

This model will be analyzed in three aspects: medical insurance fee, department annual fee and fee compositions.

7.2.2 Related work and results

Based on the model built above, the embedded real-time DeepSee is used to carry out data mining for the model of inpatient fee theme. It will be analyzed from three aspects.

7.2.2.1 *Analysis of medical insurance fee*

The data mining result of medical insurance fee from 2001 to 2007 is illustrated in Fig. 6.6. It can be seen that public retire staff occupies the maximum proportion of 22.35%, and the unemployed people least, 0.72%. It is important for the hospital to receive public retire staffs due to the high proportion. Hospital managers should adjust guidelines accordingly to provide higher quality treatment for these patients in advance.

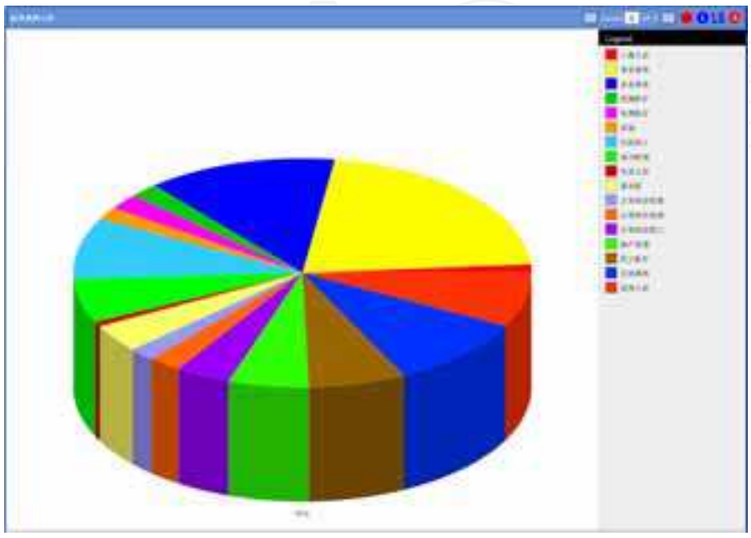


Fig. 6.6 Medical insurance fees from 2001 to 2007

7.2.2.2 *Analysis of the annual fee*

From Fig. 6.7, the conclusion can be reached that the department annual fee rose continuously from the year 2001 to 2007, with the amount rising from 9,346.86 million to 19,788.85 million and the growth rate is about 1.17.

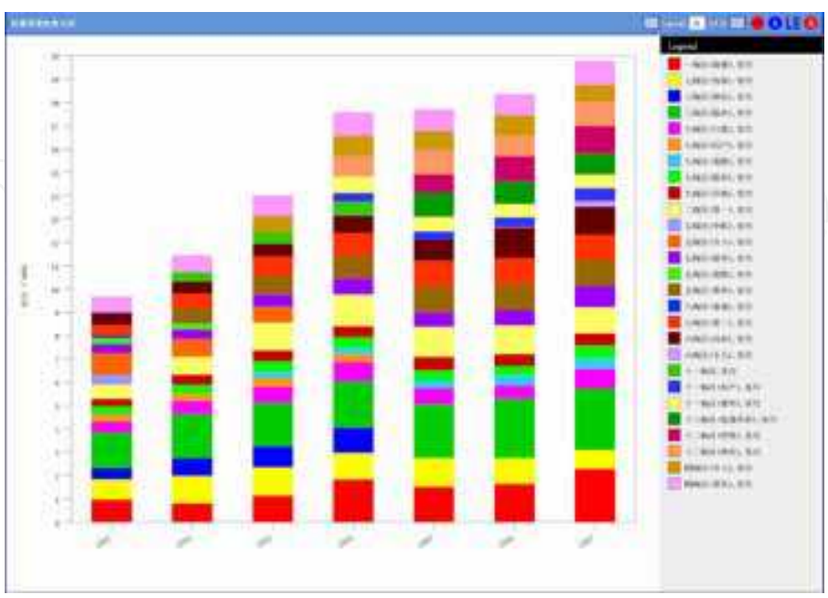


Fig. 6.7 Department annual fee from 2001 to 2007

Conclusion could be got from Fig. 6.8 that the third endemic area (cerebral surgery) had the most proportion, followed by the first endemic area (burn and plastic) and second endemic area (orthopaedics).

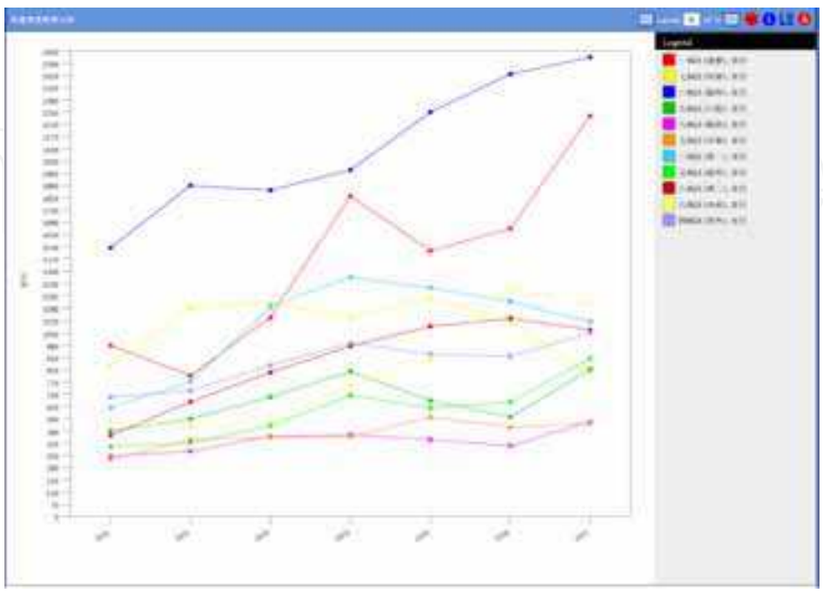


Fig. 6.8 Trend of each department annual fee from 2001 to 2007

7.2.2.3 *Analysis of fee compositions*

Inpatient fees include drugs, examinations, treatments, test and operations et al. We can reach the conclusion that the fees of western medicine, treatment and operation occupied a fairly large proportion, according to Fig. 6.9.

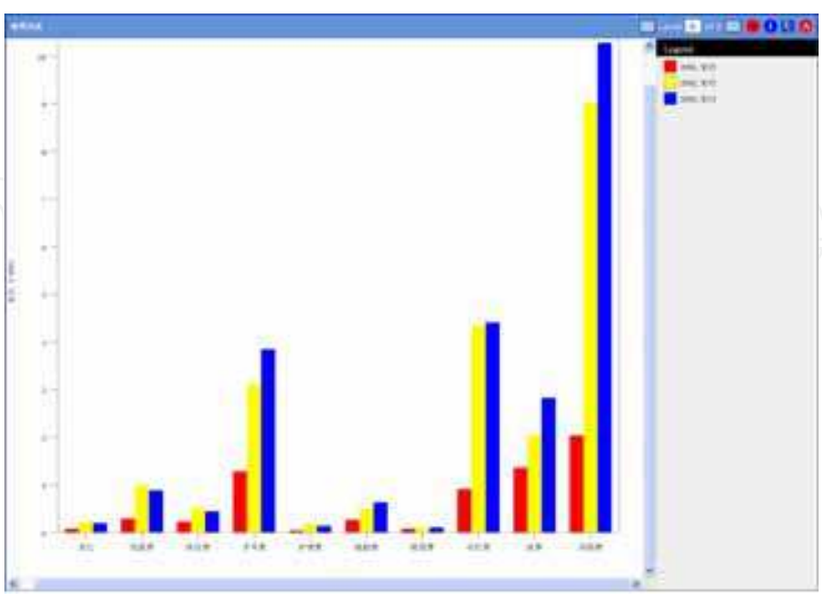


Fig. 6.9 Fee proportion from 2001-2003

When the fee proportion of drugs, examination or test is too high, managers can analyze the compositions of every inpatient fee according to data mining results, and finally control over treatment pertinently.

7.3 KDD of Pharmacy

Pharmacy is the main place of drug storage and supply base in hospital. Pharmacy managers have a responsibility to guarantee the medicinal safety, effect and abundance. Correctness of pharmaceutical expenditure accounts influences hospital operating results directly, so it has financial significance to enhance management during drug circulation. Hospital Information System (HIS) database contains all data related to pharmacy.

7.3.1 Modeling of pharmacy theme

A model is built based on the pharmacy data (in the duration from 2001 to 2005) of a hospital in Zhejiang Province in China. The dimensions and data source defined, accordingly, are in Table 6.2 as follow:

Theme	**Pharmacy**
Dimensionalities	Date dimensionalities
Related indexes	Drug name, Stock name, Annual inventory, Inventory profit, Stock amount, Supplier
Dataset and the sources (summary)	Drug dictionary: Drug_Dict Drug supplier catalog:DRUG_SUPPLIER_CATALOG Drug stock balance: DRUG_STOCK_BALANCE Drug storage dept dictionary: DRUG_STORAGE_DEPT
Data details	Drug name: drugfk.DRUGNAME Drug code: drugfk.DRUGCODE Export money: EXPORTMONEY Annual inventory: INVENTORY Profit: PROFIT Storage name: STORAGE.STORAGENAME Supplier: FIRMID.SUPPLIER Time: YEARMONTH (ps: drugfk is the foreign key of DRUG_CODE and DRUG_SPEC)

Table 6.2 Theme and dimensions

Then pharmacy model is established according to the data source in the HIS database. It's an object model in Fig. 6.10 to explain the relationship between the tables. The model includes four tables: Drug stock balance, Drug storage dept dictionary, Drug dictionary and Drug supplier catalog. Drug stock balance is the main table for analysis, and it includes 246,016 records data involving 11,655 kinds of drugs. The relationship between these tables is one to one correspondence. Subsequent data mining is all based on this pharmacy model, analyzing in four aspects: delivery trend, stocks, stock department profitability and profitability from different supplier.

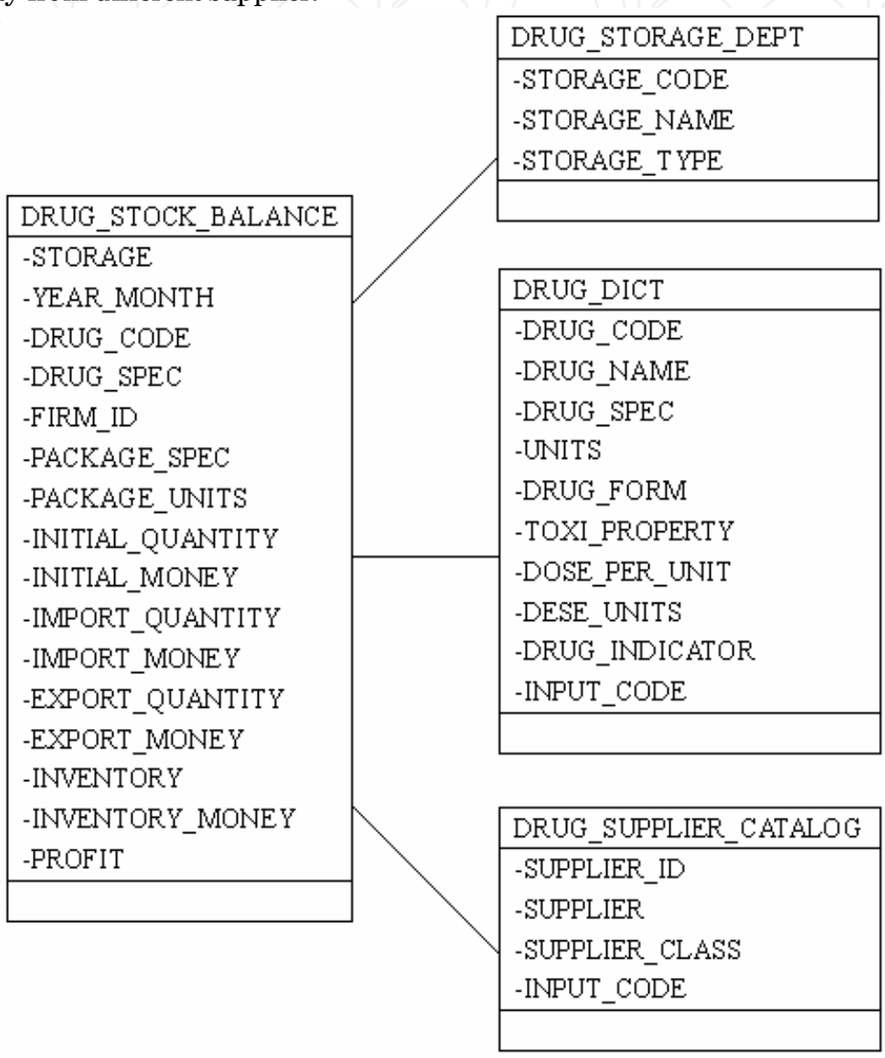


Fig. 6.10 Object model based on pharmacy theme

7.3.2 Related work and results

Data mining analysis is established based on the object model above. Main data mining dashboard (Fig. 6.11) contains three parts: project selection, main panel and filter. Analysis will be expatiated from four aspects: delivery trend, stocks, stock department profitability and profitability from different supplier. The main panel can be chosen to display the topic in statistical data or graphics. Data will be filtered by time, drug name and stock units.

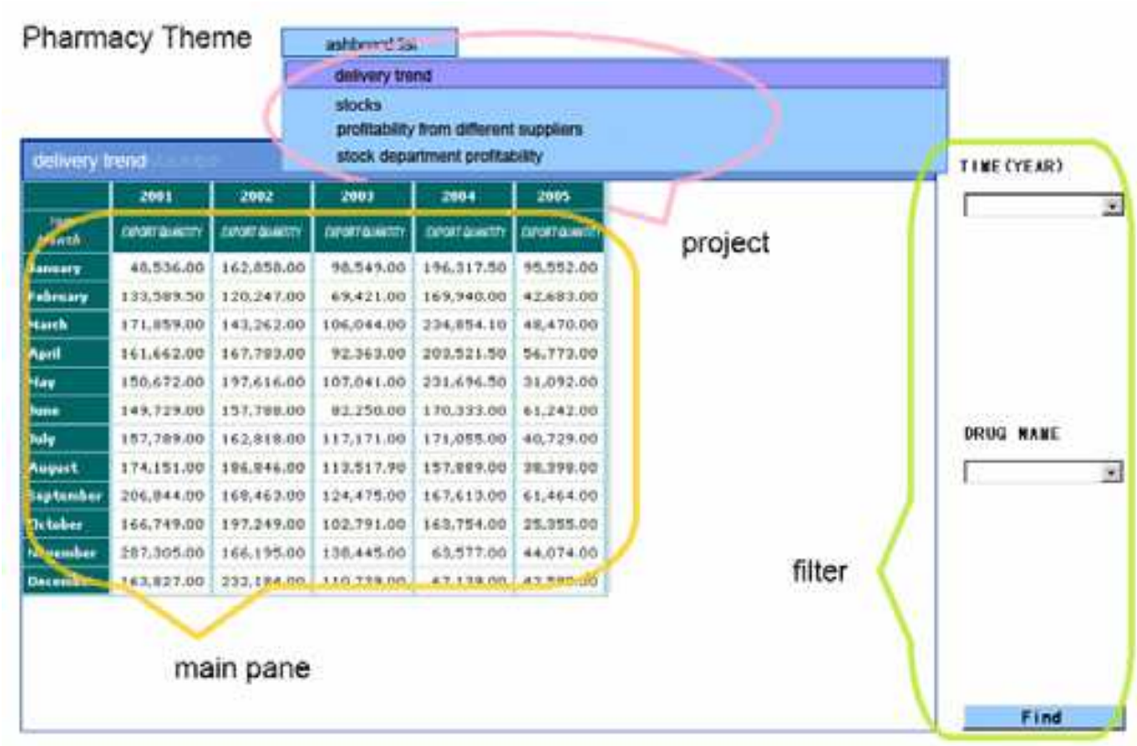


Fig. 6.11 Main data mining dashboard

7.3.2.1 *Analysis of Delivery Trend*

By observing delivery trend graphical, managers can adjust the drug inventory and warehousing in the next year. On the other hand, managers can speculate epidemics while there are fluctuation or peak valley in the trend graphics.

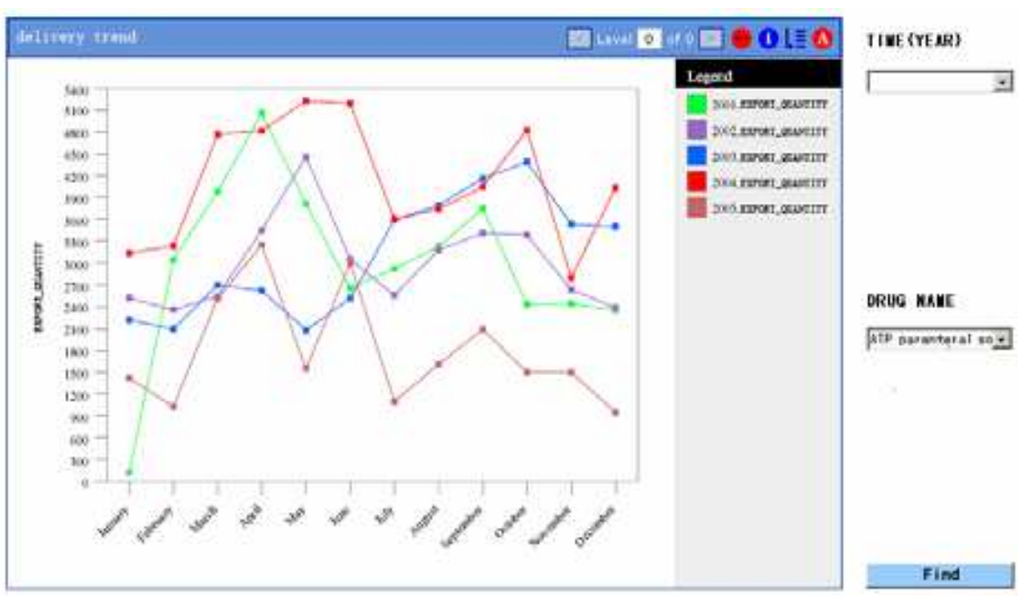


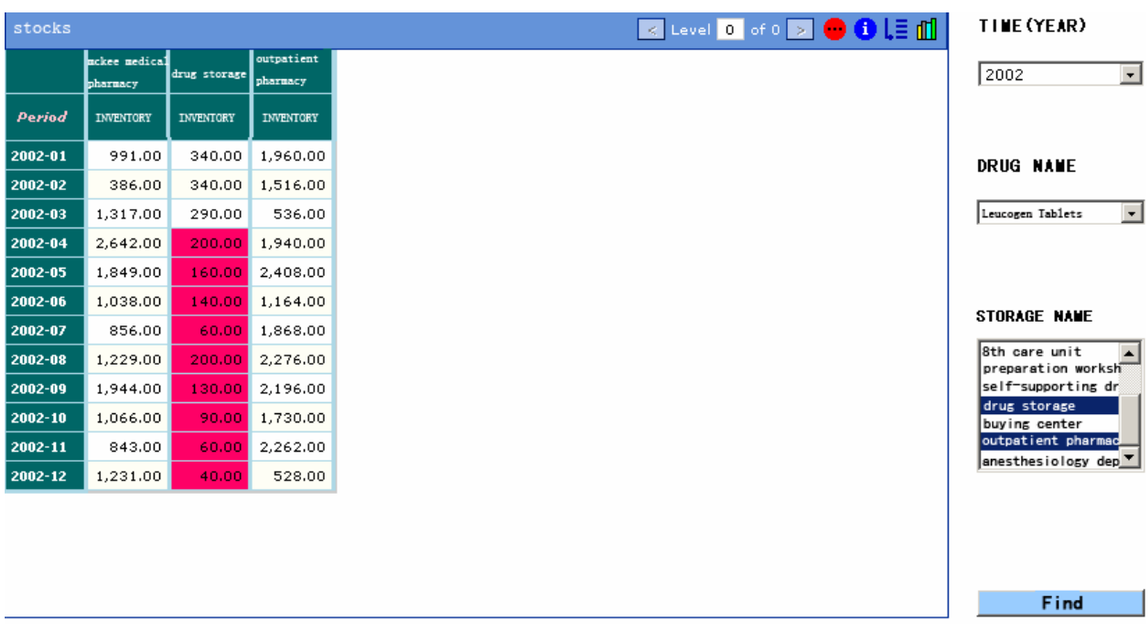
Fig. 6.12 Delivery trend of ATP Injection

The delivery trend of ATP Injection from 2001 to 2005 is illustrated in Fig. 6.12, horizontal axis as the time month, vertical axis as delivery volume, different colors corresponding to different year. Analysis will be carried through two aspects: vertically the global delivery trends from 2001 to 2005, and horizontally the delivery trend of ATP Injection in the assigned year. The global delivery volume of ATP Injection from 2001 to 2005 keeps in a relatively stable state. In most years, there is a delivery peak around April, and then the trend line declines slowly, and reaches a new peak around October. Pharmacy managers can investigate the actual situation according to these fluctuations and peaks, and adjust the pharmacy more reasonably.

7.3.2.2 *Analysis of Stock*

Stock analysis will be carried out in table and graphics tow forms. In table, red data indicates warning signal. Maximum stock and minimum stock were presupposed. There will be a warning signal any time when stock is higher than maximum stock or lower than minimum stock. Warning signal reminds pharmacy mangers to stop stocking in or stock in time. In graphics, horizontal axis is the time month and vertical axis is stock volume, different colors representing different stock department.

Fig. 6.13 is the table form stock of Leucogen in 2002. It has been below the minimum stock from March in the whole year. According to the delivery trend of Leucogen in the main data mining dashboard, it is the large amount of delivery which causes the warning signal. Pharmacy managers should stock in Leucogen in time to avoid emergency storage. Fig. 6.14 is the graphics form of Leucogen in 2002, which is much more visual than table form.



	mckee medical pharmacy	drug storage	outpatient pharmacy
Period	INVENTORY	INVENTORY	INVENTORY
2002-01	991.00	340.00	1,960.00
2002-02	386.00	340.00	1,516.00
2002-03	1,317.00	290.00	536.00
2002-04	2,642.00	200.00	1,940.00
2002-05	1,849.00	160.00	2,408.00
2002-06	1,038.00	140.00	1,164.00
2002-07	856.00	60.00	1,868.00
2002-08	1,229.00	200.00	2,276.00
2002-09	1,944.00	130.00	2,196.00
2002-10	1,066.00	90.00	1,730.00
2002-11	843.00	60.00	2,262.00
2002-12	1,231.00	40.00	528.00

Fig. 6.13 Stock table of Leucogen

7.3.2.3 *Analysis of Stock Department profitability*

Profitability of department has been paid lots of attention. Departments contain the center pharmacy, pharmaceutical workshop, self-supporting pharmacy, drug storage, procurement center and out-patient pharmacy. Fig. 6.15 exhibits the profitability of each department

referred above from 2001 to 2005, horizontal axis as the year, vertical axis as profitability, different colors corresponding to different departments. As can be seen, drug storage profits stably and accounts for 77.79% to 99.98% from 2001 to 2005. The center pharmacy profits secondly, and pharmaceutical workshop the least.

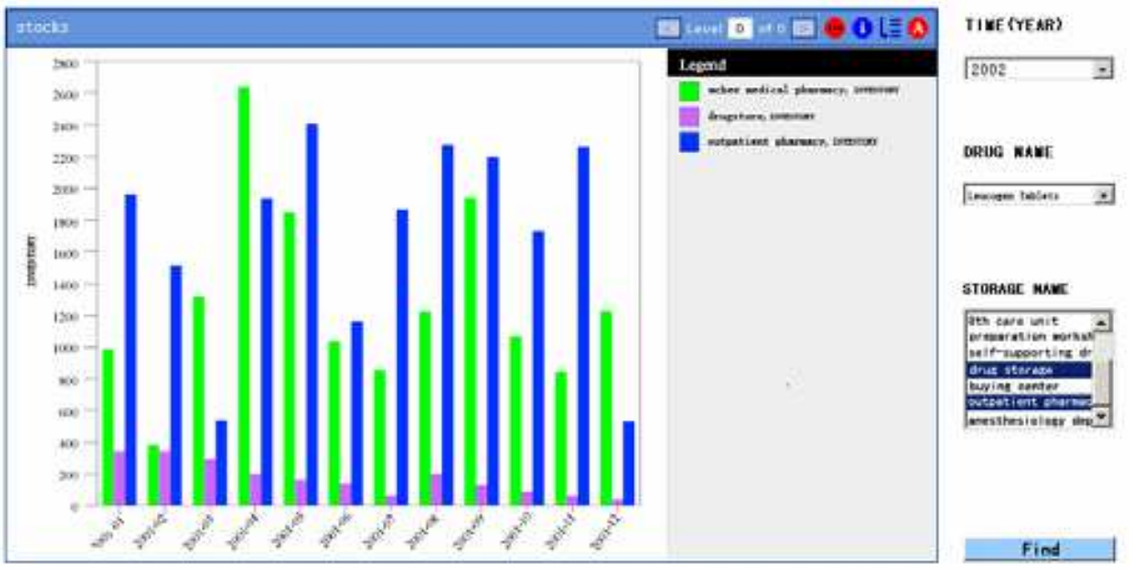


Fig. 6.14 Stock graphics of Leucogen

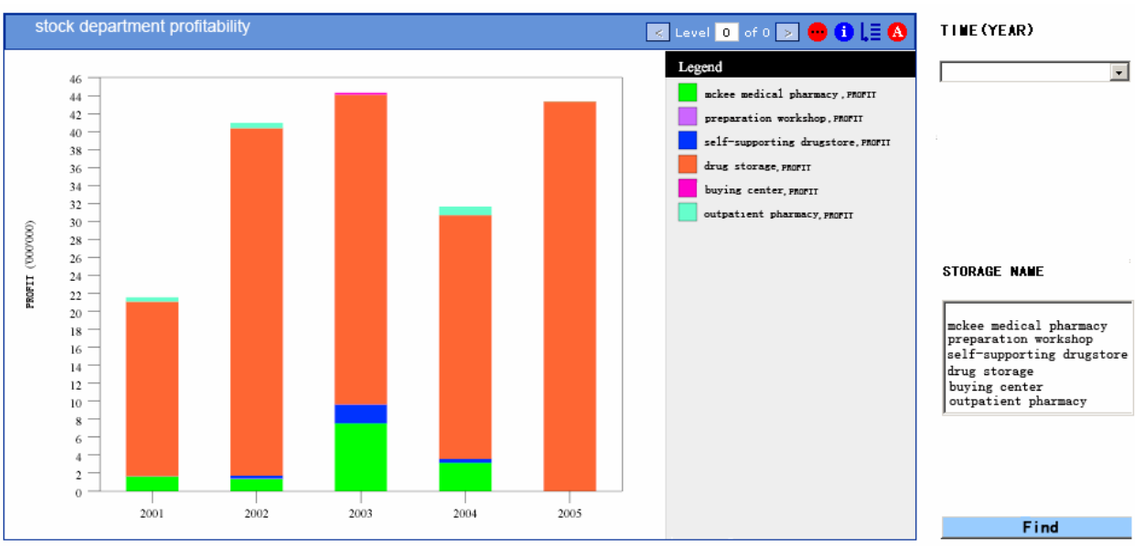


Fig. 6.15 Profitability of each stock department

7.3.2.4 *Analysis of profitability from different suppliers*

Different suppliers make different profitability, which is fairly important to hospital finance. Both the drug price and delivery volume would affect total profitability. Fig. 6.16 is profitability of Mannitol Injection from different suppliers, horizontal axis as suppliers' name, vertical axis as profitability, different colors corresponding to different years.

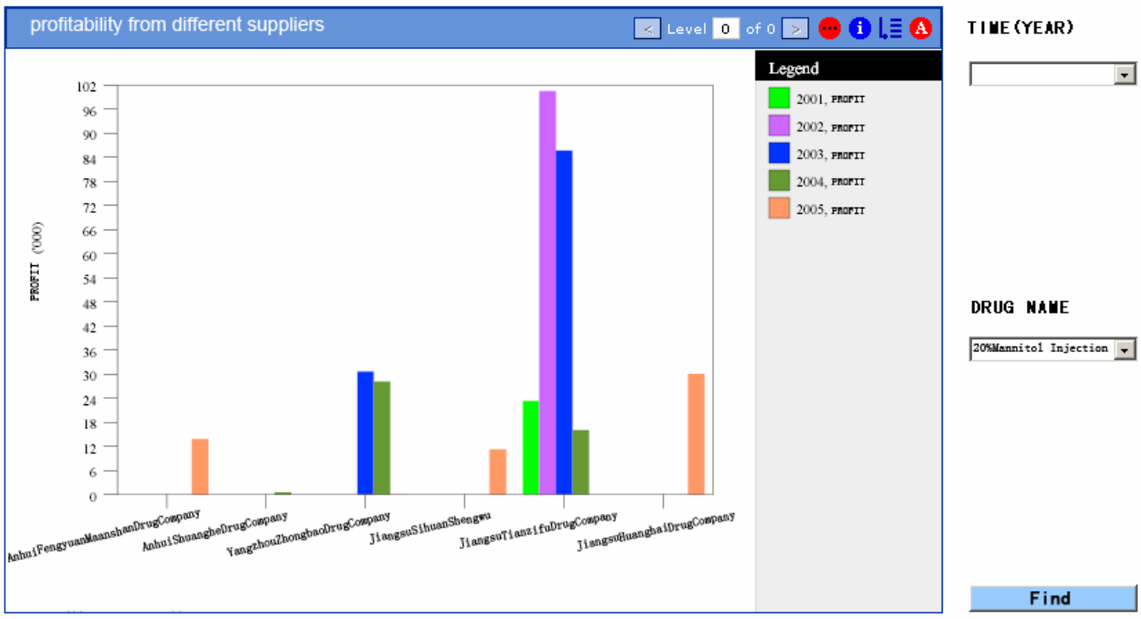


Fig. 6.16 Profitability from different suppliers

It can be seen in Fig. 6.16 that profitabilitys steadily increase from 2001 but gradually decrease from 2004. Jiangsu Tianzifu Drug Company accounts for the largest proportion which means the pharmacy buy large volume of Mannitol Injection from the supplier.

Decision-making throughout hospital activities is the core of hospital management. How to access high-quality decision-makers in time makes key deciders feel tremendous pressure. Pharmacy is the source of medication in hospital. Data mining of pharmacy in HIS database and monitoring on it is an important way to ensure safe medication and adequate stocks. In this part, we use Deepee as a data mining tool. Managers can utilize the knowledge mined sufficiently into decision-making for the hospital with the final purpose to provide the hospital better development.

8. References

[1] Krzysztof J. Cios, G. William Moore. Uniqueness of medical data mining. Artificial Intelligence in Medicine 26: 1–24, 2002.

[2] Linda Goodwin, Michele VanDyne, Simon Lin. Data mining issues and opportunities for building nursing knowledge. Journal of Biomedical Informatics 36: 379–388, 2003.

[3] Thomas M. Lehmanna, Mark O. Gu¨ld, Thomas Deselaers. Automatic categorization of medical images for content-based retrieval and data mining. Computerized Medical Imaging and Graphics 29: 143–155, 2005.

[4] Sean N. Ghazavi, Thunshun W. Liao. Medical data mining by fuzzy modeling with selected features. Artificial Intelligence in Medicine: 43, 195—206, 2008.

[5] Belacel N, Boulassel MR. Multicriteria fuzzy assignment method: a useful tool to assist medical diagnosis. Artificial Intelligence in Medicine 21: 201—207, 2001.

[6] M.R. Smith, X. Wang, R.M. Rangayyan. Evaluation of the sensitivity of a medical data-mining application to the number of elements in small databases. Biomedical Signal Processing and Control 4: 262–268, 2009.

[7] Wen-Tsann Lin, Shen-Tsu Wang, Ta-Cheng Chiang. Abnormal diagnosis of Emergency Department triage explored with data mining technology: An Emergency Department at a Medical Center in Taiwan taken as an example. Expert Systems with Applications 37: 2733–2741, 2010.

[8] Janez Demsˇar, Blazˇ Zupan, Noriaki Aoki. Feature mining and predictive model construction from severe trauma patient's data. International Journal of Medical Informatics 63: 41–50, 2001.

[9] AndrewKusiak, Bradley Dixon, Shital Shah. Predicting survival time for kidney dialysis patients: a data mining approach. Computers in Biology and Medicine 35: 311–327, 2005.

[10] Gloria Phillips-Wren, Phoebe Sharkey, Sydney Morss Dy. Mining lung cancer patient data to assess healthcare resource utilization. Expert Systems with Applications 35: 1611–1619, 2008.

[11] Miguel Delgado, Daniel SaÂnchez, MarõÂa J. MartõÂn-Bautista. Mining association rules with improved semantics in medical databases. Artificial Intelligence in Medicine 21: 241—245, 2001.

[12] Po Shun Ngan, Man Leung Wong, Wai Lam. Medical data mining using evolutionary computation. Artificial Intelligence in Medicine 16: 73–96, 1999.

[13] Richard J. Roiger, Michael W. Geatz.DATA MINING A TUTORIAL-BASED PRIMER. Pearson Education.2003.

[14] Margaret H. Dunham. DATA MINING Introductory and Advanced Topics. Pearson Education.2003.

[15] Soukup, T., & Dabifdon, I. Visual data mining: Techniques and tools for data visualization and mining. New York: Wiley.2002.
http:/ / www.intersystems.com/

[16] Yu Hai-Yan, Li Jing-Song. Data mining analysis of inpatient fees in hospital information system. ITME2009, August 14.

[17] Chae, Y.M., Kim, H.S. Analysis of healthcare quality indicator using data mining and decision support system. Expert Systems with Applications, 2003, 24-, 167-172.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[18] Chae, Y.M., Kim, H.S. et al. Analysis of healthcare quality indicator using data mining and decision support system. Expert Systems with Applications, 2003, 24-, 167-172.

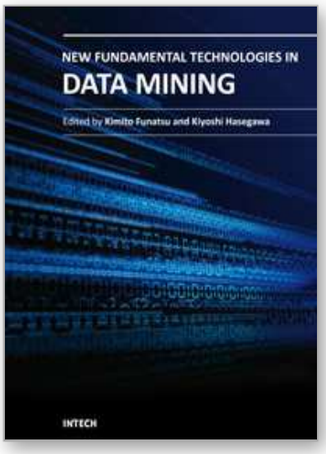
[19] Se-Chul Chun, Jin Kim, Ki-Baik Hahm, Yoon-Joo Park and Se-Hak Chun, Data mining technique for medical informatics: detecting gastric cancer using case-based reasoning and single nucleotide polymorphisms. Expert Systems, May 2008, Vol.25, No.2

[20] Delen, D., Walker, G. Predicting breast cancer survivability: A comparison of three data mining methods. Artificial Intelligence in Medicine, 2005, 34(2), 113-127.

[21] Koh H, Tan G. Data mining applications in healthcare. JHealthc Inf Manag, 2005, 19(2).

[22] Milley A. Healthcare and data mining. Health Manag Technol, 2000, 21(8).

[23] Ozcan YA.Forecasting. In: Quantitative methods in health care management, California: Josey-Bass; 2005.p.10-44 [Chapter 2].



New Fundamental Technologies in Data Mining

Edited by Prof. Kimito Funatsu

ISBN 978-953-307-547-1

Hard cover, 584 pages

Publisher InTech

Published online 21, January, 2011

Published in print edition January, 2011

The progress of data mining technology and large public popularity establish a need for a comprehensive text on the subject. The series of books entitled by "Data Mining" address the need by presenting in-depth description of novel mining algorithms and many useful applications. In addition to understanding each section deeply, the two books present useful hints and strategies to solving problems in the following chapters. The contributing authors have highlighted many future research directions that will foster multi-disciplinary collaborations and hence will lead to significant development in the field of data mining.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Jing-Song Li, Hai-Yan Yu and Xiao-Guang Zhang (2011). Data Mining in Hospital Information System, New Fundamental Technologies in Data Mining, Prof. Kimito Funatsu (Ed.), ISBN: 978-953-307-547-1, InTech, Available from: http://www.intechopen.com/books/new-fundamental-technologies-in-data-mining/data-mining-in-hospital-information-system

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821