Decision Support System In Medical Science
Using OLAP & Data Mining

Gajanan S. Raipure
M.E. (Computer science &Engg. Dept.) P.R.M.I.T,
College of Engg & Technology, Badnera,
Amravati.

Abstract

The clinical industry collects large amounts of data which, unfortunately, are not turned into useful information for effective decision making. For this purpose Decision support systems (DSS) use advanced technologies such as On-Line Analytical Processing (OLAP) and data mining to deliver advanced capabilities. In this paper a prototype clinical decision support system which combines the strengths of both OLAP and data mining. In this system will predict the future state and generate useful information for effective decision-making. With data mining, doctors can predict patients who might be diagnosed. OLAP provides a focused answer using historical data of concerned patients.

Key words:
Clinical decision support system, OLAP, Data mining, hybrid approach, Decision support system

1. Introduction

The healthcare industry faces strong pressures to reduce costs while increasing quality of services delivered. Oftentimes, information produced is excessive, incomplete, in the wrong place, inaccurate, disjointed or difficult to make sense [02]. A critical problem facing the industry is the lack of relevant and timely information [10]. As information costs money, it must adopt innovative approaches to attain operationally efficiently [15].

Decision Support Systems (DSS) have been developed to overcome these limitations. It supports business or organizational decision-making activities. However, they still do not provide advanced features to help doctors to perform complex queries [2, 17]. Advanced technologies can now generate a rich knowledge environment for effective clinical decision making. This paper presents a combined approach to diagnose the problem which combines the strengths of both OLAP and data mining.

2. Problem Statement

For recording business transactions On-Line Transaction Processing (OLTP) systems based on relational databases are suitable. They record information in two dimensions and automate repetitive tasks. Structured Query Language (SQL) is typically used to access information and results are presented in the form of reports which doctors use to make clinical decisions. Fig. 1 shows a Entity-Relationship Diagram (ERD) of OLTP schema consisting of six tables and Fig. 2 shows a SQL command for analyzing...
the relationship between hospitals and patients. OLTP has some major drawbacks. Large amounts of data in normalized form require many joins even to answer simple queries. For example, to analyze relationships between hospital and patients (Fig. 1), the query may require several table scans and multi-way table joins which can degrade performance significantly [3]. It requires at least four inner joins across (Fig. 2). A real-life database will have many tables and the time taken to process the joins will be unacceptable.

3. The Model

Any computer system that deals with clinical data or knowledge is intended to provide decision support. As OLAP uses several preprocessing operations such as data cleaning, data transformation, data integration, its output can serve as valuable data for data mining [3, 11]. OLAP operations (e.g., drilling, dicing, slicing, pivoting, filtering) enable users to navigate data flexibly, define relevant data sets, analyze data at different granularities.

**On-Line Analytical Processing (OLAP)** was introduced to overcome this problem. Whereas OLTP uses two-dimensional tables, OLAP uses multidimensional tables called data cubes. OLTP focuses on the automation of data collection procedure. Keeping detailed data, consistent and modern, is the most important condition for the application of OLTP [12]. However, in spite of OLAP is able to provide summary information efficiently, and how to take the final decision is still an art application of knowledge and common sense in some cases, the decision maker few quantitative data mining methods, such as regression or classification is introduced in OLAP. Data mining with OLAP integrated model system is divided into two parts:

a) **The server-side**: To build an integrated model.

b) **The Client-side**: For inquiries and for the results.

And with the help of OLAP user can filter, slice-and-dice, drill-down and roll-up data to search for relevant information efficiently. This paper presents a model for clinical decision support system based on OLAP and data mining to solve the problem of data association.
and visualize results in different structures [12, 8, 25]. Applying these operations can make data mining more exploratory.

### 3.1 Data Mining & OLAP Integration

Data mining and knowledge discovery in databases relate to the process of extracting valid, previously unknown and potentially useful patterns and information from raw data in large databases. “The analogy of “mining” suggests the sifting through of large amounts of low grade or data to find something valuable. The motivation for an integrated model, OLAP with datamining, is the concept hierarchy. Data in OLAP and decision tree are organized into multiple dimensions where each dimension contains multiple levels of abstraction defined by the concept hierarchy [8, 29]. The concept hierarchy is illustrated in Fig. 3, where each member has one root and all members between roots have parents and every branch ends with a leaf member. OLAP data cubes which store concept hierarchies can be used to induce decision trees at different levels of abstraction [29, 1]. Once the decision tree mining model is built, the concept hierarchies can be used to generalize individual nodes in the tree, which can then be accessed by OLAP operations and viewed at different levels of abstraction.

<table>
<thead>
<tr>
<th>Location</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Country</td>
<td>nepal, india</td>
</tr>
<tr>
<td>State</td>
<td>goa, assam</td>
</tr>
<tr>
<td>City</td>
<td>panji, malwan</td>
</tr>
</tbody>
</table>

**Fig. 3:** A concept hierarchy for the dimension location

### 4. Research and Observation

This research demonstrates how the integrated approach, Data mining & OLAP with data mining, provides advanced Decision Support. The research shows that by using the integrated model (OLAP with data mining) it is possible to:

1. Enhance real time indicators like bottlenecks.
2. Improve visualization to uncover patterns/trends that are likely to be missed.
3. Find out more subtle patterns in data over capabilities provided by OLAP or data mining alone.

**Fig. 4** shows the architecture of the integrated model (OLAP with data mining).

This architecture shows integrated model comparing with various components. This architecture has Four layers:

- **Layer 1:** It contains data repository where actual data is stored.
- **Layer 2:** It contains multidimensional database i.e. data cubes.
- **Layer 3:** It has rollup, drilldown, slice, dice, pivot, filter operations which is also called as OLAP operations.
Layer 4: It contain procedures & queries of data mining.

In this data is collected and it is first validate and cleaned then it is stored in data warehouse. Then this data in data warehouse is again filtered & integrated to form the data cube which is also called as OLAP cube.

From the data cube data then accessed using cube API. After that the data mining algorithm is used to predict & diagnosis probability of patient. It uses OLAP operations and the decision tree mining algorithm C4.5. The test data validates the effectiveness of the model.

6. System Implementation

Is the first creation of the data mining cube and then began the process of data collection. The cube keep the information and allows browsers on different theoretical levels. Serves as the source of the data for the task of data mining. Can be performed to extract the data on any level or dimension of the cube. After building a model is stored in a cube OLAP. Each representing a dimension of the rule corresponding to the node in the decision tree mining model (Fig.4). OLAP operations explain the different states of the system.

6.1 Data preparation

Data preparation process is roughly divided into data selection, data cleaning, formation of new data and data formatting.

6.1.1 Data Selection. A subset of data acquired is selected based on the following criteria:

a) Data quality properties: completeness and correctness.

b) Technical constraints such as limits on data volume or data type: this is basically related to data mining tools which are planned earlier to be used for modeling.

6.1.2 Data cleaning. The techniques used for data cleaning include:

a) Data normalization: Decimal scaling into the range (0,1), is used but even standard deviation normalization can also be used as a data normalization technique.

b) Data smoothing. Discretization of numeric attributes is used as data smoothing technique. This is helpful or even necessary for logic based methods.

c) Treatment of missing values. There is not simple and safe solution for the cases where some of the attributes have significant number of missing values. Generally, it is good to experiment with and without these attributes in the modeling phase, in order to find out the importance of the missing values. Simple solutions are: a) Replacing all missing values with a single global constant, b) replace a missing value with its feature mean, c) replace a missing value with its feature and class mean. If the missing values can be isolated to only a few features, then we can try a solution by deleting examples containing missing values, or delete attributes containing most of the missing values. The presented system uses the third method i.e. replace the

Fig 5: Overview of implementation of the system
missing value with its feature and class mean.
d) Data reduction. Reasons for data reduction are in most cases twofold: either the data may be too big for the program, or expected time for obtaining the solution might be too long. The techniques for data reduction are usually effective but imperfect. The most usual step for data dimension reduction is to examine the attributes and consider their predictive potential. Some of the attributes can usually be discarded, either because they are poor predictors or are redundant relative to some other good attribute. Some of the methods for data dimension reduction through attribute removal are: a) attribute selection from means and variances, b) using principal component analysis c) merging features using linear transform. The presented system uses the first approach i.e. attribute selection from means and variances.

6.1.3 New data construction. This step represents constructive operations on selected data which includes:
a) Derivation of new attributes from two or more existing attributes.
b) Generation of new records (samples).
c) Data transformation: data normalization (numerical attributes), data smoothing.
d) Merging tables: joining together two or more tables having different attributes for same objects.
e) Aggregations: operations in which new attributes are produced by summarizing information from multiple records and/or tables into new tables with “summary” attributes.

6.1.4 Data formatting. Final data preparation step which represents syntactic modifications to the data that do not change its meaning, but are required for the data mining task. These include:
a) Reordering of the attributes or records.
b) Changes related to the constraints of modeling tools: removing commas or tabs, special characters, trimming strings to maximum allowed number of characters, replacing special characters with allowed set of special characters.

6.2 Construct OLAP cube

The general idea of the approach is to materialize certain expensive computations that are frequently inquired, especially those involving aggregate functions, such as count, sum, average, max, etc., and to store such materialized views in a multidimensional database (called a “data cube”) for decision support, knowledge discovery, and many other applications. Aggregate functions can be pre-computed according to the grouping by different sets or subsets of attributes[18]. Values in each attribute may also be grouped into a hierarchy or a lattice structure. For example, “date” can be grouped into “day”, “month”, “quarter”, “year” or “week” which form a lattice structure. Generalization and specialization can be performed on a multiple dimensional data cube by “roll-up” or “drill-down” operations, where a roll-up operation reduces the number of dimensions in a data cube or generalizes attribute values to high-level concepts, whereas a drill-down operation does the reverse. Since many aggregate functions may often need to be computed repeatedly in data analysis, the storage of pre-computed results in a multiple dimensional data cube may ensure fast response time and flexible views of data from different angles and at different abstraction levels. Once the data is validated and cleaned, a data cube is built from the data.

6.3 Construct decision tree
Decision tree algorithms ID3 and C4.5 are chosen for building decision tree and predicting the probability and type of diabetes in a patient. The algorithm C4.5 represents “supervised learning” models with a known output used for comparison of the model output. It in fact, „prunes” away certain branches of the tree based on their significance. It also adds the discrimination of continuous attributes, the treatment of unknown attribute and production regulation, etc. The training data is a set $S = s_1, s_2, \ldots$ of already classified samples. Each sample $s_i = x_1, x_2, \ldots$ is a vector where $x_1, x_2, \ldots$ represent attributes or features of the sample. The training data is augmented with a vector $C = c_1, c_2, \ldots$ where $c_1, c_2, \ldots$ represent the class to which each sample belongs. At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. Its criterion is the normalized information gain (difference in entropy) that results from choosing an attribute for splitting the data. The attribute with the highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurs on the smaller sub lists[21]. The system also generates the decision tree using the ID3 algorithm and compares the trees generated by both the algorithms.

8. Conclusion and Future Work

This paper has presented a DSS based on OLAP with datamining. The system is powerful because (1) it discovers hidden patterns in the data, (2) it enhances real-time indicators and discovers bottlenecks and (3) it improves information visualization. Further work can be done to enhance the system. For example, features can be added to allow doctors to query data cubes on business questions and automatically translate these questions to Multi Dimensional expression (MDX) queries.

Besides decision tree, the use of other data mining techniques can also be explored.

References

[7] Ming-Syan, C., Jiawei, H. and Philip, S.Y., Data Mining: An Overview From a Database Perspective, IEEE Transactions on Knowledge and Data Engineering, 8(6), 866-883, 1996, December.


