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.

Dr. S. R.Gupta
Faculty of Computer science &Engg. Department 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 effectivedecision making. For this purpose Decision support systems (DSS) useadvanced technologies such as On-Line Analytical Processing(OLAP) and data mining to deliver advanced capabilities. In this paper a prototype clinical decision support system whichcombines 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, Datamining, hybrid approach, Decision support system

1. Introduction

The healthcare Industry faces strong pressures to reduce costs while increasing quality of services delivered. Oftentimes, information producedis excessive, incomplete, in thewrong place, inaccurate, disjointed or difficult to make sense [02]. A criticalproblem facing the industry is the lack of relevant andtimely information [10]. As information costs money, itmust adopt innovative approaches to attain operationalefficiently [15].

Decision Support Systems (DSS) have been developed toovercome these limitations. It supports business or organizational decision-making activities. However, they still do notprovide advanced features to help doctors to performcomplexqueries [2, 17].Advanced technologies can now generate a richknowledge environment for effective clinical decisionmaking. 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 onrelational databases are suitable. They record information in two dimensions and automate repetitive tasks. Structured Query Language(SQL) is typically used to access information and resultsare presented in the form of reports which doctors use tomake clinical decisions.Fig. 1 shows a Entity-RelationshipDiagram (ERD) of OLTP schema consisting six tables and Fig.2 shows aSQL command for analyzing the relationship betweenhospitals and Patients.

OLTP has some major drawback. Largeamounts of data innormalized form require many joins even to answer simplequeries. For example, to analyze relationships betweenhospital and patients (Fig. 1), the query may requireseveral table scans and multi-way table joins which candegrade performance significantly [3]. It requires at leastfour inner joins across (Fig. 2). A real-lifedatabase will have many tables and the time taken toprocess the joins will be unacceptable.

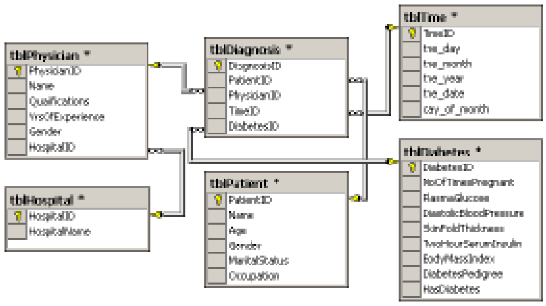

Fig.1: (a) ER diagram of OLTP schema

Question: Find the diabetics admitted to “WCK hospital” in “Jul” and “Aug”

Solution:

SELECT p.Name, h.HospitalName
FROM tblPatient p, tblHospital h,
tblTime t, tblDiagnosis d,
tblPhysician y
WHERE p.PatientID = d.PatientID AND
d.TimeID = t.TimeID AND
d.PhysicianID = y.PhysicianID AND
y.HospitalID = h.HospitalID AND
t.day_of_month="**JUL**" AND
t.day_of_month="**AUG**" AND
h.HospitalName="**WKC**"
GROUP BY h.HospitalName

Fig 2: SQL command for analyzing the relationship between hospitals& patients

On-Line Analytical Processing (OLAP) was introduced toovercome this problem. Whereas OLTP uses twodimensional 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 takethe final decision is still an art application of knowledge and common sense in some cases, thedecision 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 paperpresents a model for clinical decision support system based on OLAP and data mining to solvethe problem of data association.

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, itsoutput 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, definerelevant data sets, analyze data at different granularities and visualize results in different structures [12, 8, 25].Applying these operations can make data mining moreexploratory.

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 whereeach dimension contains multiple levels of abstractiondefined by the concept hierarchy [8, 29]. The concepthierarchy is illustrated in Fig. 3, where each member hasone root and all members between roots have parents andevery branch ends with a leaf member.OLAP data cubes which store concept hierarchies can beused to induce decision trees at different levels of abstraction [29, 1]. Once the decision tree mining model isbuilt, the concept hierarchies can be used to generalize individual nodes in the tree, which can then be accessed byOLAP operations and viewed at different levels ofabstraction.

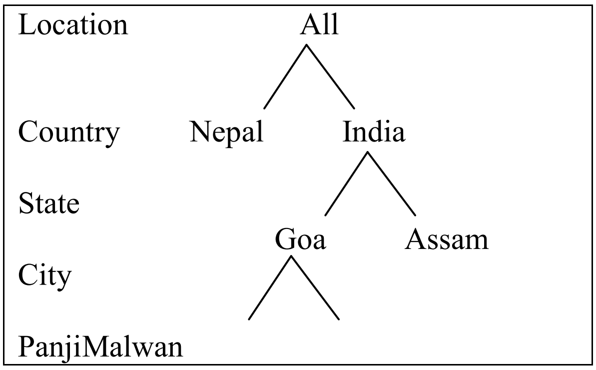

Fig. 3: A concept hierarchy for the dimension location

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

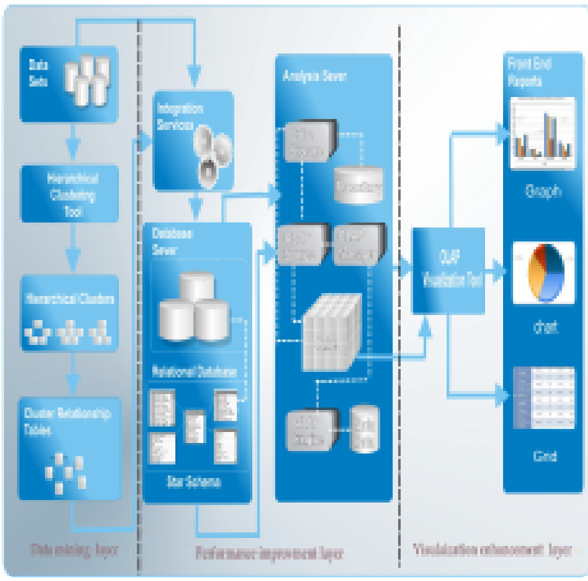

Fig. 4 shows the architecture of the integrated model (OLAP with data mining)

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.

Layer3: 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. Thecube keep the information and allows browsers on different theoretical levels. Serves as thesource of the data for the task of data mining.Can be performed to extract the data on any level ordimension of the cube. After building a model is stored in a cube OLAP. Each representing adimension of the rule corresponding to the node in the decision tree mining model (Fig.4). OLAP operationsexplain the different states of the system.

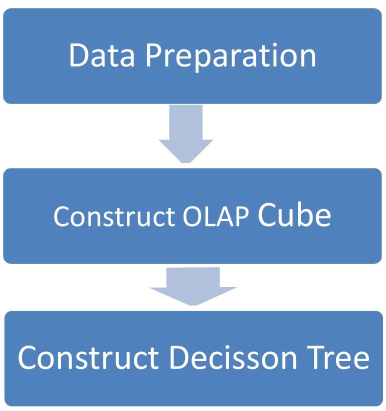

Fig 5: Overview of implementation of the system

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 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 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 = s1, s2,... of already classified samples. Each sample si = xl, x2,... is a vector where xl,x2,... represent attributes or features of the sample. The training data is augmented with a vector C = cl, c2,... where cl,c2,... 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 discovershidden patterns in the data, (2) it enhances real-timeindicators and discovers bottlenecks and (3) it improvesinformation visualization.Further work can be done to enhance the system. Forexample, features can be added to allow doctors to querydata cubes on business questions and automaticallytranslate these questions to Multi Dimensional expression(MDX) queries.

Besides decisiontree, the use of other data mining techniques can also beexplored.

References

[1] Helen, H. and Peter, H., Using OLAP and MultidimensionalData for Decision Making, IEEE IT Professional, 44-50,2001, October.

[2] Robert, S.C., Joseph, A.V. and David, B., Microsoft DataWarehousing: Building Distributed Decision SupportSystems, London: Idea Group Publishing, 1999.

[3] Surajit, C. and Umeshwar, D., An Overview of DataWarehousing and OLAP Technology, ACM Sigmod Record, *26*(1), 65-74, 1997.

[4] Ralph, K. and Margy, R., The Data Warehouse Toolkit. TheComplete Guide to Dimensional Modeling (2nd ed.),Canada: John Wiley & Sons, Inc, 2002.

[5] Torben, B.P. and Christian, S.J., Multidimensional DatabaseTechnology, IEEE Computer, *34*(12), 40-46, 2001, December.

[6] Usama, M. F., Data Mining and Knowledge Discovery:Making Sense Out of Data, IEEE Expert, 20-25, 1996,October.

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

[8] Han, J., OLAP Mining: An Integration of OLAP with DataMining, Proceedings of 1997 IFIP Conference on DataSemantics (DS-7), Leysin, Switzerland, 1-11, 1997, October.

[9] Blake, C.L. &Merz, C.J., UCI Repository of MachineLearning Databases, University of California, Departmentof Information and Computer Science, 1998.

[10] Sellappan, P., Chua, S.L., Ng, Y.H., Ng, Y.M., HealthcareInformation Services - The Application Service Provider(ASP) Model, Proceedings of SEARCC Conference 2004,Kuala Lumpur, October, 2004.

[11] Fayyad, U., Gregory, P.-S. and Smyth, P., From DataMining to Knowledge Discovery

in Databases, AI Magazine, *37*(3), 37-54, 1996.

[12] Parseye, K., OLAP and Data Mining: Bridging the Gap.Database Programming and Design, *10*, 30-37, 1998.

[13] Surajit, C., Umeshwar, D., and Ganti, V., DatabaseTechnology for Decision Support Systems, IEEE Computer, *34*(12), 48-55, Dec. 2001.

[14] Panos, V., and Timos, S., A Survey on Logical Models forOLAP Databases. ACM Sigmod Record, *28*(4), 64-69, Dec.1999.

[15] Donald, J.B., John, W.F., Alan, R.H., James, S., HealthcareData Warehousing and Quality Assurance, IEEE Computer,56-65, 2001, December.

[16] George, C., OLAP, Relational and MultidimensionalDatabase Systems, AcmSigmod Record, *25*(30), 64-69,Sept. 1996.

[17] Shim, J.P., Warkentin, M., Courtney, J.F., Power, D.J.,Ramesh, S., and Christer, C., Past, Present and Future ofDecision Support Technology, Elsevier Science B. V., *33*,111-126, 2002.

[18] Jonathan, C.P., Lobach, D.F., Goodwin, L.K., Hales, J.W.,Hage, M.L. and Hammond, W.E., Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse,Proceedings of the American Medical InformaticsAssociation Symposium, Philadelphia, United States of America, 101-105, 1997.

[19] Bansal, K., Vadhavkar, S., and Gupta, A., Neural NetworksBased Data Mining Applications for Medical InventoryProblems, International Journal of Agile Manufacturing,*2*(1), 187-200, 1998.

[20] Margaret, R.K., Kevin, C.D., and Ida, A., Data Mining inHealthcare Information Systems: Case Study of a Veterans' Administration Spinal Cord Injury Population, Proceedingsof the 36th Hawaii International Conference on System Sciences (HICSS'03), Hawaii, United States of America,IEEE Computer, 159-167, 2002.

[21] Hedger, S.R., The Data Gold Rush, Byte, *20*(10), 83-88, 1995.

[22] Bill, G. F., Huigang, L. and Kem, P. K., Data Mining for theHealth System Pharmacist. Hospital Pharmacy, *38*(9), 845 850, 2003.

[23] Usama F., Data Mining and Knowledge Discovery inDatabases: Implications for Scientific Databases.Proceedings of the 9th International Conference on Scientificand Statistical Database Management (SSDBM '97),Olympia, WA., 2-11, 1997.

[24] Raymond P.D., Knowledge Management as a PrecursorAchieving Successful Information Systems in Complex Environments.Proceedings of SEARCC Conference 2004,127-134, Kuala Lumpur, Malaysia.

[25] Han, J., Chiang, J.Y., Chee, S., Chen, J., Chen, Q., Cheng, S. & et al., DBMiner: A System for Data Mining in RelationalDatabases and Data Warehouses, Proceedings of the 1997Conference of the Centre for Advanced Studies onCollaborative research, Ontario, Canada, 1-12, November,1997.

[26] Sarwagi, S., Explaining Differences in MultidimensionalAggregate, Proceedings of the 25th InternationalConference on Very Large Data Bases, Scotland, UnitedKingdom, 42-53, September, 1999.IJCSNS International Journal of Computer Science and Network Security, 296 VOL.8 No.9, September 2008

[27] Fong, A.C.M, Hui, S.C., and Jha, G., Data Mining forDecision Support, IEEE IT Professional, *4*(2), 9-17,March/April, 2002.

[28] David K. and Daniel O'Leary, Intelligent ExecutiveInformation Systems. IEEE Expert, *11*(6), 30-35, Dec. 1996.

[29] Han, J., Kamber, M., Data Mining Concepts and Techniques,San Diego, USA: Morgan Kaufmann Publishers, pp. 294- 296.