More Than a Database: Mining Your Data for Decision-making Success

Save to myBoK

by Mehnaz Farishta, MS, BS

Are you drowning in data instead of sailing on the power of knowledge? Data mining technology can uncover hidden and unexpected patterns in data for strategic decision making in healthcare.

Here's how it works.

In today's environment, healthcare organizations need more than a database to be successful. To make the most of their information, organizations need staff who are empowered with critical thinking and who have the tools to manage data, information, and knowledge to make strategic decisions.

Even the most common of tools, an online transaction processing (OLTP) system, is not integrated for decision making and pattern analysis. As a result, many healthcare organizations struggle with the use of data collected.

Do you know what your top 10 paid procedures are? Do you know the DRGs that have the lowest reimbursement from Medicare/Medicaid? Are you receiving timely and accurate reports? Is your legacy system effective in identifying patterns to maximize your operational output? Are you drowning in data instead of sailing on the power of knowledge? This article will help you understand data mining technology and how it can uncover hidden and unexpected patterns in data for strategic decision making in healthcare.

Healthcare Data 101

Healthcare data can be classified into four categories:

- **patient-centric data**, which are directly related to patients
- **aggregate data**, which are based on performance and utilization/resource management data
- **transformed-based data** for planning, clinical, and management decision support
- **comparative data** for health services research and outcomes measurement

The data source is usually the OLTP system or operational support system of an enterprise. After years of accumulation of disparate data, these systems are data rich but information poor.[1]

With the adoption of data warehousing and data analysis/ online analytical processing (OLAP) tools, an organization can make strides in leveraging data for better decision making.

It should be noted here that these tools depend on users to guide the data investigation process. Moreover, the tools require a predefined starting point, such as a hypothesis, query, procedure, or program that dictates the nature of data analysis. These tools do not "uncover previously unknown business facts."[2] Therefore, there is a possibility that certain important information can remain untapped because no one knows of its existence.

Why Is Data Mining Important?

Data mining, also known as database exploration and information discovery, brings the practice of information processing closer to providing the real answers organizations are seeking from their data.[3] Healthcare organizations generate massive data, and these data have no organizational value until converted into information and knowledge, which can help control costs, increase profits, and maintain high-quality patient care.

Data mining provides automated pattern recognition and attempts to uncover patterns in data that are difficult to detect with traditional statistical methods. Without data mining, it is difficult to realize the full potential of data collected within an organization, as data under analysis are massive, highly dimensional, distributed, and uncertain.[4]

For healthcare organizations to succeed, they must have the ability to capture, store, and analyze data.[5] OLAP provides one way for data to be analyzed in a multidimensional capacity.[6] To understand the pivotal role data mining plays in a healthcare setting, it is important to first understand the role of a data warehouse, its various components, and how data mining methods can harness its power.

Introduction to Data Warehousing

A data warehouse can be considered a set of hardware and software components to analyze data for better decision making. Author and speaker Bill Inmon, known as "the father of the data warehouse," defines a data warehouse as a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management's decision-making process.[7] Typically, a data warehouse consists of a set of programs that extract data from the operational environment, a database that maintains warehouse data, and systems that provide data to users.[8]

A data warehouse enables users to tap into knowledge hidden in the massive amounts of data to understand business trends and make timely strategic decisions. If implemented correctly, a data warehouse can be the platform that contains an organization's data in a centralized and normalized form for deployment to users. This will enable them to perform simple reporting, complicated analysis, multidimensional analysis, population analysis, physician analysis and benchmarking, data mart deployment, clinical protocol development, reimbursement/cost analysis, and much more.[9]

Data Warehousing and Analysis Tools

To leverage an organization's operational data to assist in strategic decision making, data warehousing can be supported by decision support tools such as OLAP, data mining tools, and data marts. A data mart is a restriction of the data warehouse to a single business process or to a group of related business processes targeted toward a particular business group.[10]

OLAP solutions provide a multidimensional view of the data found in relational databases, which store data in a two-dimensional format.11 OLAP makes it possible to analyze potentially large amounts of data with very fast response times and provides the ability for users to "slice and dice" through the data and drill down or roll up through various dimensions as defined by the data structure.[12] Data mining tools automate the analysis of data to find unexpected patterns or rules that management can use to tailor business operations. With data mining, the system researches the data and determines hidden patterns and associations, while the analyst determines what to do with the results.[13]

Data Mining-A Closer Look

In the early days of data warehousing, data mining was viewed as a subset of the activities associated with the warehouse. Today, while a warehouse may be a good source for the data to be mined, data mining is recognized as an independent activity that is paramount for success with data warehouse-based decision support systems.

It should be noted that although warehousing and mining are related activities that reinforce each other, data mining does rely on a different set of data structures and processes and caters to a different group of users than the typical warehouse.[14]

Data mining enables users to discover hidden patterns without a predetermined idea or hypothesis about what the pattern may be. The data mining process can be divided into two categories: discovering patterns and associations and predicting future trends and behaviors using the patterns.[15] The power of data mining is evident, as it can bring forward patterns for which the user may not even consider searching. As a result, you have the answer to a question that was never asked. This is especially helpful when you are dealing with a large database in which there may be an infinite number of patterns to identify.

It is interesting to note that "the more data in the warehouse, the more patterns there are, and the more data we analyze the fewer patterns we find."16 What this means is that when there is richness of data and data patterns, it may be best to data mine different data segments separately, so that the influence of one pattern does not dilute the effect of another pattern in a large database.

At this junction, the user should not fear the loss of certain indicators or patterns in a smaller segment. If a trend or pattern is applicable in an entire database, then it is also evident in a smaller segment. Thus, depending on the circumstances and the nature of data in a warehouse, data mining can be generally performed on a segment of data that fits a business objective, rather than the whole warehouse.17

Types of Analysis Sessions in a Data Mine

In a data mine, analysis for pattern identification is done on a data segment, and the process of mining this data segment is called an "analysis session." For example, management may want to predict the response to a community flu vaccination initiative for the elderly by analyzing previous such initiatives, or they may want to know the patient sample over various geographic regions.

If a user starts the analysis with a specific task, for instance, analyzing reimbursement by DRGs, then this analysis is a "structured analysis." Structured analysis is generally done on a routine basis, like analysis of quarterly costs, revenues, and expenses, to identify and forecast trends. When the user wanders through a database without a specific goal, he or she may discover interesting patterns that were not conceived of previously. This type of analysis is termed "unstructured."18

Data Mining Techniques

A variety of types of techniques are used in data mining-including decision trees, genetic algorithms, neural networks, predictive modeling, rule induction, fuzzy logic, k-NN, and so forth. Considering the scope of this article, these techniques are discussed in a simple and general overview. Readers are encouraged to explore further depending on their interest.

A **decision tree** is a tree-shaped structure that visually describes a set of rules that caused a decision to be made. For example, it can help determine the factors that affect kidney transplant survival rates. **Genetic algorithms** are optimization techniques that can be used to improve other data mining algorithms so that they derive the best model for a given set of data. For example, algorithms can help determine the optimal treatment plan for a particular diagnosis.

Neural networks are nonlinear predictive models that learn how to detect a pattern to match a particular profile through a training process that involves interactive learning, using a set of data that describes what you want to find. For example, they can help determine what disease a susceptible patient is likely to contract. Neural networks are good for clustering, sequencing, and predicting patterns, but they do not explain why they have reached a particular conclusion.

Predictive modeling can be used to identify patterns, which can then be used to predict the odds of a particular outcome based upon the observed data. **Rule induction** is the process of extracting useful if/then rules from data based on statistical significance. **Fuzzy logic** handles imprecise concepts and is more flexible than other techniques. For example, it can help determine which patients are likely to respond to hospital/community initiatives to raise awareness of AIDS prevention. Finally, k-NN, or **k-nearest neighbor**, is a classic technique for discovering associations and sequences when the data attributes are numeric.19

Data Mining Architecture

In the early days, the "build a warehouse first, mine later" paradigm was generally followed by organizations. Classified by data mining expert Kamran Parsaye as the "sandwich paradigm," this "build first, think later" approach can be considered a "data-dump paradigm," because in many cases it leads to the construction of a "toxic data dump" whose contents are not easily salvageable. A much better method is to "sandwich" the warehousing effort between two layers of mining, thus understanding the data before warehousing it. This approach ensures that data mining complements a data warehouse.[20, 21]

Data mining can exist in three basic forms: *above the warehouse* where SQL statements are used to build a set of conceptual views above the warehouse tables; beside the warehouse where the data mining server follows the "three-level computing" approach; and within the warehouse as a separate repository where the mine uses a portion of the physical warehouse but is independent of the warehouse structure. Data mining above the warehouse provides a minimal architecture for analysis and identifying patterns, and is prescribed only when data mining is not a key objective for the warehouse.

Data mining beside the warehouse allows mining to be done in specific data segments with specific business objectives, and the exploratory nature of the mining exercise does not interfere with the warehouse's routine processes of query and reporting. Data are moved from the warehouse to the mine, restructured during the transformation, and then analyzed.

Data mining within the warehouse works like a "republic within a republic" and is independent of the warehouse structures but usually leads to loss of flexibility without any significant benefits. Thus, in most cases, the data mine beside the warehouse approach is prescribed. Stand-alone data mines can also exist in cases where it takes too long for an organization to build its corporate warehouse. In fact, the stand-alone data mine can be used later as a guide to design the warehouse architecture following the sandwich paradigm.[22, 23]

Requirements for a Data Mining Tool

When selecting a data mining tool or product, it is important to consider how well the product integrates with other components of a data warehouse. A two-way exchange interface with an OLAP tool is recommended; this will allow the OLAP tool to pass selected subsets of a data warehouse to the data mining tool. The data mining tool can mine for patterns that are difficult to observe using an OLAP tool, and it can forward its findings to the OLAP tool to be verified in a larger database.

It is also important for a data mining tool to support rule conversion. Converting mined rules into structured query language (SQL) queries allows OLAP and other analysis tools to use the mined rules immediately for data segmentation purposes. What's more, it makes the mined rules available for reuse by other OLAP applications.[24]

It is equally important for a data mining tool to have visualization capabilities. The success of data mining relies on how well the user can view, evaluate, and act on discovered patterns through the user interface tool. There must be a natural way for users to view the results of a data mining activity, and the user interface must integrate seamlessly into the user's current environment for ease of use. The power of automatic pattern discovery is enhanced if patterns can be viewed via charts and tables using, for example, color, position, or size differentiators.[25]

Other requirements need to be considered as well. The data mining product architecture must account for scalability and run-time performance in its design, and it should use multiprocessors to take full advantage of performance-enhancing technology.

The product must support a full range of data mining activities and techniques to enable users to see the same problem from many different angles for a thorough investigation. It must also support various possible data sources, both internal and external. Finally, it is important for the tool to support data preparation activities, such as data cleansing, data description, data transformation, data sampling, and data pruning.[26]

Using data mining technology and techniques can allow healthcare management and staff to analyze clinical and financial data for strategic decision making. This technology can enable organizations to meet their business objectives for cost control and revenue generation while maintaining high quality of care for their patients. Moreover, data mining tools and techniques can ease the burden of growing pressures of lower payer reimbursements, increased labor costs, and a highly competitive healthcare marketplace. These tools can work as strategic weapons to convert data into information, creating business intelligence in the process.

Notes

1. Shams, Kamruddin, and Mehnaz Farishta. “Data Warehousing: Toward Knowledge Management.” *Topics in Health Information Management* 21, no. 3 (2001): 24-32.

2. Ferguson, Mike. “Evaluating and Selecting Data Mining Tools.” *InfoDB* 11, no. 2 (1997).

3. Ibid.

4. Ibid.

5. Shams, K., and M. Farishta. “Data Warehousing: Toward Knowledge Management.”

6. Hettler, Mark. “Data Mining Goes Multidimensional.” *Healthcare Informatics* 13, no. 3 (1997): 43-46.

7. Lambert, Bob. “Data Warehousing Fundamentals: What You Need to Know to Succeed.” *Data Management Review* 6, no. 3 (1996): 24-30.

8. Ibid.

9. Shams, K., and Farishta, M. “Data Warehousing: Toward Knowledge Management.”

10. Ibid.

11. Hettler, M. “Data Mining Goes Multidimensional.”

12. Ibid.

13. Ferguson, M. “Evaluating and Selecting Data Mining Tools.”

14. Parsaye, K. “Data Mines for Data Warehouses.” *Database Programming and Design*, September 1996.

15. Ferguson, M. “Evaluating and Selecting Data Mining Tools.”

16. Parsaye, K. “Data Mines for Data Warehouses.”

17. Ibid.

18. Ibid.

19. Ferguson, M. “Evaluating and Selecting Data Mining Tools.”

20. Parsaye, K. “The Sandwich Paradigm for Data Warehousing and Mining.” *Database Programming and Design*, April 1995.

21. Parsaye, K. “Data Mines for Data Warehouses.”

22. Parsaye, K. “The Sandwich Paradigm for Data Warehousing and Mining.”

23. Parsaye, K. “Data Mines for Data Warehouses.”

24. Ferguson, M. “Evaluating and Selecting Data Mining Tools.”

25. Ibid.

26. Ibid.

Mehnaz Farishta *is a project manager for marketing and communications with The Shams Group, Inc. based in Coppell, TX. She can be reached at mehnazf@shamsgroup.com.*

Article citation:
Farishta, Mehnaz. "More than a Database: Mining Your Data for Decision-Making Success." *Journal of AHIMA* 72, no.10 (2001): 28-32.