IRACST – International Journal of Computer Networks and Wireless Communications (IJCNWC), ISSN: 2250-3501
Vol.2, No.3, June 2012

Study of Data Mining Architecture

Bharati Sanjay,
Student of Ph.D., Singhania University, Pacheri Bari, Dist. Jhunjhunu (Rajasthan), India

Dr. Sagar Jambhorkar

Dr. A.R. Dani
International Institute of Information Technology, Hinzwadi, Pune

Shinde Bhausaheb
Student of Ph.D., Singhania University, Pacheri Bari, Dist. Jhunjhunu (Rajasthan), India

ABSTRACT-Student, and a professional data mining consultant. The tools ran under the Microsoft Windows95, Microsoft WindowsNT, or Macintosh System 7.5 operating systems, and employed Decision Trees, Rule Induction, Neural Networks, or Polynomial Networks to solve two binary classification problems, a multi-class classification problem, and a noiseless estimation problem. Twenty evaluation criteria and a standardized procedurefor assessing tool qualities were developed and applied. The traits were collected in five categories: Capability, Learnability/Usability, Interoperability, Flexibility, and Accuracy. Performance in each of these categories was rated on a six-point ordinal scale, to summarize their relative strengths and weaknesses. This paper summarizes a lengthy technical report [1], which details the evaluation procedure and the scoring of all component criteria. This information should be useful to analysts selecting data mining tools to employ, as well as to.

INTRODUCTION: Data mining, *the extraction of hidden predictive information from large databases*, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.

Most companies already collect and refine massive quantities of data. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources, and can be integrated with new products and systems as they are brought on-line. When implemented on high performance client/server or parallel processing computers, data mining tools can analyze massive databases to deliver answers to questions such as, "Which clients are most likely to respond to my next promotional mailing, and why?"

This white paper provides an introduction to the basic technologies of data mining. Examples of profitable applications illustrate its relevance to today's business environment as well as a basic description of how data warehouse architectures can evolve to deliver the value of data mining to end users.

Scope of Data Mining: Data mining derives its name from the similarities between searching for valuable business information in a large database — for example, finding linked products in gigabytes of store scanner data — and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

- **Automated prediction of trends and behaviors**. Data mining automates the process of finding predictive information in large databases. Questions that traditionally required extensive hands-on analysis can now be answered directly from the data — quickly. A typical example of a predictive problem is targeted marketing. Data mining uses data on past promotional mailings to identify the targets most likely to maximize return on investment in future mailings. Other predictive problems include forecasting bankruptcy and other forms of default, and identifying segments of a population likely to respond similarly to given events.

- **Automated discovery of previously unknown patterns**. Data mining tools sweep through databases and identify previously hidden patterns in one step. An example of pattern discovery is the analysis of

retail sales data to identify seemingly unrelated products that are often purchased together. Other pattern discovery problems include detecting fraudulent credit card transactions and identifying anomalous data that could represent data entry keying errors.

Method Of Data Mining : To best apply these advanced techniques, they must be fully integrated with a data warehouse as well as flexible interactive business analysis tools. Many data mining tools currently operate outside of the warehouse, requiring extra steps for extracting, importing, and analyzing the data. Furthermore, when new insights require operational implementation, integration with the warehouse simplifies the application of results from data mining. The resulting analytic data warehouse can be applied to improve business processes throughout the organization, in areas such as promotional campaign management, fraud detection, new product rollout, and so on. Figure 1 illustrates an architecture for advanced analysis in a large data warehouse.

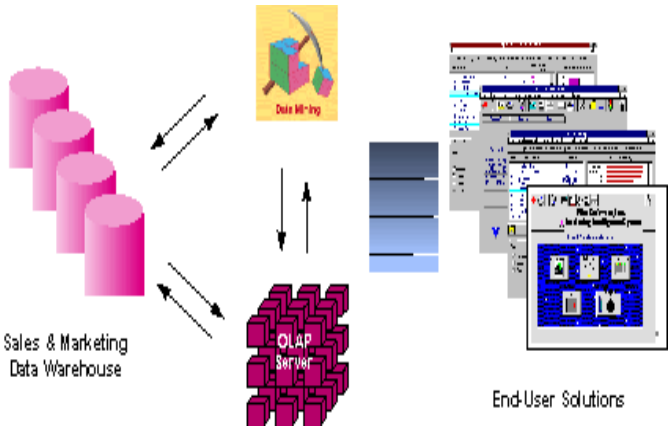

Figure 1 - Integrated Data Mining Architecture

The ideal starting point is a data warehouse containing a combination of internal data tracking all customer contact coupled with external market data about competitor activity. Background information on potential customers also provides an excellent basis for prospecting. This warehouse can be implemented in a variety of relational database systems: Sybase, Oracle, Redbrick, and so on, and should be optimized for flexible and fast data access.

An OLAP (On-Line Analytical Processing) server enables a more sophisticated end-user business model to be applied when navigating the data warehouse. The multidimensional structures allow the user to analyze the data as they want to view their business – summarizing by product line, region, and other key perspectives of their business. The Data Mining Server must be integrated with the data warehouse and the OLAP server to embed ROI-focused business analysis directly into this infrastructure. An advanced, process-centric metadata template defines the data mining objectives for specific business issues like campaign management, prospecting, and promotion optimization. Integration with the data warehouse enables operational decisions to be directly implemented and tracked. As the warehouse grows with new decisions and results, the organization can continually mine the best practices and apply them to future decisions.

This design represents a fundamental shift from conventional decision support systems. Rather than simply delivering data to the end user through query and reporting software, the Advanced Analysis Server applies users' business models directly to the warehouse and returns a proactive analysis of the most relevant information. These results enhance the metadata in the OLAP Server by providing a dynamic metadata layer that represents a distilled view of the data. Reporting, visualization, and other analysis tools can then be applied to plan future actions and confirm the impact of those plans..

CONCLUSION: Comprehensive data warehouses that integrate operational data with customer, supplier, and market information have resulted in an explosion of information. Competition requires timely and sophisticated analysis on an integrated view of the data. However, there is a growing gap between more powerful storage and retrieval systems and the users' ability to effectively analyze and act on the information they contain. Both relational and OLAP technologies have tremendous capabilities for navigating massive data warehouses, but brute force navigation of data is not enough. A new technological leap is needed to structure and prioritize information for specific end-user problems. The data mining tools can make this leap. Quantifiable business benefits have been proven through the integration of data mining with current information systems, and new products are on the horizon that will bring this integration to an even wider audience of users.

DISCUSSION: Clearly, data mining software can be designed to do well on both estimation and classification problems, as attested to by the performance of several of the tools across categories. It's puzzling that some developers would restrict their tools to one or the other realm alone. The asterisked cells in Table 2 are all situations where the training score was at least 0.10 better than the evaluation score. Often, that signals *overfit*, where over-training leads to worse generalization (performance on new data). In other cases, "more" training (using options which led to better training results) also led to better evaluation results, even though the evaluation was still much worse than training. A welcome

enhancement to future products would involve helping users to identify improperly fit models and, if possible, improve them. Evaluations of software applications are unavoidably subjective. Even using our scoring for the individual components, one could combine them with different emphases and arrive at different conclusions as to fitness to purpose. (Indeed, issues not considered here, such as computer environment, database connectivity, stability of vendor, etc., might also need to factor into a purchase decision.)

Acknowledgement:

Shinde Bhausaheb : I have completed my M.C.S.(Master Of Computer Science), M.Phil. Also Register to Ph.D. in Singhaniya Uniersity, Rajasthan. I am working in R.B.N.B. College as Head of Computer Science Department having 12 years of expert as well as Lecturer experience.

REFERENCES

[1] Gomolka, B., Schmidt, S., Summers, M., and Toop, K., *Data Mining Tool Evaluation: An Evaluation of Fourteen Tools Using Decision Trees, Rule Induction, Neural Networks, or Polynomial Networks*, Capstone rpt.

[2] http://www.salford-systems.com.

[3] http://www.cognos.com.

[4] http://www.rulequest.com.

[5] http://www.mathsoft.com.

[6] *WizWhy for Windows® User's Guide*. WizSoft

[7] http//www.datamindcorp.com.