THE ANNALS OF "DUNĂREA DE JOS" UNIVERSITY OF GALAȚI
FASCICLE I - 2006, Economics and Applied Informatics, Year XII, ISSN 1584-0409

INTEGRATING DATA MINING INTO BUSINESS INTELLIGENCE

Maria Cristina ENACHE

"Dunărea de Jos" University of Galati
mpodoleanu@ugal.ro

Data Mining is a broad term often used to describe the process of using database technology, modeling techniques, statistical analysis, and machine learning to analyze large amounts of data in an automated fashion to discover hidden patterns and predictive information in the data. By building highly complex and sophisticated statistical and mathematical models, organizations can gain new insight into their activities. The purpose of this document is to provide users with a background of a few key data mining concepts and business intelligence and about benefits of integrating business intelligence and data mining

Keywords: *Business Intelligence, platform, data mining*

Introduction

In general, data mining software assists and automates the process of building and training highly sophisticated data mining models, and applying these models to larger datasets. The data mining process involves the following steps:

1. Creating a predictive model from a data sample. A sample of data with a known outcome is extracted from the enterprise data store and pre-processed for the development of the predictive model. Advanced statistical and mathematical models are used to identify the significant characteristics and trends using the pre-processed fields as inputs, resulting in a predictive model. Generally, only a small subset of the all characteristics and trends in the sample data is used in the model.
2. Training the model against the dataset and its known results. The new predictive model is applied to additional data samples with known outcomes to validate whether the model is reasonably successful at predicting the known results. This gives a good indication of the accuracy of the model.
3. Applying the predictive model to a new dataset with an unknown outcome. Once the predictive model is validated against the known data, it is used for scoring, which is defined as the application of a data mining model to forecast an outcome.

For example, a data mining model that predicts the likelihood of a customer responding to a marketing campaign will generate a score for each customer that indicates his or her likelihood to respond. This score can be a simple binary result, such as a "Yes" or "No," or it could be a number indicating the propensity or confidence in that customer responding, say "97%." As mentioned earlier, the "Create-Train-Apply" process is typically the domain of the statistician or the data mining analyst. A solid understanding of data mining concepts, statistical concepts, techniques, and data mining tools is necessary in the "Create" and "Train" steps. Applying the predictive model requires less expertise, and is available for all business users.

Scoring Data for a Business Intelligence Application

There are three main approaches to integrating predictive insight into a BI [Business Intelligence] application:

This paper was recommended for publication by Assoc. Prof. Gabriela VIRLAN, PhD

1. Data mining tool scores the database. The data mining tool scores the records in a batch process, and saves them as columns in database tables. The BI application references the scored columns as required.
2. Database does the scoring. Database uses embedded scoring algorithms to score records in response to SQL queries from a BI application.
3. BI application does the scoring The BI Platform scores records using scoring metrics and reports.

All three approaches are viable methods for deploying data mining results throughout the enterprise. Determining which approach to use depends greatly on the business need for predictive analysis, and the IT infrastructure and philosophy.

The starting point for most data mining implementations is to use the data mining tool for scoring. Although it is very common for the data mining analyst to provide scores in standalone flat files or spreadsheets, integrating scored results into databases has long been a common practice.

When scoring is required on a real time basis, or when predictive models are created, and changed faster than scores can be calculated and stored in the database, one of the other approaches must be adopted. If the database supports data mining, deploying models in the database is a possible next step. If the BI Platform contains data mining capabilities, deploying models directly in BI applications can speed the adoption of predictive analysis by business users.

An BI platform such as Microsoft Strategy can integrate predictive insight into all BI applications used by business users. The following sections discuss the benefits and drawbacks of each approach.

1. Data Mining Tool Does the Scoring.
In this approach, a data mining scoring application calculates, and inserts scores into the database as new tables, new columns in existing tables, or updates to existing (old) scores. Once the scores are part of the database, the BI application reads these scores just like any other data, directly from the database. Historically, this approach has been the most common, and has the following benefits and drawbacks:

Benefits:
- Since a data mining tool does the scoring, model complexity, and performance is hidden within the scoring engine. The scoring process does not require any BI resources, and should not impact other concurrent BI processes.
- At runtime, BI applications simply read the scores from the database without having to calculate scores on the fly.

Drawbacks:
- Requires database space and database administrator [DBA] support.
- Large datasets can take a very long time to score.
- New records inserted after batch scoring are not scored.
- Updating the data mining model or scores requires more database and DBA overhead.
- Adding new or changing existing models requires rescoring the data.

2. Database Does the Scoring.
In this approach, data mining features inside the database management system perform the scoring. Several major databases have the ability to score data mining models. The most common approach is to import the predictive model into the database, and then generate scores by using extensions to SQL queries. A key feature of this approach is that the model can be imported and stored in the database. Several standards, such as the Predictive Model Markup Language (PMML), OLE DB for Data Mining, and the JSR-73 Java standards, enable the database to import of predictive models. The sophisticated techniques needed to create the model are not required to score the data. Scoring simply involves mathematical calculations on a set of inputs to generate a result.
This approach has the following benefits and drawbacks:

Benefits:
- Scores can be done “on the fly” even if new records are added.
- Updating the model is easier than having to re-score the entire database.

- Requires less database space than scoring the database since scores do not have to be persisted in the database.
- BI applications can take advantage of this approach by using the database's data mining capabilities directly.

Drawbacks:
- Requires database space and database administrator support.
- Requires application knowledge of the database's data mining capabilities. Typically, this is different from the database administration skills.
- BI applications must be customized for each database's data mining implementation.

3. Business Intelligence Tool Does the Scoring.

The third approach for integrating data mining uses enterprise data resources without significantly increasing the database overhead. This is accomplished by importing predictive models into the BI platform as standard metrics. Deploying predictive models in the BI platform allows sophisticated data mining techniques to be applied directly within the business intelligence environment on only the data that has been requested. Like the other approaches, it also has benefits and drawbacks:

Benefits:
- Scores can be done "on the fly" even if new records are added.
- Adding a new model or updating an existing model is simply a matter
- Does not require database space or database administrator support.

Drawbacks:
- Input characteristics need to be passed to the BI application even if they are not displayed on the report.
- Very large datasets may use a large amount of BI resources.

Applications of Data Mining Integrated with Business Intelligence

To understand the power of data mining and how business intelligence allows this information to be distributed to all relevant decision makers, it is helpful to look at various different use cases and business examples.

- **Market Basket Analysis** – This effective data mining modeling technique is used to determine items that are frequently sold together. Using association rules, a nationwide grocery store identified hidden patterns in buying behavior that had been previously overlooked. The implication of these findings suggested that store managers should place items that are often purchased together in key strategic locations across the store to promote the sales of these items.
- **Fraud Detection through Purchase Sequences** – A major credit card company introduced a new offering to protect their customers against fraud. They used a sequence association model to detect fraudulent purchases.

By analyzing historical data, they noticed that when a transaction for a gas purchase was followed by transactions for expensive luxury items, there was a high probability that these purchases were fraudulent. The new product offering used a series of these rules that identified potential fraudulent activity which protected their customers against unauthorized purchases.

- **Campaign Management** – A mail-order retailer wanted to improve the effectiveness of its direct mail marketing campaigns, with the goals of reducing costs and increasing the percent of positive responses. The retailer knew that it is too costly to send direct mail to all of its customers. Using a neural network model, they analyzed all of the factors that affect their customer's propensity to respond. The model included many variables, such as past purchase history, purchase frequency, customer age, gender, marital status, location, etc. and it was trained on a number of historical mailing campaigns. The model was then applied to the full list of customers and the probability of them responding to the campaign was predicted. Customers marked as most likely to respond were targeted in the new campaign.
- **Instant Credit Scoring** – A commercial bank wanted to automate the process of approving loan applications to save costs. Through a series or if-then rules using a decision tree, a credit score was generated for each new loan application which helped to identify whether it should be approved. This application decreased their costs by

employing fewer customer service representatives and increased customer service ratings by allowing customers to know instantly whether their loan had been approved.

• **New Restaurant Locations** – A nationwide fast food restaurant chain used data mining models to determine the best place to establish new restaurants. They grouped together all of the variables that are likely to influence the sales of a new restaurant. These included variables like: population size and demographics, competition, distance from other franchises, etc. Using a regression model, they were able to input a prospective restaurant location and estimate the potential growth and profitability of this location. They compared the outputs of various prospective locations to identify which had the highest profitability potential. Prior to using this data mining model, the company relied on educated guesses backed by the demographic data of the location.

• **Television Audience Share Prediction** – A nationwide television programming station needed to predict the audience share of a new TV program which was scheduled for broadcast at a particular time. With years of historical data containing audience share for each program shown in each time slot, a neural network model based on a large number of variables was developed to predict the audience share.

These variables included: the characteristics of the new program, such as genre, time of showing, target audience, cast, etc., the preceding and following programs with their characteristic information, other programs shown at the same time and the audience share, time of year, major public and sporting events, weather, etc. The model was able to predict audience share accurately which resulted in better sales opportunities for advertising slots.

• **Online Sales Improvement** – Online merchants rely on cross-selling and up-selling to increase their revenues. By relying on historical sales and user ratings of specific items, buyers are provided a choice of "similar" products when browsing specific items in the store. A nearest neighbor model generates "similarity' metrics which browse the product data warehouse for products nearest to a selected item, enticing the buyer to purchase additional items.

Benefits of Integrating Business Intelligence and Data Mining

While adding data mining scores and predictive models directly in the database is beneficial, there is additional value to be gained by integrating data mining scoring inside the BI platform.

- Business users can view predictive reports in a wide variety of user interfaces.
- Highly formatted predictive reports provide the easiest possible user consumption and professional presentation.
- Personalized messages and predictive reports can be delivered to very large user populations based on alerts or schedules.
- Ad hoc query and analysis that includes predictive metrics is possible without requiring knowledge of SQL, table structures, or predictive models.
- Business analysts can perform further analysis, such as slice-and-dicing data, ad hoc report creation, drilling, pivoting, and sorting, on predictive reports.
- Strict security is applied to users within, and outside the organization.

References:

- *[http1] www.microsoft.com - "An Architecture for Enterprise Business Intelligence - A Review of the MicroStrategy Platform Architecture for Reporting, Analysis, and Monitoring Applications"*
- *[http2] www.watchit.com*
- *[http3] www.patentstorm.us/*
- *[http4] www.microsoft.com – "MicroStrategy and Database support for functions"*