Using Standard APIs for Data Mining in Prediction

Jaroslav Zendulka
Brno University of Technology, Czech Republic

INTRODUCTION

There are three standardization initiatives concerning application programming interfaces (API) for data mining — OLE DB for Data Mining (OLE DB for DM), SQL/MM Data Mining (SQL/MM DM), and Java Data Mining (JDM) (Schwenkreis, 2001; Grossman et al., 2002; Grossman, 2004). Their goal is to make it possible for different data mining algorithm providers from various software vendors to be easily plugged into applications. Although the goal is the same for all the APIs, the approach applied in each of them is different. OLE DB for DM is a language-based interface developed by Microsoft, SQL/MM DM is an ISO/IEC standard based on user-defined data types of SQL:1999, and JDM, which is being developed under SUN’s Java Community Process, contains packages of data mining oriented Java interfaces and classes.

This short paper presents a simple example that shows how the APIs can be used in an application for prediction based on classification. The objective is to demonstrate basic steps that the application must include if we decided to use a given API. The example also helps to understand better different approaches the APIs are based on.

BACKGROUND

A brief characterization of all three APIs is presented in another article in this book (Zendulka, 2005). There are several introductory publications that deal with OLE DB for DM (Han & Kamber, 2001; Netz et al., 2001) and its referential implementation in Microsoft SQL Server 2000 (de Ville, 2001; Whitney, 2000; Tang & Kim, 2001). Keeping (2002) presented a scenario for using OLE DB for DM. The full specification was published by Microsoft in 2000 (OLE DB, 2000). Melton & Eisenberg (2001) presented an introductory paper to the SQL Multimedia and Application Package (SQL/MM), part of which the specification of data mining support known as SQL/MM DM is. The standard was accepted as ISO/IEC standard in 2002 (SQL 2002). JDM is developed as a Java Specification Request JSR-73. At the time of writing this paper, it was in the stage of final draft public review (Hornick et al., 2004).

Classification is a two-step process. In the first step, a model is built describing training data. The model can have a form of, for example, a decision tree. In the second step, the model is used for prediction. First, the predictive accuracy of the model is estimated using testing data. If the accuracy is considered acceptable, the model can be used to classify future (previously unseen) data (Han & Kamber, 2001).

These two steps are refined in the application according to the API used for implementation.

MAIN THRUST

Consider that based on a debt level, income level and employment type we want to predict the credit risk of a customer. A set of data is stored in a Customers table with columns CustomerID, DebtLevel, Income, EmploymentType, and CreditRisk. The CreditRisk column will be a target for prediction.

OLE DB for Data Mining

OLE DB for DM uses SQL CREATE, INSERT and SELECT statements with extended syntax and semantics in some cases to provide a language-based API for data mining services provided by a data mining provider that implements the required data mining technique. There are four basic steps that must be performed by our prediction application:

1. Define a data mining model.
2. Populate the data mining model.
3. Test the data mining model.
4. Apply the data mining model.

First, it is necessary to define a data mining model. OLE DB for DM provides a CREATE statement for this (the square brackets are name delimiters by convention for Microsoft SQL server):

```
CREATE MINING MODEL [RiskPrediction] %Model name
( [CustomerId] LONG KEY, %Source columns
  [DebtLevel] TEXT DISCRETE,
```

Copyright © 2006, Idea Group Inc., distributing in print or electronic forms without written permission of IGI is prohibited.
Related Content

Count Models for Software Quality Estimation
www.igi-global.com/chapter/count-models-software-quality-estimation/10843?camid=4v1a

Sentiment Analysis of Product Reviews
www.igi-global.com/chapter/sentiment-analysis-product-reviews/11061?camid=4v1a

Data Cleaning Based on Entity Resolution
www.igi-global.com/chapter/data-cleaning-based-on-entity-resolution/103253?camid=4v1a

Unsupervised Mining of Genes Classifying Leukemia
Diego Liberati, Sergio Bittanti and Simone Garatti (2005). Encyclopedia of Data Warehousing and Mining (pp. 1155-1159).
www.igi-global.com/chapter/unsupervised-mining-genes-classifying-leukemia/10771?camid=4v1a