Lacking standards for statistical and data mining models, applications cannot leverage the benefits of data mining.

DATA MINING STANDARDS INITIATIVES

The data mining and statistical models generated by commercial data mining applications are often used as components in other systems, including those in customer relationship management, enterprise resource planning, risk management, and intrusion detection. In the research community, data mining is used in systems processing scientific and engineering data. Employing common data mining standards greatly simplifies the integration, updating, and maintenance of the applications and systems containing the models. Established and emerging standards address various aspects of data mining, including:

Models. For representing data mining and statistical data.
Attributes. For representing the cleaning, transforming, and aggregating of attributes used as input in the models.
Interfaces and APIs. For linking to other languages and systems.
Settings. For representing the internal parameters required for building and using the models.
Process. For producing, deploying, and using the models.
Remote and distributed data. For analyzing and mining remote and distributed data.

The parameters of a parameterized data mining model, such as a neural network, can be represented using the Extensible Markup Language (XML); for example, the tag

```
<Neuron id="10">
    <Con from="0"
    weight="-2.08148"/>
```

indicates that a neural network node with id 10 has an input from a node with id 0 and a weight of −2.08148. The standards for defining parameterized models using XML are relatively mature. They assume the inputs to the models are given explicitly, as in the example. In practice, however, inputs are generally not explicit; the data must first be cleaned and transformed. But standards for cleaning and transforming data are only beginning to emerge. Standards related to the broader process of using data mining in operational processes and systems are relatively immature; for example, what is the business implication of a particular credit risk score produced by a credit card fraud model?

XML Standards

The Predictive Model Markup Language (PMML) is an XML standard being developed by the Data Mining Group (www.dmg.org), a vendor-led consortium established in 1998 to develop data mining standards [7]. PMML represents and describes data mining and statistical models, as well as some of the operations required for cleaning and transforming data prior to modeling. PMML aims to provide enough infrastructure for an application to be able to produce a model (the PMML producer) and another application to consume it (the PMML consumer) simply by reading the PMML XML data file.

BY ROBERT L. GROSSMAN, MARK F. HORNICK, AND GREGOR MEYER

TWO MAJOR CHALLENGES

top the data mining standards agenda: agreeing on a common standard for cleaning, transforming, and preparing data for data mining; and agreeing on a common set of Web services for working with remote and distributed data.

PMML consists of the following components:

Data dictionary. Defines the input attributes to models and specifies each one's type and value range.

Mining schema. Precisely one in each model, listing the schema's attributes and their role in the model; these attributes are a subset of the attributes in the data dictionary. The schema contains information specific to a certain model, while the data dictionary contains data definitions that do not vary by model. It also specifies an attribute's usage type, which can be active (an input of the model), predicted (an output of the model), or supplementary (holding descriptive information and ignored by the model).

Transformation dictionary. Can contain any of the following transformations: normalization (mapping continuous or discrete values to numbers); discretization (mapping continuous values to discrete values); value mapping (mapping discrete values to discrete values); and aggregation (summarizing or collecting groups of values, such as by computing averages).

Model statistics. Univariate statistics about the attributes in the model.

Models. Model parameters specified by tags. PMML v.2.0 includes regression models, cluster models, trees, neural networks, Bayesian models, association rules, and sequence models.

The first major release of PMML (v.1.0 in 1999) focused on defining XML representations for some of the most common statistical and data mining models. The assumption built into PMML v.1.0 was that the inputs to the models (called DataFields) were already defined. In practice, however, defining such inputs can be highly complex. The next major release of PMML (v.2.0 in 2001) introduced a mechanism, the transformation dictionary, to more flexibly define model inputs. In PMML v.2.0, inputs to PMML models can be DataFields defined in a data dictionary or DerivedFields defined in the transformation dictionary. The consensus among Data Mining Group members is that the transformation dictionary is powerful enough for capturing the process of preparing data for statistical and data mining models.

Standard APIs

To facilitate integration of data mining with application software, several data mining APIs have been developed for the following types of application:

SQL. The SQL Multimedia and Applications Packages Standard (SQL/MM) includes a specification called SQL/MM Part 6: Data Mining, which specifies a SQL interface to data mining applications and services. It provides an API for data mining applications to access data from SQL/MM-compliant relational databases.

Java. The Java Specification Request-73 (JSR-73) defines a pure Java API supporting the building of data mining models and the scoring of data-using models, as well as the creation, storage, and maintenance of and access to data and metadata supporting data mining results [5].

Microsoft. The Microsoft-supported OLE DB for DM defines an API for data mining for Microsoft-based applications [6]. Released in 2000, OLE DB for DM was especially noteworthy for introducing several new capabilities, variants of which are now part of other standards, including PMML v.2.0; included are taxonomies for data and a mechanism for transforming data. Earlier this year, however, OLE DB for DM was subsumed by Microsoft's Analysis Services for SQL Server 2000 [9]; Analysis Services provide APIs to Microsoft's SQL Server 2000 for data transformations, data mining, and online analytical processing (OLAP).

Other Standards Efforts

Standards have also been developed for defining the software objects used in data mining, the business processes used in data mining, and Web-based services for mining remote and distributed data.

Data mining metadata. In 2000, the Object Management Group defined the Common Warehouse Model for Data Mining (CWM DM) [1] for metadata specifying model building settings, model representations, and results from model operations, along with other data mining-related objects. Models are defined through the Unified Modeling Language [10] using tools to generate XML Document Type Definitions, which are used to specify formally XML documents.

Process standards. The CRoss-Industry Standard Process for Data Mining (CRISP-DM) was developed in 1997 by two vendors (ISL and NCR) along with two industrial partners. Designed to capture the data mining process, it begins with business problems, then captures and understands data, applies data mining techniques, interprets results, and deploys the knowledge gained in operations [2].

Web standards. The semantic Web includes the open standards being developed by the World Wide Web Consortium (W3C) for defining and working with knowledge through XML, the Resource Description Framework (RDF), and related standards [8]. RDF can be thought of informally as a way to code triples consisting of subjects, verbs, and objects. The semantic Web can in principle be used to store knowledge extracted from data though data mining systems, though this capability is, today, more a goal than an achievement.

The W3C is also standardizing Web services based on XML and a protocol for working with remote objects called the Simple Object Access Protocol (SOAP). The services describe themselves to applications using the Web Services Description Language [11].

Data webs are Web-based infrastructures employing Web services and other open Web protocols and standards for analyzing and mining remote and distributed data [4]. In addition to standard Web protocols, some data webs also use protocols designed to transport remote and distributed data, such as the Data Space Transport Protocol (DSTP) [3] being developed by the National Center for Data Mining at the University of Illinois at Chicago and standardized by the Data Mining Group.

Meanwhile, earlier this year, Hyperion, a software vendor, and Microsoft announced a set of XML message interfaces using SOAP to define the data-access interaction between a client application and OLAP or other data mining data provider [12].

Conclusion

The main reason so many different data representation and data communication standards exist today is that data mining is used in so many different ways and in combination with a so many different systems and services, many requiring their own separate often-incompatible standards. Although some vendor-led efforts have sought to homogenize terminology and concepts among standards, more work is indeed required.

Relatively narrow XML standards, such as PMML, serve as common ground for several emerging standards. For example, SQL/MM Part 6: Data Mining, JSR-73, CWM, and Microsoft's Analysis Services all use PMML in their specifications, providing a base level of compatibility among them all.

Meanwhile, two major challenges top the data mining standards agenda: agreeing on a common standard for cleaning, transforming, and preparing data for data mining (PMML v.2.0 represents a first step in this direction); and agreeing on a common set of Web services for working with remote and distributed data (an effort only just beginning). ■

References

1. Common Warehouse Metamodel: Data Mining. Object Management Group; see cgi.omg.org/cgi-bin/doclist.pl.
2. Cross Industry Standard Process for Data Mining (CRISP-DM); see www.crisp-dm.org.
3. Data Space Transfer Protocol. National Center for Data Mining; see www.ncdm.uic.edu.
4. Grossman, R. and Mazzucco, M. DataSpace: A data Web for the exploratory analysis and mining of data. *IEEE Comput. Sci. Eng.* (2002).
5. Java Specification Request 73; see jcp.org/jsr/detail/073.jsp.
6. OLE DB for Data Mining Specification 1.0. Microsoft; see www.microsoft.com/data/oledb/default.htm.
7. Predictive Model Markup Language (PMML). Data Mining Group, see www.dmg.org.
8. Semantic Web. World Wide Web Consortium; see www.w3c.org/2001/sw.
9. SQL Server 2000 Analysis Services. Microsoft; see www.microsoft.com/SQL/techinfo/bi/analysis.asp.
10. Unified Modeling Language. Object Management Group; see www.uml.org.
11. Web Services Activity. World Wide Web Consortium; see www.w3c.org/2002/ws.
12. XML for Analysis (XMLA); see www.xmla.org.

Robert L. Grossman (grossman@uic.edu) is director of the Laboratory of Advanced Computing and the National Center for Data Mining at the University of Illinois at Chicago and president of the Two Cultures Group, Chicago.

Mark F. Hornick (mark.hornick@oracle.com) is a senior manager in the Data Mining Technologies unit of Oracle Corp., Burlington, MA.

Gregor Meyer (gregorm@us.ibm.com) is a senior software engineer in the Business Intelligence unit of IBM Corp., San Jose, CA.