KDD-2003 Workshop on Data Mining Standards, Services and Platforms (DM-SSP 03)

Robert Grossman
National Center for Data Mining, University of Illinois at Chicago
and
Open Data Partners
Chicago, IL
grossman@uic.edu

ABSTRACT

At KDD 2003 a half day workshop was held on data mining standards and data mining services based on them. A theme of the workshop was that data mining standards have matured sufficiently that standards-based services and applications can now be deployed easily.

1. INTRODUCTION

Over the past several years, various data mining standards have matured and today are used by many of the data mining vendors, as well as by others building data mining applications. With the maturity of data mining standards, a variety of standards-based data mining services and platforms can now be much more easily developed and deployed. Related fields such as data grids, web services, and the semantic web have also developed standards based infrastructures and services relevant to KDD. These new standards and standards based services and platforms have the potential for changing the way the data mining is used.

Broadly speaking data mining standards and services have passed through three eras: in the first era, algorithms sat over flat files and no standards were necessary. In the second era, the development of languages such as PMML enabled a clean separation of the off-line development of the model and the on-line scoring or deployment. Today, we are entering a third era, in which languages such as PMML Version 2.0 are rich enough to support much of the data cleaning, normalizations, and transformations that often are more time consuming than the modeling per se. With standards rich enough to cover both the data preparation as well as the scoring, interest is turning to coordinating with the standards in near-by communities such as grid computing and web-services based computing.

Talks in the workshop covered current and emerging standards for statistical and data mining models, for data transformations, for building models, for workflow, and for related topics. In addition, the workshop included talks on requirements and on standards based data mining services and platforms.

2. DATA MINING STANDARDS

Gregor Meyer from IBM gave a talk on the Predictive Model Markup Language or PMML. In particular, he described the data transformations which are supported by PMML Version 2.1 and designed to cover the type of cleaning, normalization, binning, aggregating, etc. that are so important when preparing data for data mining. Gregor Meyer also spoke on the SQL/MM data mining standard. Mark Hornick from Oracle gave a talk on the Java Data Mining Standards JSR-73. Robert Chu from SAS and Zhaohui Tang from Microsoft gave a talk on the XML for Analysis, an emerging web services standard for data mining.

3. DATA MINING SERVICES

Robert Grossman from the University of Illinois at Chicago and Open Data Partners gave a survey talk introducing some of the data mining services and platforms that are now being developed, including data mining services for data grids, data webs, knowledge grids, semantic webs, and web-service based computing platforms. Michael Thess from prudsys AG gave a talk describing the XELOPES Data Mining Library, an open source data mining library available under the GNU GPL. XELOPES supports a variety of data mining standards and open source libraries, including, CWM, PMML, OLE DB for DM, (JDMAPI), MLC++, and Weka. Joesph Bugajski from Visa International described how standards based data mining services are used in financial services. There were also two papers which described data mining in the context of data grid applications.

4. FUTURE DIRECTIONS

One theme of the workshop is that web services provide a common ground for a variety of computing platforms and architectures, including web-service based distributed computing, data grids, data webs, semantic webs and knowledge grids. The Data Mining Group is exploring the creation a working group to provide a common ground for those interested in the standardization of web services for data mining.

Another message of the workshop is that although the data mining standards themselves have matured, building applications using them is still a challenge due to the relative lack of documentation that is available.

As data mining applications built on web services become more common, there will be a growing need for research whose goal is to develop web services which can scale gracefully with the large, distributed data sets that are becoming more common.