Data Mining Standards, Services and Platforms 2005 Workshop Report

Rick Pechter
MicroStrategy, Inc.
Carlsbad, CA
rpechter@microstrategy.com

ABSTRACT

This report is a summary of the workshop on Data Mining Standards, Services and Platforms, which was held at the KDD 2005 in Chicago on August 21, 2005. The workshop included several presentations on the application of data mining standards in various systems and platforms, as well as presentations on recent changes to the PMML standard, an opening panel discussion and a concluding round-table forum.

Keywords

KDD 2005, DM-SSP, Workshop, PMML, JSR-73, OLE DB for Data Mining, XML for Analysis

1. INTRODUCTION

The fifth workshop on Data Mining Standards, Services and Platforms was held as part of KDD 2005, the eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Since the first KDD conference, the use of data mining techniques has grown significantly, driven by applications in many fields, from astronomy and medical diagnostics to internet and business applications. Along with increased usage has come a need for standards that allow data mining models to be stored, shared and implemented in a consistent way. This workshop started as a forum for promulgating the Predictive Model Markup Language (PMML) data mining standard and over the years has helped to drive acceptance of this standard by the data mining community [1]. More recently, it has also been a forum for other data mining standards, including Java and service oriented approaches. This 2005 workshop was attended, as usual, by key software vendors in the industry. But it also saw the best representation of non-vendor organizations looking to apply these standards to their enterprises.

The on-line proceedings can be found at:

`http://www.thearling.com/dm-ssp-05.htm`

2. Future Directions for Data Mining Standards, Services and Platforms.

The workshop opened with a panel discussion on the future of data mining standards, services and platforms. Moderated by Dave Selinger from OverStock.com, the panel consisted of several authors who submitted papers for this year's workshop along with workshop chair Kurt Thearling from Capital One. It was clear from this discussion that acceptance of data mining standards is evident, from numerous vendors who support the PMML standard and use it to store models and exchange them with other applications. As a result, more customers of these products are taking advantage of these standard approaches by making them part of their operational processes. For example, Capital One is making PMML support a key vendor selection criterion as it allows this financial services company the ability to store and share models within their environment, and the ability to explain their internal processes to government regulators. One consensus was that the standards that allow data mining models to be instantiated in a consistent manner would allow greater use of data mining techniques and applications.

3. Data Mining Applications

Representatives from several vendors presented talks on how they and their customers are using data mining standards. I work for MicroStrategy, a business intelligence company that helps organizations turn their data into actionable knowledge for improve their performance and profitability. We have adopted the PMML standard so that data mining models from other vendors can be used to provide predictive forecasts in additional the historical perspective usually provided by BI applications. While organizations have usually scored records "off-line" in a batch processing approach, today there are more options available including scoring in the database and scoring the business intelligence "query-and-reporting" application. These new approaches take advantage of standards so the scoring can be done in real-time "on the fly" and without requiring data mining knowledge or tools to use and apply data mining models.

ZhaoHui Tang from Microsoft presented the new data mining features that will become available when Microsoft's SQL Server 2005 database is released later this year. This product will include nearly a dozen data mining algorithms and proprietary OLE DB for Data Mining extensions to SQL that allow data mining models to be created and applied. Microsoft also supports the XML for Analysis standard and, by using this service oriented standard, SQL Server 2005 data mining features can be accessed by XML for Analysis consumers.

Two vendors presented talks on deploying and managing models after they've been created. Tom Khabaza discussed the "Predictive Enterprise," where organizations make use of predictive analytics whenever they can be used in decision making situations. These models therefore represent a valuable asset to an organization, one that needs to be managed and protected like any other asset. SPSS provides a Predictive Enterprise Repository for just this purpose. This product includes version control, scheduling and auditing of predictive models.

Wayne Thompson and David Duling from SAS discussed how their product can help integrate the model development environment with the production infrastructure of an organization. SAS supports a number of ways to represent the various aspects of the data mining development environment in formats that are interoperable with production systems. While it often takes

literally months to successful deploy some models from development to production, SAS supports the generation of code, in the form of C, Java, PMML and SQL, that can greatly improve an organizations ability to leverage predictive analytics.

4. PMML Applications

Not surprisingly, several talks focused on the use and future direction of PMML. Stefan Raspl showed how IBM uses PMML to store models and also visualize these models in a number of ways. From neural networks to decision trees, these visualization techniques allow for better understanding of the model behavior. IBM includes additional data that, while not required for scoring the model, allows for better understanding of the data used to create the model, including histograms, charts and graphs of predictive variables.

Since PMML 3.0 was announced at KDD 2004 and released two months later, the Data Mining Group industry consortium that controls the standard has been working on the next release, PMML 3.1. Svetlana Levitan from SPSS discussed the recent changes in PMML, including improvements in missing confidence and missing value handling for Tree models, and various improvements to cluster models, transformations, data dictionary and the mining schema.

Bob Grossman presented a new proposal that is being considered for PMML 3.1. This new model type is an event based approach for change detection applications. The need for this model has been driven by business that need to detect changes in dozens, hundreds and even thoughts of different variables, whether they be continuous or discrete in nature.

5. The Evolution of PMML: Version 4.0 and Beyond

The workshop concluded with a round table discussion on the future of PMML. This discussion covered a wide variety of topics, including:

- The need for PMML to be responsive to the needs of vendors so users can be more successful in deploying models that use sophisticated features of data mining tools.
- The need for a PMML Reference Test or Stability Kit that would ensure model compatibility across vendors.
- Include supplement info that would be useful for tracking model behavior and understand conditions when the model was created (for example, date of training data collection, histograms on variables, lift charts, ROC curves, Confusion matrices, and economic data like the inflation rate, energy costs, interest rate.
- Exception handling for calculation errors (like divide by zero).
- Support for transformations that are beyond PMML's transformations and built-in functions (perhaps using inline SQL or C code).

By far, the most resounding comment was for "PMML to get serious!" Producers need ways to verify their PMML models conform to the standard, and consumers need the confidence to know that models deployed to one environment generate the same results as in the model creation environment. As PMML becomes richer in its support of predictive analytics and the data mining process, organizations need the confidence that interoperability between vendors is reliable and trustworthy.

6. References

[1] R. L. Grossman. Data Mining Standards Services and Platforms 2004 (DM-SSP 2004). SIGKDD Explorations, Volume 6, Issue 2

About the author:

Rick Pechter is director of the MicroStrategy Pacific Technology Center in Carlsbad, CA where he leads the team responsible for MicroStrategy's data mining features. He has been active with the Data Mining Group in the advancement of the PMML standard since early 2004. He previously worked at NCR's Teradata unit and has over twenty years experience in the data processing industry. He has a Bachelors of Science degree in Electrical Engineering from UC Irvine and a Masters of Science degree in Engineering Management from National Technological University.