IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITP4569

Analysis of Current Data Mining Standards

Janusz Swierzowicz
Rzeszów University of Technology
2 W.Pola, 35-959 Rzeszow, Poland
tel:+4817-8651424, fax: +4817-856-2519
jswierz@prz.rzeszow.pl

ABSTRACT

This paper examines the objective assumptions for Data Mining Process standardization, which simplifies integration of Information Systems with Data Mining models. In doing so it provides an overview of the more important characteristics of Cross Industry Standard Process Model for Data Mining (CRISP-DM), Application Programming Interface OLE DB for Data Mining (API OLE DB DM), and Predictive Model Markup Language (PMML).

INTRODUCTION

Information Technology development has strong effects on data resources. In this fast rising volumes of data environment, human abilities in memory capacities and low data complexity or dimensionality analysis cause data overload problem. It is impossible to solve this issue in a human manner – it takes strong effort to use intelligent and automatic software tools for turning rough data into valuable information [2-7,9-10]. One of the central activities associated with understanding, navigating and exploring the world of digital data is Data Mining. It is an intelligent and automatic process of identifying and discovering useful structures in data such as patterns, models and relations. We can consider Data Mining as a part of the overall Knowledge Discovery in Data process, which is defined as "the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" [4], it should support us as we struggle to solve data overload and complexity issues.

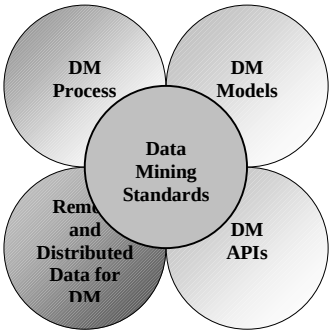
Data mining applications have to process data of diverse nature, drawn from different storage architectures, then use multiple data-specific exploration algorithms, and present results in a variety of forms. Data mining processes and models are used as a part of commercial Information Systems including those in enterprise resource planing, customer relationship management and in processing engineering and scientific data as well. With the fastest acceleration of online data resources in the Internet, the World Wide Web is a natural domain for using data mining techniques to automatically discover and extract actionable information from Web documents and services, especially in e-business. We have named those techniques as Web Mining. We also consider text mining as a data-mining task that helps us summarize, cluster, classify and find similar text documents.

Technological standards play an important role in Information Technology development [7]. Now, many organizations are developing technological standards for various aspects of data mining. Several standardization efforts [6] are undertaken on models, attributes, application programming interfaces, processing of remote and distributed data as depicted in Figure 1.

This issue is discussed in following chapters
Cross Industry Standard Process Model for Data Mining – CRISP DM

CRISP DM was developed in the year 2000 by a consortium of data mining vendors and advanced users (e.g. SPSS, NCR Daimler-Benz, Mercedes-Benz and OHRA) [3]. The CRISP-DM applies across different industry sectors (e.g. automotive, aerospace, insurance) was designed to make data mining projects easily adopted as a key part of business processes. The main assumption in this model preparation was its neutrality with respect to industry, method, tool and application. It consists of task described at four levels of abstraction: *phases, generic tasks, specialized tasks and process instances*.

Figure 1. Data mining standards in various aspects



At the top level, the data mining process is organized into the following phases:

- *Business understanding* that focuses on understanding the project objectives and requirements from business perspective,
- *Data understanding* that includes initial data collection, identification of data quality problems and detection interesting data subset to form hypotheses for hidden valuable information,
- *Data preparation* that covers construction of the data set for modeling tools. This phase focuses on tables, records and attributes selection as well as transformation and cleaning of data.
- *Modeling* that focuses on selection of various modeling techniques and on tuning for values of optimal parameters,
- *Evaluation* of the model quality with respect to achieving the business objectives,
- *Deployment* that involves applying models within decision making process in organization. It takes simple forms as reports generation as well as repeatable mining process.

The second level is the level of *generic tasks*. It was introduced to cover whole data mining process, all possible data mining applications and new modeling techniques e.g. [1].

The third level is the *specialized task*s. It describes how the general task differed in various situations.

The last but not least is the *process instance* level. It is a record of the actions, decision and results of an actual data mining engagement.

CRISP-DM distinguishes between following dimensions of data mining context:

- The *application domain* (e.g. banking, education, customer relationship management [2, 6, 13, 14]) is the area in which project take place,
- The *data mining problem type* (e.g. data description and summarization, segmentation, concept descriptions, classification, prediction, dependency analysis, etc.) describes the specific classes of objectives that the mining process deals with,
- The *technical aspect* (e.g. missing values) describes technical challenges that usually occur during data mining,
- The *tool and technique* that specifies which DM tools and/or techniques (e.g. Clementine, Poly Analyst, Weka [14]) are applied during the DM project.

APPLICATION PROGRAMMING INTERFACE OLE DB FOR DATA MINING

The API OLE DB DM is an example of a new protocol that simplifies communication and provides better integration of data mining tools with data based management applications. A virtual object that is similar to a table (the Data Mining Model DMM) can be created with CREATE statement, browsed with SELECT, populated with INSERT INTO, refined or used to derive prediction. A fundamental operation is the training of DMM, follow by use of the model to derive prediction [11,12]. The operation is executed in the following steps:

- Create an OLEDB data source and obtain an OLE DB session object
- CREATE MINING MODEL ...
- INSERT INTO //training data into the model
- SELECT ...
  FROM .....
  PREDICTION JOIN

PREDICTIVE MODEL MARKUP LANGUAGE (PMML)

Predictive Model Markup Language (PMML), managed by the Data Mining Group [6,15] is the most widely deployed data mining standard. It is based on an XML mark up language to describe statistical and data mining models. It describes the inputs to data mining models, the transformations used prior to prepare data for data mining, and the parameters that define the models themselves. It is used for a wide variety of applications, including applications in e-business, direct marketing, finance, manufacturing, and defense in products released by such vendors as Angoss, IBM, Magnify, Microsoft, MINEIT, NCDM, NCR, Oracle, Salford Systems, SPSS, SAS and Xchange. The current standard - PMML 2.0 - supports several predictive model types: Tree Model, Neural Network, Clustering Model, Regression Model, General Regression Model, Naïve Bayes Models, Association Rules Model, and Sequence Model. These categories cover the most popular data mining methods that are likely to find in contemporary data mining tools.

CONCLUSION

User participation in the standardization process is becoming more important. This issue should be also considered in the process of selection of data mining methods and tools.

REFERENCES

1. Abbass H.A., Sarker R.A., Newton C.S.: Data Mining. A Heuristic Approach, Idea Group Publishing, Hershey, London, 2002,
2. Berry M., Linoff G.: Data Mining Techniques, John Wiley & Sons, Inc, New York, 1997
3. Chapman P., Clinton J., Kerber R., Khabaza T., Reinartz T., Shaerer C., Wirth R. : CRISP-DM 1.0. Step -by - step data mining guide, CRISP-DM Consortium, 2000
4. Fayyad, U., M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy R.: From data mining to knowledge discovery: An overview. Fayyad, U., M. et al. (ed): Advances in Knowledge Discovery and Data Mining, AAAI Press/ The MIT Press, Menlo Park, CA, pp.1-34, 1996
5. Fayyad, U.: The Digital Physics of Data Mining, Communication of the ACM, March, 2001/Vol.44, No.3 pp.62-65
6. Grossman R.,L., Hornick M.F., Meyer G.: Data Mining Standards Initiatives, Communication of the ACM, August, 2002/Vol.45, No.8 pp.59-61
7. Jacobs K.: Global Aspect of Information Technology Standards and Standardization, Information Management, Vol. 15. No.1/2, 2002, pp.8-35
8. Landauer T. K.; "How much do people remember? Some estimates of the quantity of learned information in long-term memory," Cognitive Science, 10 (4) pp. 477-493 (Oct-Dec 1986).
9. Leavitt N: Data Mining for the Corporate Masses?, Computer, May, 2002/Vol.35, No.5 pp. 22-24
10. Liautaud B.: e-Business Intelligence: Turning Information into Knowledge into Profit, McGraw-Hill, New York, 2001
11. OLE DB for Data Mining Specification, Version 1.0, Microsoft Corporation, July 2000
12. Seidman C. : Data Mining with Microsoft® SQL Server 2000 Technical Reference, Microsoft Press, 2000
13. Swierzowicz J.: A Management Information System for Classification of Scientific Achievements, Evolution and Challenges in System Development, Zupancic et al (ed), Kluwer Academic/Plenum Publishers, New York, pp.735-740, 1999.
14. Swierzowicz J.: Decision Support System for Data and Web Mining Tools Selection, Issues and Trends of Information Technology Management in Contemporary Organizations, Khosrow-Pour M. (ed), Idea Group Publishing, Hershey, London, 2002, pp.1118-1120
15. www.dmg/org