CERIAS Tech Report 2001-80
Emerging standards for data mining
by Christopher Clifton
Center for Education and Research
Information Assurance and Security
Purdue University, West Lafayette, IN 47907-2086

ELSEVIER

Computer Standards & Interfaces 23 (2001) 187–193

COMPUTER STANDARDS & INTERFACES

www.elsevier.com/locate/csi

Emerging standards for data mining

Chris Clifton, Bhavani Thuraisingham *

The MITRE Corporation, 202 Burlington Rd., Bedford, MA 01730-1420, USA

Received 7 January 2001; received in revised form 7 March 2001; accepted 27 March 2001

Abstract

This paper presents an overview of data mining, then discusses standards (both existing and proposed) that are relevant to data mining. This includes standards that affect several stages of a data mining project. Summaries of several emerging standards are given, as well as proposals that have the potential to change the way data mining tools are built.

Keywords: Data mining; SQL; OLE DB; PMML; CRISP-DM

1. Introduction

Data mining is the process of extracting patterns as well as predicting (previously unknown) trends from large quantities of data by posing (automatically) repeated queries [1]. While various forms of data mining have existed for quite a while, it is only during the past decade that data mining has emerged as a technology area for a wide range of applications. For example, for decades, various organizations have been carrying out data analysis using statistical packages. Furthermore, neural networks and other machine-learning techniques have been applied to predict trends and extract patterns. While many of these techniques have become quite sophisticated, they have not scaled well. It is only recently that they are being applied to large quantities of data managed by database management systems. The merging of statistics, machine learning and database management has resulted in the emerging technology area called data mining. Various texts have appeared on data mining [1–4]. In addition, data mining research papers are published in various conference proceedings including the three major ones: Knowledge Discovery in Databases (held in North America), Principles of Knowledge Discovery and Data Mining Conference (Europe) and the Pacific Asia Knowledge Discovery and Data Mining Conference (Australasia).

Since data mining is now becoming a mature technology, it is important that appropriate standards be established for various aspects of data mining. For example, data mining processes have been developed. These processes are yet to be standardized. One needs to examine whether the various processes model could be applied for modeling the data mining

* Corresponding author. Tel.: +1-781-271-8873; fax: +1-781-271-2352.

E-mail addresses: clifton@mitre.org (C. Clifton), thura@mitre.org (B. Thuraisingham).

0920-5489/01/$ - see front matter
PII: S0920-5489(01)00072-1

process. Another group has developed various languages for data mining. For example, Structured Query Language (SQL) extensions are being proposed. However, these extensions are yet to be made standards for data mining. Architecture for data mining is also being examined. One needs to determine whether the various standards emerging from consortiums may be applied for data mining. Finally, data mining is becoming a key technology for e-business. The various standards for e-business need to be examined for relevance to data mining. In summary, as we make more and more progress in data mining, we cannot avoid standardization. Standardization will enable standard methods and procedures to be developed for data mining so that the entire process of data mining could be made easier for different types of users.

This paper addresses how standards may be applied to data mining. In Section 2, we discuss what data mining is, including data mining technologies, process, and directions. Section 3 discusses emerging standards for relevant to data mining tools and process. Sections 4 and 5 discuss other areas where standardization could affect data mining. The paper is summarized in Section 6.

2. Overview of data mining

For data mining to be effective, several technologies have to work together. First of all, statistical analysis and machine-learning techniques have to be applied successfully to databases to extract patterns and to predict trends. Visualization techniques are important to provide visual understanding of data, patterns and trends and subsequently guide the user in carrying out further data mining. Data warehousing is a critical technology for organizing and cleaning the data to prepare for mining. Parallel processing techniques provide important enabling technology to speed up the mining process for large-scale data sets. Network-computing infrastructures are an important consideration especially for distributed data mining. That is, various technologies have to be integrated to carry out successful data mining, leading to a need for standards.

Before carrying out data mining, there are several steps that one needs to consider. First of all, what is expected of the mining process? Do we want to form clusters, make associations, or classify the data? Are there commercial tools that can be applied? If so, what techniques do we use to get the desired outcomes? For example, should these techniques be decision trees or neural networks? If not, do we develop the tools in-house? If we do not want to develop the tools in-house, then can we contract the work outside? Once we get results, how do we know that the results are good? How do we prune and only get the useful results? All these questions have to be addressed to carry out successful mining. In addition, we also need to have good quality data. Therefore, it is widely recognized that a high-quality data warehouse is a necessary condition of successful mining.

As a result of the developments in data mining during the past decade, numerous commercial products and research prototypes have been developed. Most major database management system vendors as well as data analysis vendors are now marketing data mining tools. Many of these tools work on relational databases. That is, they assume that the data are placed in tables and the tools are geared towards manipulating the tables (or in many cases, a single-table view of the data). Various success stories have also been reported (see for example Ref. [5]). One of the major challenges now is to mine unstructured databases where it could be text, image, video or a combination of all of these. Another challenge in data mining is web mining. There are two aspects here. One is to mine the vast quantities of multimedia data on the web and extract meaningful information (web content mining). The other is to mine the usage patterns and give advice to the users as well as to those who want to carry out commerce/business on the web (web usage mining). A third challenge is mining distributed and heterogeneous databases. This is because databases are scattered within and across many organizations and it may be infeasible to bring them together into a centralized warehouse. Therefore, the distributed and sometimes disparate data sources have to be mined. While data mining has many valuable applications in many areas, there are also some negative aspects and that is compromising privacy. Data mining tools may be applied to deduce sensitive information and therefore compromise privacy and security. This is another major challenge facing data mining as well as security technologists.

3. Standards applicable to data mining

To handle all of these challenges and make progress in data mining, one needs effective standards for various aspects of data mining. What do we mean by a "data mining standard"? As we have seen, there are many different tasks involved in a data mining project. Standardizing the task and results becomes difficult. For example, if we define a standard classification model, we ignore a variety of other types of pattern discovery, such as rule discovery or clustering, that also qualify as data mining. The result is that there is currently no attempt for a single "standard" for data mining, but instead standards to support different aspects of data mining. These can broadly be divided into:

- Standards for the task to be performed (e.g. a formal definition of inputs to and outputs from the training and use phases of a classifier);
- Standards for supporting technology (e.g. SQL as a standard for data access); and
- Process standards (e.g. what is the sequence of events in performing a data mining project?).

Other areas for applying standards include developing standard architectures for data mining and web-data standards. These two areas will be addressed in Sections 4 and 5, respectively.

3.1. SQL and data mining

Typically, when we think of standards in data management, we think of languages and protocols that allow information exchange between applications and systems. Within the database community, the premier standard is SQL. By itself, and in conjunction with related standards such as Java DataBase Connectivity (JDBC) and Open DataBase Connectivity (ODBC), SQL is having an influence on the data mining community. Historically, data mining tools operated on flat-file data in fielded or comma-separated value formats. Most are moving to support data access through ODBC or JDBC, however. These provide a way for tools to get at data. However, SQL is designed for transaction-oriented access to data: retrieval or update of small data sets based on a query. Data mining operates over large data sets. While SQL can easily generate such data sets, the actual APIs such as ODBC and JDBC are poorly equipped for retrieval of huge quantities of data.

Many data mining tools operate by copying all the relevant data either into memory, or into their own disk storage, then operating on that data. This is wasteful—better would be to make use of the database for storage and retrieval of the data (even during the running of the data mining algorithm), and perhaps to offload some of the algorithmic tasks to the database, where prebuilt indexes and the like may enable better optimization. Some tools (such as IBM's Intelligent Miner for Data) are beginning to do this—but through tight, proprietary integration with the database. A standards-based approach would be better, but requires an understanding of the types of access patterns made by data mining tools.

There have been several proposals to add operations to SQL to support data mining. The most common of these are based on the notion of a "data cube" [6]. A data cube is a collection of data, where each axis represents a particular "selection criteria", and a point in the cube is the value where all selection criteria meet. For example, Fig. 1 shows a three-dimensional data cube, where the dimensions correspond to month, region, and department. The values within the cube correspond to "total sales"—for example, for February in the south, meat had sales of US$150, and in January in the Northeast, produce had sales of US$100.

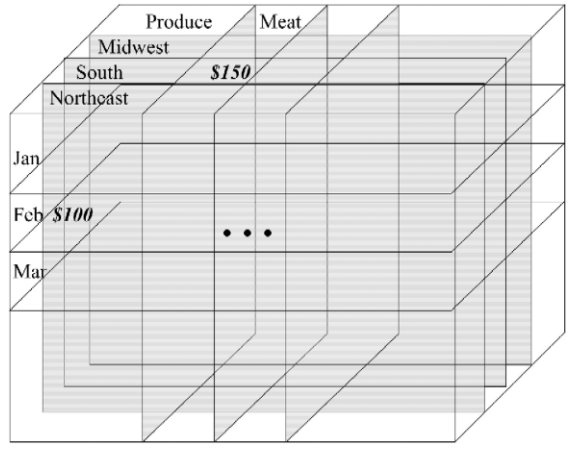

Fig. 1. Sample data cube.

The key to a data cube is that it quickly provides answers to aggregates—for example, we may want sales for all months for New England and the Produce department. This sort of aggregation is a useful building block in many data mining algorithms. The idea is that data mining algorithms could use the data cube to get only the needed aggregate information, instead of retrieving the entire set of information from the database. Commercial products (particularly those intended for data warehouse applications) are beginning to include data cube concepts; however, a standard for these extensions does not yet exist, and without such standards, it is unlikely that data mining tools will take advantage of these features.

3.2. Data mining model standards

Existing database standards provide a way for data mining tools to get to the data—improvements will make tools more efficient, but will not really add new capabilities. At the other end of the tool, however, we have a different story. The output of today's data mining tools vary widely—text-based reports, lists of rules in a pseudo-English format, visual displays, or even binary files that enable proprietary tools to do certain tasks (e.g. classification). There is no way to build "generic" applications that make use of the output of data mining tools. To some extent, this is reasonable—since data mining may be used to produce widely different things (e.g. a classifier that divides new items into known groups based on the training source data, or a list of unusual entries in the source data), a single standard for data mining results is unlikely. However, progress is being made in this area.

The Data Mining Group has developed Predictive Model Markup Language [7], an eXtensible Markup Language (XML)-based specification language for predictive models (classifiers). A PMML specification consists of several parts.

- A Data Dictionary. This names and defines the types of the input and output fields of the model (e.g. "Salary", continuous, range 0–10,000,000). Note that multiple models can share a single data dictionary.
- Mining Schema. This defines the particular entries in the data dictionary used as input and output by a particular model. In addition to specifying which are input, and which are the predicted (output) values, it also may specify a range of accepted values and how values outside that range are to be treated (e.g. an unusually low salary may be treated as missing.)
- Statistics. Contains statistics about a single field. Examples would be the minimum, maximum, mean, standard deviation, and median for numeric attributes. This is not required, but is relevant for some models.
- Normalization. Some tools may expect inputs to be in a fixed range (e.g. 0–1). If so, the normalization component of the model describes how this is to be done for each field in the Mining Schema.
- The actual model. There are several types of models—Tree Classification, Polynomial Regression, General Regression, Association Rules, Neural Networks, and clustering. For example, a tree model consists of a Node, each containing a predicate (a Boolean expression determining if that node is selected), a list of subsidiary nodes (evaluated if the predicate is true), and a score (the result if the predicate is true and none of the subsidiary nodes are selected.)

A coalition of organizations headed by Microsoft is supporting this with OLE DB for Data Mining [8], an extension of Microsoft's OLE DB database access standard. The idea is to represent the output of a data mining model as a table. This "prediction table" is created by providing a prediction model and an input table (the Prediction join operation). The prediction model can either be created directly by an OLE DB DM compliant data mining tool, using a "create model" statement in the database than runs the tool on the chosen input data, or from an XML model specification given in a variant of PMML. The structure of the "prediction table" is defined by two things: the input table, and a formal definition associated with the tool that defines the output in terms of input for that tool. This works well for predictive modeling (classification), but extension to other types of data mining may need work.

The Java community is working on a similar standard, the Java Data Mining API [9]. This is also expected to be compliant with PMML.

While of great benefit, these standards do pose the risk of limiting the scope of data mining. Vigilance is required to ensure that as data mining tools develop new capabilities, the standards are extended

(or new standards are created) to support those new capabilities.

3.3. Process standards for data mining

While tool interoperation is a valuable goal, it is not the only area where standardization can benefit data mining. Actual tool use is a relatively small cost in typical data mining projects—even with the effort required to connect those tools to the data. Other factors in the overall process (see Fig. 2) dominate the total cost. For example, deciding what data should be mined, and bringing it all together in a data warehouse so that related items have a common semantics, can be a multi-year effort. Data cleansing (often part of the warehousing process) is also difficult—and knowing when the data is "clean enough" can be difficult (for example, association rule learning can be quite tolerant of randomly distributed errors). Interpreting the results is also difficult—for example, an intriguing result may actually arise from common errors in the data (such as always entering January 1 if the actual date is not known), and may require further cleansing of the data—tool support can help here (such as means of visualizing the results). Perhaps the most difficult step of all is putting the results into practice—changing business processes based on the results. Finally, we need to analyze the effect and determine whether, and how, to proceed with the next cycle. Currently, these steps are carried out in an ad-hoc fashion. There are no software engineering methods for mining. The questions is, can we apply various models such as the waterfall model or the spiral model for data mining?

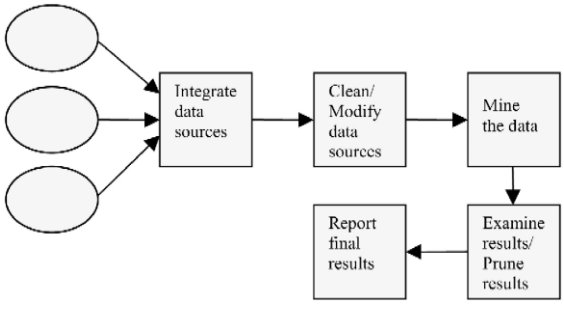

Fig. 2. Data mining process.

There is one notable effort in this area. A consortium of data mining vendors and early adopters of data mining technology, through a European Commission funded effort, have developed the Cross-Industry Standard Process for Data Mining [10]. This is a hierarchical process model that breaks the data mining process into several phases, each with a variety of tasks. These phases are:

1. Business understanding. Determine business objective, assess situation, determine data mining goals, and produce a project plan.
2. Data understanding. Collect initial data, describe the data, explore data, and verify data quality.
3. Data preparation. Select data, clean data, construct data, integrate data, and format data.
4. Modeling. Select modeling technique, generate test design, build model, and assess model.
5. Evaluation. Evaluate results, review process, and determine next steps.
6. Deployment. Plan deployment, plan monitoring and maintenance, produce final report, and review project.

The CRISP-DM user manual further subdivides the tasks in each phase, defines the output and required activities for each, and provides hints on potential pitfalls along the way. While not yet to the level and detail of some such standards (e.g. ISO 9000 quality standards), it serves as a good base. More formal metrics, such as ways of measuring data quality to determine when data cleansing is sufficient, could be useful here. Such formal efforts would also help in developing tools to support the various tasks.

4. Architecture standards for data mining

In the area of architecture standards for data mining, there are various dimensions. One is the relationship between data mining and related technologies such as database systems, decision support, and data warehousing. What are the interfaces say between a data manager and a data miner? Can one

Fig. 3. Three-tier architecture for data mining.

standardize these interfaces? Another area is to standardize the functional architecture for data mining. What are the data mining functions and how can we develop standards? The third area is to develop a three-tier middleware system. The front-tier is the client-tier. The middle tier is the business objects tier and consists of business object for data mining. The third tier may be the database server tier. One could use distributed object systems to integrate the various layers.

Fig. 3 illustrates the three-tier architecture for mining. There is still very little discussions about standardizing the data mining architecture. However, the Object Management Group is involved in specifying object-based standards for data mining. For further details, we refer to Ref. [11].

5. E-business standards and data mining

While data mining has been developing over the past decade, there has been an explosion during the last few years. Much of this is due to the rise of the Web and e-commerce. E-commerce generates large transaction databases—fertile ground for data mining—and competitive pressures drive the desire to obtain knowledge from these data. We are now hearing the term e-business. Many companies prefer to be doing e-business rather than e-commerce, as e-commerce is perceived to be too narrow. Those who differentiate between e-business and e-commerce state that e-commerce is all about carrying out transactions on the web. But e-business is much broader and includes learning and training, entertainment, putting up web pages and hosting web sites, conducting procurement, carrying out supply chain management, help desk services—any business model making use of the web as a core component of extra-company interaction.

The result of this is an explosion in the amount of data being gathered. We are now seeing not only transaction data, but web content, usage patterns ("click-stream data"), and text records of interaction (e.g. chat rooms, help desk records.) Corporations want to maintain a competitive edge and are exploring numerous ways to market effectively. Major corporations including retail stores have e-business sites and customers can now order products from books to clothing to toys through these sites. E-business sites collect massive amounts of data on customer purchases, browsing patterns, usage times, and preferences; each site can also collect information on competitors' offerings and prices. Based on the information, a site can adjust its assortments, prices and promotion quickly and dynamically to respond to the changing trends, competitor's strategy and personalization rules. As an example, companies can now sponsor "chat rooms" and analyze the text streams to improve marketing—like a focus group, but on a much grander scale. This is being done today, but with manual analysis of the text. The opportunities for data mining technology are obvious —but where is the structured, tabular data? Although some of the data are structured, much of the data on the web are either free-form (text, images). Even the tabular data (such as product/price lists) are formatted for display rather than processing—extracting tabular data from web pages, in a form suitable for further processing, is a challenging task.

There are attempts to develop standards for e-business. eXtensible Markup Language (XML), in particular, is hyped as a panacea for the interoperability problems of the web. While XML by itself solves few problems, metadata standards based on XML do provide hope. The challenge for the data mining community is ensuring that these standards will capture the information needed to support data mining, and in a form that supports feeding data mining tools from data captured in that standard. This is a particular difficult problem, as data mining of textual data is a novelty (with only a few vendors in the market), so knowing if a data standard is

"good enough" for data mining is difficult. Mining of other media (images, video, audio) is even less well understood.

6. Summary and directions

Data mining is a new and rapidly developing technology. Given the wide variety of tasks data mining can perform, it is difficult to come up with a data mining "standard". However, standards can help push the acceptance of data mining technology without compromising the speed and direction of new technology development. The key is to avoid trying to standardize what data mining is or what it does, but instead push standards that support the data mining process. First among this is standards for data description. Although data access standards are widely accepted and used, the definition of what data means is typically captured in prose and paper documents. XML is a step in the right direction—it ensures that some metadata are kept with the data. However, work remains in this area.

Another big area where standards can support data mining is in the general architecture of a data mining process. Understanding in advance what must be done at any stage in a data mining effort helps ensure success of that effort. In addition, developing an architecture for the data mining process helps to identify areas within that architecture where standards are needed.

Perhaps the biggest challenge is to standardize the definition of data mining tasks. We need to be able to standardize the results of data mining to support (for example) visualization tools that operate on the results. However, we must not limit our ability to extend the types of analyses that can be performed.

References

[1] B. Thuraisingham, Data Mining: Technologies, Techniques, Tools and Trends. CRC Press, Boca Raton, FL, 1998.

[2] P. Adriaans, D. Zantinge, Data Mining. Addison Wesley, Massachusetts, 1996.

[3] M. Berry, G. Linoff, Data Mining Techniques for Marketing, Sales, and Customer Support. Wiley, New York, 1997.

[4] J. Han, M. Kamber, Data Mining: Concepts and Techniques. Morgan Kaufmann, San Francisco, CA, 2000.

[5] F. Grupe, M. Owrang, Database mining tools. in: B. Thuraisingham (Ed.), The Handbook of Data Management Supplement. Auerbach Publications, New York, 1998, pp. 625–636.

[6] J. Gray, A. Bosworth, A. Layman, H. Pirahesh, Data cube: a relational aggregation operator generalizing group-by, cross-tab, and sub-totals. Proceedings of the 12th International Conference on Data Engineering. IEEE Computer Society Press, Los Alaminos, CA, USA, 1996, pp. 152–159.

[7] Data Mining Group, PMML 1.1—Predictive Model Markup Language (2000). http://www.dmg.org/html/pmml_v1_1.html.

[8] Microsoft, Introduction to OLE DB for Data Mining (Jul. 2000). http://www.microsoft.com/data/oledb/dm.htm.

[9] Java data mining API expert group, jsr 000073 (Aug. 2000). http://java.sun.com/aboutJava/communityprocess/jsr/jsr_073_dmapi.html.

[10] Cross industry standard process for data mining (Dec. 1999). http://www.crisp-dm.org.

[11] Object management group. http://www.omg.org.

Chris Clifton is a Principal Scientist in the Information Technology Center at the MITRE. He has a PhD from Princeton University, and Bachelor's and Master's degrees from the Massachussetts Institute of Technology. Prior to joining MITRE in 1995, he was an Assistant Professor of Computer Science at Northwestern University. His research interests include data mining, database support for text, and heterogeneous databases.

Bhavani Thuraisingham is a chief scientist in data management in MITRE's Information Technology Directorate. She has a MSc from the University of Bristol and a PhD from the University of Wales, both in the United Kingdom. Her research interests are in data mining, electronic commerce, and information security.