Ideal Modelling of Data Mining
Harshil Joshi¹ Keval Parikh²
¹M.E. ²B.E.
Computer Science & Engineering

Abstract— The study of data mining has focused primarily on the mining algorithms and their applications, while relies its foundations on established fields, such as logic statistical analysis, machine learning, database. Motivated by the practical needs of specific types of real world data analysis problems, many mining algorithms are designed and studied. They include association rule mining, classification rule mining, exception and peculiarity rule mining, sequence mining, stream mining, text mining, web mining, and others. A review of data mining literature suggests that there does not exist a well-accepted and non-controversial conceptual framework. A lack of conceptual modelling may jeopardize further development of data mining.

Our discussions are unique and differ from existing studies in several perspectives. First, we treat data mining as a field of study and emphasize the study of the nature, the scope, and philosophical foundations of data mining. We stress on the understanding of data mining as a scientific inquiry, in addition to simply empirical investigations. We pay more attention to the effectiveness of data mining methods, rather than only to the efficiency. Second, we view data mining in a wide context of scientific research, in terms of their goals, processes and methods. Third, we search for a unified and general framework, or at least general principles and guidelines, rather than a family of isolated algorithms. The framework aims at finding answers to what and why questions, as well as how questions. Forth, with the help of conceptual modelling, we attempt to move beyond trial and error, or ad hoc, applications of data mining algorithms, which dominate most current applied studies of data mining.

Key words: Data Mining, Processes, Patterns, Framework for data mining.

I. MODELLING
The needs for modelling and foundations of data mining, it is necessary first to present the current situation of the field and to identify the associated problem. Potential solutions can then be found.

A. Summary of Data Mining Research
A definition of data mining defines it as “The non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns from data” The definition concisely summarizes three views of data mining, the function-oriented, the procedure-oriented, and the application-oriented views.

The function-oriented views focus on the requirements and goals of data mining tasks. Data mining attempts to extract knowledge from data. Goal-driven approaches establish a between data mining research and real world applications.

The theory-oriented views focused on the theoretical studies of data mining, in relation to the other disciplines. Many theories and models of data mining have been nominating analytically investigated and examined. Fields include in the theoretical study include logic, statistics, machine learning, databases, pattern recognition, visualization etc.

The procedure-oriented views cover two parts, namely, data mining algorithms and multiple phases of data mining process. Data mining algorithms deal with specific methods for mining particular types of knowledge. A multi-phase process describes the main steps involved in data mining.

The application-oriented views deal with the utilization of data mining algorithms and techniques in various domains. Applications are in fact the driving market of data mining research.

B. Motivations for Data Mining Modelling
The needs for the studying data mining, let us first quote the following comments from Salthe:

“Functioning as a scientist means functioning within the rules learned during an apprenticeship in which examination of the philosophic foundations of the game plays a characteristically tiny role. One strives to become a member, not to potentially undermine the club by examining its structure from outside. Only when commitment to a way of life is secure is it possible for some to examine its foundations with sympathy.”

C. Foundations of Data Mining
The study of foundations of data mining deals with conceptual modelling of data mining as a field of scientific inquiry. It examines into the nature of data mining and the scope of data mining methods. It treats data mining as an integrated whole and a subject of study, rather than an isolated family of algorithms and applications. It studies the conceptual structures of data mining.

II. DATA MINING AND SCIENTIFIC RESEARCH
Extracting knowledge from data or making sense out of data has been, and is still, a basic endeavour of any scientist. The term data is used here in a very broad sense, covering any format and any content. It is therefore constructive to examine data mining in a wide context of scientific research.

A. Purposes and Goals
 Scientific research is affected by the perceptsions and the purposes of science. The main purpose of science is to describe and predict, to improve or manipulate the world, and to explain our world around us. The results of the research process provide a illustration of an event or a phenomenon. The fact received from research helps us to make predictions about what will happen in the future.

B. Common Processes
 Research is a highly complex and precise human activity. Some lessons and general principles can be learnt from the experience of scientists. There are some basic principles and techniques that are used in scientific investigations.
“The particular observations made vary from one discipline to another because each discipline is interested in observing and understanding different phenomena. But the basic processes and the systematic way of studying problems are common elements of science, regardless of each discipline’s particular subject matter. It is the process and not the content that distinguishes science from other ways of knowing, and it is the content – the particular phenomena and fact of interest – that distinguishes one scientific discipline from another.”

C. Data Mining Process

The commonly used data mining process is similar to the research process. The following is a summary of a typical data mining process.

- Data pre-processing phase: to select and clean working data.
- Data transformation phase: to change the working data into the required form.
- Pattern discovery and evaluation phase: to apply algorithms to identify knowledge embedded in data, and to evaluate the discovered knowledge.
- Explanation construction and evaluation phase: to construct plausible explanations for discovered knowledge, and to evaluate different explanations.
- Pattern presentation: to present the extracted knowledge and explanations.

D. Suggestion

The examination of data mining in a wider context of scientific research lends itself to a bi-directional interaction of data mining and scientific research.

From the viewpoint of scientific research, it is possible to have an in-depth understanding of data mining. The four dimensions of growth of data mining research is indeed closely related to some aspects in the evolution of science in general, namely, the application of old ideas to new data or new types of data, the exploration of new scientific ideas, and the emergence of new branches of science. The experiences and results from the studies of research methods can be applied to data mining. Many different ways, methods, tools, and ideas have been experiment by scientists again and again in many different domains. This opens new doors to data mining research. In carrying out data mining tasks, one can borrow or adopt scientific research methods and ideas that have been used either explicitly or implicitly by scientists.

III. MULTI-LEVEL MODELLING OF DATA MINING

The needs for a ideal modelling of data mining, as well as its benefits, have been argued early. Then discuss three levels for the understanding of data mining.

A. Understanding of An Information System

Understanding of an information processing system involves explanations at various levels or layers. A special feature of this hierarchical analysis is the consideration of computational issues, such as representation and process, and implementation. As a basic component of information processing systems, a process can be understood at three levels.

The most abstract level deals with what the process does and why. One builds a theory that explains internal working principles of the process, and defines the operations by specifying constraints that must be satisfied by the process.

The second level deals with the realization of the process in an abstract way. One needs to choose a representation for the input and for the expected output of the process, and to specify an algorithm for the transformation from input to output. The choices of representation and algorithm are closely tied together. There usually exist many alternative representations. For a given representation, there are also many possible algorithms. A representation and an algorithm should be chosen so that advantages of the representation are fully exploited by the algorithm and, at the same time, the disadvantages of the representation are avoided.

The third level deals with the physical realization of the process. The devices that physically realize a process may not be unique. The advances in technologies imply that the same process may be implemented again with the invention of new physical devices.

The three-level understanding of a process can be generalized to an information processing system. While the first level answers questions of what and why, the other two levels answer two types of questions of how.

Investigation at the computational theory level is independent of representations and investigations at representation and algorithm level is independent of physical devices. The levels are ordered and interpreted as levels of abstraction. There are in fact two implementation levels. The representation and algorithm level may be viewed as the logical implementation level, and the hardware implementation level as the physical implementation level.

B. A Three-layered Framework of Data Mining

A three-layered conceptual framework consists of the philosophy layer, the technique layer, and the application layer. The layered framework represents the understanding, discovery, and utilization of knowledge. The framework basically adds an application level to Sowa’s [Sow84] information processing model, and at the same time taking consideration of Marr’s [Mar82] multi-level understanding of information processing systems.

The philosophy layer: The philosophy layer investigates the basic issues of knowledge. One attempts to answer the fundamental question, namely, what is knowledge? There are many related issues, such as the representation of knowledge, the expression and communication of knowledge in languages, the storage and processing of knowledge in mind, the relationship between knowledge in abstract, in the mind and in the external real world, and the classification and organization of knowledge [Sow84]. Philosophical study of data mining serves as a precursor to technology and applications. It generates knowledge and the understanding of our world, with or without establishing operational boundaries of knowledge.

The technique layer: The technique layer is the study of knowledge discovery methods and their implementations in machine. Two levels can be further formed, i.e., the logical implementation and physical implementation. One attempts to answer the fundamental question, how to discover knowledge? Logical analysis and mathematical modeling are more relevant at the logical
implementation level. One is interested in searching for new
calgorithms and improving existing algorithms. The physical
implementation expresses knowledge discovery methods
through programming languages, which involves the coding,
storage and retrieval of information. The main streams of
research in machine learning, data mining, and knowledge
discovery have concentrated on the technique layer.

The application layer: The ultimate goal of
knowledge discovery is to effectively use the discovered
knowledge. At this layer, one needs to answer the question,
how to utilize the discovered knowledge? This layer
attempts to make explicit and precise the intuitive notions of
usefulness and meaningfulness of discovered knowledge,
based on domain specific background knowledge.

The three layers are relatively independent and
loosely connected. The inner or lower layers establish a
foundation for the outer or upper layers, while the outer
layers may raise questions for the inner layers. This explicit
division, although may be artificial, enables us to see the
basic issues of data mining more clearly. Useful and
insightful remarks can be made regarding the three-level
framework. The results from philosophy level will provide
guideline and set the stage for the algorithm and application
levels. Philosophical study does not depend on the
availability of specific techniques. More specifically, the
existence of a particular type of knowledge, as well as its
usefulness, is not determined by the existence of a mining
algorithm for such knowledge. Knowledge is a matter of
existence. Data mining algorithms merely reveal the
knowledge embedded in the data. The technique level study
is not constrained by a particular application. The existence
of an algorithm does not necessarily imply that the
discovered knowledge is meaningful and useful.

Fig. 1: The Three Layer Ideal Framework of Data Mining

C. Suggestion

It may be debatable regarding the appropriateness of the
three-level framework, such as the number of levels, the
division between levels, the interaction of different levels,
and the issues at each level. The usefulness of the
framework is evident. It represents a three-level description
of data mining characterized by three fundamental issues,
namely, the understanding, discovery, and utilization of
knowledge. Each of them is indispensable, and jointly they
present a framework within which a multi-disciplinary study
of data mining is possible.

A significant implication of the framework lies on
its division of the understanding of a complex problem into
different levels, which leads to a division of basic issues of
data mining into levels. It also gives the proper context in
which a particular type of questions can be answered. One
can clearly see the scope and limitations of various types of
data mining research.

It may be argued that the similar testing method
can be used to evaluate the usefulness of the ideal modeling
of data mining, or in particular the three-level framework. In
fact, initial classroom instruction does show that students
gasp fast the basic issues of data mining with the three-
level framework. Moreover, students can better formulate a
research problem, in terms of its scope, basic issues, and
potential solutions, with the help of multi-level conceptual
description.

IV. CONCLUSIONS

In this paper, we elaborate on several issues in the ideal
modeling of data mining. Since detailed description of data
mining, in terms of theory, techniques, and algorithms, can
be found easily in the literature, we only touch it briefly and
superficially. Instead, we concentrate on a high-level ideal
modeling of data mining. An exposition of data mining is
made in a wide context of scientific research. A three-
layered conceptual framework is used for the understanding
of data mining as a field of study. Our aim is to stimulate
more researchers to look into conceptual modelling, a less
studied, but extremely crucial, area of data mining research.

The content of the paper is both old and new. It is
old in the sense that many results are drawn from other
fields, including cognitive science, philosophy of science,
research methods, education, hierarchy theory, granular
computing, and information processing systems. It is also
new in the sense that these results are associated with new,
domain specific meaning and are applied to data mining. In
particular, we review the relevant results, interpret them, and
apply them to data mining.

The content of the chapter is both old and new. It is
old in the sense that many results are drawn from other
fields, including cognitive science, philosophy of science,
research methods, education, hierarchy theory, granular
computing, and information processing systems. It is also
new in the sense that these results are associated with new,
domain specific meaning and are applied to data mining. In
particular, we review the relevant results, interpret them, and
apply them to data mining.

REFERENCES

[1] [Man97] Mannila, H. Methods and problems in
data mining. Proceedings of International
Conference on Database Theory,
Investigation into Human Representation and
Processing of Visual Information, W.H. Freeman
[3] [ML02] Matloff, N. and Lin, T.Y. Toward
statistical foundation for data mining, Proceedings
of IEEE ICDM’02 Workshop on Foundation of
Data Mining and Knowledge Discovery, 125-130,
2002.
[4] [MM99] Martella, R.C., Nelson, R. and
Marchand-Martella, N.E. Research Methods:
Learning to Become a Critical Research Consumer,
Allyn and Bacon, Boston, 1999.
[5] [Pat73] Pattee, H.H. Unsolved problems and
potential applications of hierarchy theory, in:


