Formatted: Width: 8.27", Height: 11.69"

INTRODUCTION TO KNOWLEDGE DISCOVERY AND DATA MINING

HO Tu Bao
Institute of Information Technology
National Center for Natural Science and Technology

Contents

Preface

Knowledge Discovery and Data mining (KDD) emerged as a rapidly growing interdisciplinary field that merges together databases, statistics, machine learning and related areas in order to extract valuable information and knowledge in large volumes of data.

With the rapid computerization in the past two decades, almost all organizations have collected huge amounts of data in their databases. These organizations need to understand their data and/or to discover useful knowledge as patterns and/or models from their data.

This course aims at providing fundamental techniques of KDD as well as issues in practical use of KDD tools. It will show how to achieve success in understanding and exploiting large databases by: uncovering valuable information hidden in data; learn what data has real meaning and what data simply takes up space; examining which data methods and tools are most effective for the practical needs; and how to analyze and evaluate obtained results.

The course is designed for the target audience such as specialists, trainers and IT users. It does not assume any special knowledge as background. Understanding of computer use, databases and statistics will be helpful.

The main KDD resource can be found from http://www.kdnutggets.com. The selected books and papers used to design this course are followings: Chapter 1 is with material from [7] and [5], Chapter 2 is with [6], [8] and [14], Chapter 3 is with [11] and [12], Chapters 4 and 5 are with [4], Chapter 6 is with [3], and Chapter 7 is with [13].

Chapter 1

Overview of knowledge discovery and data mining

1.1 What is Knowledge Discovery and Data Mining?

Just as electrons and waves became the substance of classical electrical engineering, we see data, information, and knowledge as being the focus of a new field of research and application—knowledge discovery and data mining (KDD) —that we will study in this course.

In general, we often see *data* as a string of bits, or numbers and symbols, or "objects" which are meaningful when sent to a program in a given format (but still un-interpreted). We use bits to measure *information*, and see it as data stripped of redundancy, and reduced to the minimum necessary to make the binary decisions that essentially characterize the data (interpreted data). We can see *knowledge* as integrated information, including facts and their relations, which have been perceived, discovered, or learned as our "mental pictures". In other words, knowledge can be considered data at a high level of abstraction and generalization.

Knowledge discovery and data mining (KDD)—the rapidly growing interdisciplinary field which merges together database management, statistics, machine learning and related areas—aims at extracting useful knowledge from large collections of data.

There is a difference in understanding the terms "knowledge discovery" and "data mining" between people from different areas contributing to this new field. In this chapter we adopt the following definition of these terms [7]:

Knowledge discovery in databases is the process of identifying valid, novel, potentially useful, and ultimately understandable patterns/models in data**.** *Data mining* is a step in the knowledge discovery process consisting of particular data mining algorithms that, under some acceptable computational efficiency limitations, finds patterns or models in data.

In other words, the goal of knowledge discovery and data mining is to find interesting patterns and/or models that exist in databases but are hidden among the volumes of data.

Category	Type of Attributes	# Attributes
Present History	Numerical and Categorical	07
Physical Examination	Numerical and Categorical	08
Laboratory Examination	Numerical	11
Diagnosis	Categorical	02
Therapy	Categorical	02
Clinical Course	Categorical	04
Final Status	Categorical	02
Risk Factor	Categorical	02
Total		38

Table 1.1: Attributes in the meningitis database

Throughout this chapter we will illustrate the different notions with a real-world database on meningitis collected at the Medical Research Institute, Tokyo Medical and Dental University from 1979 to 1993. This database contains data of patients who suffered from meningitis and who were admitted to the department of emergency and neurology in several hospitals. Table 1.1 presents attributes used in this database. Below are two data records of patients in this database that have mixed numerical and categorical data, as well as missing values (denoted by "?"):

10, M, ABSCESS, BACTERIA, 0, 10, 10, 0, 0, 0, SUBACUTE, 37,2, 1, 0, 15, -, -6000, 2, 0, abnormal, abnormal, -, 2852, 2148, 712, 97, 49, F, -, multiple, ?, 2137, negative, n, n, n

12, M, BACTERIA, VIRUS, 0, 5, 5, 0, 0, 0, ACUTE, 38.5, 2,1, 0, 15, -, -, 10700, 4, 0, normal, abnormal, +, 1080, 680, 400, 71, 59, F, -, ABPC+CZX, ?, 70, negative, n, n, n

A pattern discovered from this database in the language of IF-THEN rules is given below where the pattern's quality is measured by the confidence (87.5%):

IF Poly-nuclear cell count in CFS <= 220
and Risk factor = n
and Loss of consciousness = positive
and When nausea starts > 15
THEN Prediction = Virus [Confidence = 87.5%]

Concerning the above definition of knowledge discovery, the 'degree of interest' is characterized by several criteria: *Evidence* indicates the significance of a finding measured by a statistical criterion. *Redundancy* amounts to the similarity of a finding with respect to other findings and measures to what degree a finding follows from another one. *Usefulness* relates a finding to the goal of the users. *Novelty* includes the deviation from prior knowledge of the user or system. *Simplicity* refers to the syntactical complexity of the presentation of a finding, and generality is determined. Let us examine these terms in more detail [7].

- *Data* comprises a set of facts F (e.g., cases in a database).
- *Pattern* is an expression E in some language L describing a subset F_E of the data F (or a *model* applicable to that subset). The term pattern goes beyond its traditional sense to include *models* or *structure* in data (relations between facts), e.g., "If (Poly-nuclear cell

count in CFS <= 220) and (Risk factor = n) and (Loss of consciousness = positive) and (When nausea starts > 15) Then (Prediction = Virus)".

- *Process:* Usually in KDD process is a multi-step process, which involves data preparation, search for patterns, knowledge evaluation, and refinement involving iteration after modification. The process is assumed to be non-trivial, that is, to have some degree of search autonomy.
- *Validity:* The discovered patterns should be valid on new data with some degree of certainty. A measure of certainty is a function C mapping expressions in L to a partially or totally ordered measurement space M_C. An expression E in L about a subset $F_E \subset F$ can be assigned a certainty measure $c = C(E, F)$.
- *Novel:* The patterns are novel (at least to the system). Novelty can be measured with respect to changes in data (by comparing current values to previous or expected values) or knowledge (how a new finding is related to old ones). In general, we assume this can be measured by a function $N(E, F)$, which can be a Boolean function or a measure of degree of novelty or unexpectedness.
- *Potentially Useful:* The patterns should potentially lead to some useful actions, as measured by some utility function. Such a function U maps expressions in L to a partially or totally ordered measure space M_U: hence, u = $U(E, F)$.
- *Ultimately Understandable:* A goal of KDD is to make patterns understandable to humans in order to facilitate a better understanding of the underlying data. While this is difficult to measure precisely, one frequent substitute is the *simplicity measure*. Several measures of simplicity exist, and they range from the purely syntactic (e.g., the size of a pattern in bits) to the semantic (e.g., easy for humans to comprehend in some setting). We assume this is measured, if possible, by a function S mapping expressions E in L to a partially or totally ordered measure space M_S: hence, $s = S(E,F)$.

An important notion, called *interestingness*, is usually taken as an overall measure of pattern value, combining validity, novelty, usefulness, and simplicity. Interestingness functions can be explicitly defined or can be manifested implicitly through an ordering placed by the KDD system on the discovered patterns or models. Some KDD systems have an explicit interestingness function $i = I(E, F, C, N, U, S)$ which maps expressions in L to a measure space M_I. Given the notions listed above, we may state our definition of knowledge as viewed from the narrow perspective of KDD as used in this book. This is by no means an attempt to define "knowledge" in the philosophical or even the popular view. The purpose of this definition is to specify what an algorithm used in a KDD process may consider knowledge.

> A pattern $E \in L$ is called *knowledge* if for some user-specified threshold $i \in M_I$, $I(E, F, C, N, U, S) > i$

Note that this definition of knowledge is by no means absolute. As a matter of fact, it is purely user-oriented, and determined by whatever functions and thresholds the user chooses. For example, one instantiation of this definition is to select some thresholds $c \in M_C$, $s \in M_S$, and $u \in M_U$, and calling a pattern E knowledge if and only if

> $C(E, F) > c$ and $S(E, F) > s$ and $U(S, F) > u$

By appropriate settings of thresholds, one can emphasize accurate predictors or useful (by some cost measure) patterns over others. Clearly, there is an infinite space of how the mapping *I* can be defined. Such decisions are left to the user and the specifics of the domain.

1.2 The Process of Knowledge Discovery

The process of knowledge discovery inherently consists of several steps as shown in Figure 1.1.

The first step is *to understand the application domain and to formulate the problem*. This step is clearly a prerequisite for extracting useful knowledge and for choosing appropriate data mining methods in the third step according to the application target and the nature of data.

The second step is *to collect and preprocess the data*, including the selection of the data sources, the removal of noise or outliers, the treatment of missing data, the transformation (discretization if necessary) and reduction of data, etc. This step usually takes the most time needed for the whole KDD process.

The third step is data mining that extracts patterns and/or models hidden in data. A model can be viewed "a global representation of a structure that summarizes the systematic component underlying the data or that describes how the data may have arisen". In contrast, "a pattern is a local structure, perhaps relating to just a handful of variables and a few cases". The major classes of *data mining methods are predictive modeling such as classification* and *regression; segmentation (clustering); dependency modeling such as* graphical models or density estimation; *summarization* such as finding the relations between fields, associations, visualization; and *change and deviation detection/modeling* in data and knowledge.

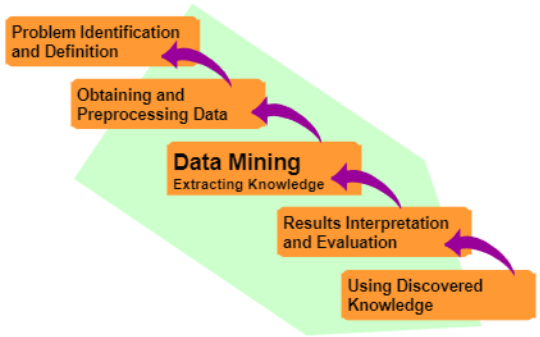

Figure 1.1: the KDD process

The fourth step is *to interpret (post-process) discovered knowledge*, especially the interpretation in terms of description and prediction—the two primary goals of discovery systems in practice. Experiments show that discovered patterns or models from data are not always of interest or direct use, and the KDD process is necessarily iterative with the judgment of discovered knowledge. One standard way to evaluate induced rules is to divide