6

Data Mining Trends and Knowledge Discovery

Objectives:

- The rapidly emerging field of knowledge discovery in databases (KDD) has grown significantly in the past few years.
- Modern database technology enables economical storage of these large streams of data.
- We do not yet have the technology to help us analyze, understand, or even visualize this stored data.
- These challenges are central to KDD and need urgent attention.
- KDD is an interdisciplinary field that brings together researchers and parishioners from a wide variety of fields.
- The major related fields include statistics, machine learning, artificial intelligence, and reasoning with uncertainty, databases, knowledge acquisition, pattern recognition, information retrieval, visualization, intelligent agents for distributed and multimedia environments, digital libraries, and management information systems.

Abstract. The rapidly emerging field of knowledge discovery in databases (KDD) has grown significantly in the past few years. A mix of daunting practical needs and strong research interest drives this growth. The technology for computing and storage has enabled people to collect and store information from a wide range of sources at rates that were considered unimaginable only a few years ago. Although modern database technology enables economical storage of these large streams of data, we do not yet have the technology to help us analyze, understand, or even visualize this stored data.

Examples of this phenomenon around in a wide spectrum of fields, finance, banking, retail sales, manufacturing, monitoring and diagnosis (be it of humans or machines), health care, marketing, and science data acquisition, among others. In science, modern instruments can easily measure and collect terabytes (10^{12} bytes) of data. For example, NASA's Earth Observing System is expected to return data at rates of several gigabytes per hour by the end of this century. Quite appropriately, the problem of how to put the torrent of data to use in analysis is often called "drinking

S. Sumathi and S.N. Sivanandam: *Data Mining Trends and Knowledge Discovery*, Studies in Computational Intelligence (SCI) **29**, 185–194 (2006)
www.springerlink.com

from the fire hose." What we mean by analysis is not well defined because it is highly context and goal dependent. However, it typically transcends by far anything achievable via simple queries, simple string matching, or mechanisms for displaying the data. The data mining trends and knowledge discovery are the subjects of this section.

Prolific sources of data are not restricted to esoteric endeavors involving spacecraft or sophisticated scientific instruments. Imagine a database receiving transactions from common daily activities such as supermarket or department store checkout-register sales, or credit card charges. On the other hand, think of the information reaching one's home television set as a stream of signals that, to be properly managed, need to be cataloged and indexed, and perhaps searched for interesting content at a higher level-channels, programs, genre, or mood, for example. The explosion in the number of resources available on the global computer network – the World Wide Web – is another challenge for indexing and searching through a continually changing and growing "database."

6.1 Getting a Handle on the Problem

Why are today's database and automated match and retrieval technologies not adequate for addressing the analysis needs? The answer lies in the fact that the patterns to be searched for and the models to be extracted are typically subtle and require significant specific domain knowledge. For example, consider a credit card company wishing to analyze its recent transactions to detect fraudulent use or to use the individual history of customers to decide online whether an incoming new charge is likely to be from an unauthorized user. This is clearly not an easy classification problem to solve.

One can imagine constructing a set of selection filters that trigger a set of queries to check if a particular customer has made similar purchases in the past, or if the amount or the purchase location is unusual, for example. However, such a mechanism must account for changing tastes, shifting trends, and perhaps travel or change of residence. Such a problem is inherently probabilistic and would require reasoning with uncertainty scheme to properly handle the trade-off between disallowing a charge and risking a false alarm, which might result in the loss of a sale (or even a customer).

In the past, we could rely on human analyst to perform the necessary analysis. Essentially, this meant transforming the problem into one of the simply retrieving data, displaying it to an analyst, and relaying an expert knowledge to reach a decision. However, with large databases, a simple query can easily return hundreds or thousands (or even more) matches. Presenting the data, letting the analyst digests it, and enabling a quick decision becomes infeasible. Data visualization techniques can significantly assist this process, but ultimately the reliance on the human in the loop becomes a major bottleneck. (Visualization works only for small sets and a small number of variables. Hence, the problem becomes one of finding the appropriate transformations and reductions – typically just as difficult as the original problem.)

Finally, there are situations where one would like to search for patterns that humans are not well suited to find. Typically, this involves statistical modeling followed by "outlier" detection, pattern recognition over large data sets, classification, or clustering. (Outliers are data points that do not fit within a hypothesis's probabilistic mode and hence are likely the result of interference from another process.) Most database management systems (DBMSs) do not allow the type of access and data manipulation that these tasks require; there are also serious computational and theoretical problems attached to performing data modeling in high dimensional spaces with large amount of data.

These challenges are central to KDD and need urgent attention. Without heavily emphasizing KDD development and research, we run the risk of forfeiting the value of most of the data that we collect and store. We would eventually drown in an ocean of massive (but valuable) data sets that are rendered useless because we cannot distil the essence from the bulk. To draw the data mining analogy: the precious nuggets of knowledge need to be extracted and the massive raw material needs to be managed appropriately (and preferably recycled effectively). Before proceeding further, let us define what we mean by KDD and data mining.

6.2 KDD and Data Mining: Background

We use the term KDD to denote the overall process of extracting high-level knowledge from low-level data. Others might use the term's *data mining and KDD* interchangeably. The multitude of names used for KDD includes data or information harvesting, data archeology, functional dependency analysis, knowledge extraction, and data pattern analysis. Historically, in statistics especially, the term *data mining or fishing* refers to sloppy exploratory data analysis with no a priori hypotheses to verify.

A simple definition. A simple high-level definition of KDD is as follows:

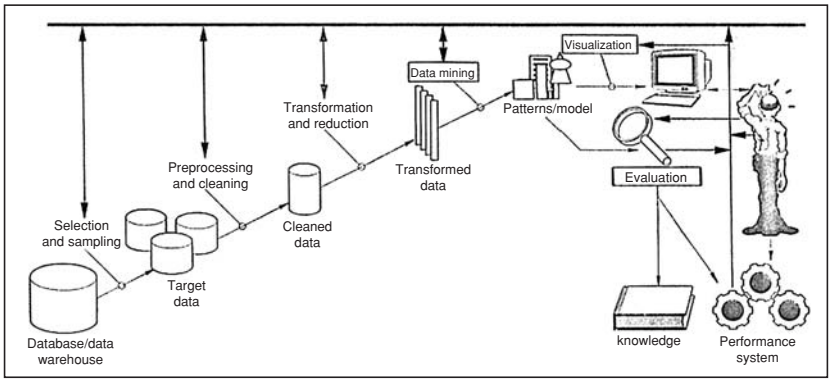
Knowledge discovery in databases in the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. Given the scope of this short section, we will not go into the definitions of each term in this high-level statement. Note, however, that the term knowledge is (and has a long history of being) difficult to define in the abstract. We adopt the view that knowledge is in the eye of the beholder, so one person's knowledge could easily be another's junk. We define knowledge in domain-dependent terms relating strongly to measures of utility, validity, novelty, and understandability. The term *patterns* in this definition loosely denote either models or patterns. In general, it designates some abstract representation of a *subset of the data*. A significant term in the definition is *process*, which indicates that knowledge discovery often involves experimentation, iteration, user interaction, and many design decision and customizations. Extracting knowledge from data can easily turn into a complicated and sometimes

arduous process. But the payoffs for success can be dramatic and rewarding, sometimes enabling people and organizations to achieve tasks that would not otherwise be possible.

We adopt the convention that data mining refers to the act of extracting patterns or models from data (be it automated or human-assisted). However, many steps precede the data-mining step: retrieving the data from a large warehouse (or some other source); selecting the appropriate subset to work with; deciding on the appropriate sampling strategy; cleaning the data and dealing with missing fields; and applying the appropriate transformation, dimensionality reduction, and projections. The data-mining step then fits models to, or extracts patterns from, the preprocessed data. However, to decide whether this extracted information does represent knowledge, one needs to evaluate this information, perhaps visualize it, and finally consolidate it with existing (and possibly contradictory) knowledge. Obviously, these steps are all on the critical path from data to knowledge. Furthermore, any one-step can result in change in the preceding or succeeding steps, often requiring starting from scratch with new choices and settings. Hence, in the definition we adopt, data mining is just a step in the overall KDD process.

Figure 6.1 outlines the KDD process; it is perhaps deceptive because it gives the impression that the steps are well defined. In fact, the interactions between the choices of techniques used in the various steps, the parameters used for those techniques, and the choice of problem representation are extremely complex. Small changes in one part can dramatically affect the rest, and consequently can make the difference between success and failure of a KDD enterprise.

The KDD Process: Table 6.1 expands on the steps outlined in Fig. 2.1. A few items in Table 2.1 warrant further comment. Step 4 is critical and can be quite involved. Indeed, in many cases, some sophisticated searching and cataloging problem must be solved before the actual subsequent analysis is performed. This transformation could require solving a significant problem in its own



I .An overview of the KDD process.2 (For simplicity, the illustration omils arrows indicuting the multilude of potential loops and iterations.)

Fig. 6.1. An overview of the KDD process. (For simplicity, the illustration omits arrows indicating the multitude of potential loops and iterations.)

Table 6.1. Steps involved in the KDD Process

Step	Explanation
1. Developing an understanding of the application domain, the relevant prior knowledge, and the goals of the end user	With today's technology, this step requires a fair bit of reliance on the user/analyst. Factors to consider include: • What are the bottlenecks in the domain? What is worth automating and what is best left for processing by humans. • What are the goals? What performance criteria are important? • Will the final product of the process be used for classification, visualization, exploration, summarization, or something else? • Is understandability an issue? What is the trade-off between simplicity and accuracy of the extracted knowledge? Is a black box model appropriate for the performance element of the system?
2. Creating a target data set, selecting a data set, or focusing on a subset of variables or data samples, on which discovery is to be performed.	This involves considerations of homogeneity of data, any dynamics and change over time, sampling strategy (such as uniform random versus stratified), sufficiency of sample, degrees of freedom, and so forth.
3. Data cleaning and preprocessing.	Involved here are basic operations such as the removal of noise or "outliers," if appropriate; collecting the necessary information to model or accounting for noise; deciding on strategies for handling missing data fields; accounting for time sequence information, known changes, and appropriate normalization; and so forth.
4. Data reduction and transformation	This involves finding useful features to represent the data, depending on the goal of the task' using dimensionally reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data; and projecting the data onto spaces in which a solution is likely to be easier to find.
5. Choosing the data mining task	This involves deciding whether the goal of the KDD process is classification, regression, clustering, summarization, dependency modeling, or change and deviation detection.
6. Choosing data mining algorithms	Here we select the methods to be used fro searching for patterns in or fitting models to the data. The choice of which models and parameters mat be appropriate is often critical. In addition, the data mining method must be compatible with the goals; the end user may be more interested in understanding the model than its predictive capabilities.

7. Data mining	This involves searching for patterns of interest in a particular form or a set of such representations: classification rules or trees, regression, clustering, and so on. The user can significantly aid the data mining method by correctly performing the preceding steps.
8. Evaluating output of step 7.	Here we decide what is to be deemed knowledge, which can be fairly a difficult task. Achieving acceptable results may involve using options (possibly in combination): • Defining an automated scheme using measures of "interestingness" and others to filter knowledge from other outputs. Such measures might be statistical measures, goodness of fir, or simplicity, among others. • Relying on visualization techniques to help the analyst decide the utility of extracted knowledge or reach conclusions about the underlying data/phenomena. • Relying entirely on the user to sift through derived patterns in hope of coming across items of interest. The outcome of this step might result in changes to any of the preceding steps and a restart of the entire process.
9. Constructing discovered knowledge: incorporating this knowledge into the performance system, or simply documenting it and reporting it to users.	This also includes checking for and resolving potential conflicts with previously believed (or extracted) knowledge.

right. In classical pattern-recognition work, this is called *feature extraction* problem. In general, its solution requires a good bit of domain knowledge and strong intuition about the problem. It typically makes the difference between success and failure of the data mining (step 7).

However not to discourage the reader, feature definition and extraction in many applications is not terribly difficult (especially for a motivated domain expert who is involved in the process). Humans find it easier to define features than to solve the data-mining problem. For instant, an expert can observe a set of low-level variables and reach initiative decision. For example, the low-level may consist of a stream of readings of voltages, currents, capacitances, loads, and so forth from a power plant; a set of pixels in multispectral images from a remote sensing instrument; or a set of transactions for a given group of bank accounts. However, the expert might not at all be capable of elucidating the reasons for reaching some decision about the state of the system being observed. This is typically a reasonable setting to use classification (supervised learning) techniques to derive classifiers from examples (the data)

directly. Hence, the expert presents the system with training data consisting of classified examples.

For a nonexpert (especially a machine), using the raw observed data to classify events is likely to result in failure: knowledge of time sequence, of properties, of instruments, of noise, of what is an important quantity, and so forth is simply a prerequisite. Experts can be asked to define features from the lower-level data. In effect, feature definition by experts lets them decompose the problem into small parts and encode significant prior knowledge implicitly in their choice of representation. This can easily result in a large number of features. Typically, the expert would not know how to use these features to solve the classification (discrimination) or modeling problem. Data mining techniques provide a way to get the solution in this feature space.

In step 8, the reliance on visualization is simply a work-around of the fact that we find it difficult to emulate human intuition and decision-making on a machine. The idea is to transform the derived knowledge into a format that is easy for humans to digest (such as images or graphs) and then rely on the speed and capability of the highly evolved human visual system to spot what is interesting. Of course, this only works in low dimensional spaces, so the choice of what to show the user to facilitate the discovery is still critical and typically not an easy problem to circumvent.

6.3 Related Fields

By definition, KDD is an interdisciplinary field that brings together researchers and parishioners from a wide variety of fields. The major related fields include statistics, machine learning, artificial intelligence and reasoning with uncertainty, databases, knowledge acquisition, pattern recognition, information retrieval, visualization, intelligent agents for distributed and multimedia environments, digital libraries, and management information systems. The remainder of this section briefly outlines how some of these relate to the various parts of the KDD process. We focus on the main fields and hope to clarify to the reader, the role of each of the fields and how they fit together naturally when unified under the goals and applications of the overall KDD process. A detailed or comprehensive coverage of how they relate to the KDD process would be too lengthy and not very useful because ultimately one can find relations to every step form each of the fields.

Statistics. Statistics plays an important role primarily in data selection and sampling, data mining, and evaluation of extracted knowledge steps. Historically, most statistics work has focused on evaluation of model fit to data land on hypothesis testing. These are clearly relevant to evaluation the results of data mining to filter the good from the bad, as well as within the data mining step itself in searching for, parametrizing, and fitting models to data. On the front end, sampling schemes play an important role in selecting which

data to feed to the data-mining step. For the data-cleaning step, statistics offers techniques for detecting "outliers," smoothing data when necessary, and estimating noise parameters. To a lesser degree, estimation techniques for dealing with missing data are also available. Finally, for exploratory data analysis, some techniques in clustering and design of experiments come into play. However, the focus of research has dealt primarily with small data sets and addressing small sample problems.

On the limitations front, work in statistics has focused mostly on theoretical aspects of techniques and models. Thus, most work focuses on linear models, additive Gaussian noise models, parameter estimation, and parametric methods for a restricted class of models. Search has received little emphasis, with emphasis on closed-form analytical solutions whenever possible. While the latter is very desirable both computationally and theoretically, in many practical situations a user might not have the necessary background statistics knowledge (which can often be substantial) to appropriately use and apply the methods. Furthermore, the typical require an a priori model and significant domain knowledge of the data as well as of the underlying mathematics for proper use and interpretation. In addition, issues having to do with interfaces to databases, dealing with massive data sets, and techniques for efficient data management have only recently begun to receive attention in statistics. John Elder and Darryl Pregibon in 1996 provide an excellent exposition of statistical perspective on KDD.

Pattern Recognition, Machine Learning, and Artificial Intelligence. In pattern recognition, work has historically focused on practical techniques with an appropriate mix of rigor and formalism. The major applicable techniques fall under the category of classification learning and clustering. There are several texts on the topic; Pattern Classification and Scene Analysis provide a good start. Hence, most pattern-recognition work contributes to the data-mining step in the process. Significant work in dimensionality reduction, transformations, and projections has relevance to the corresponding step in the KDD process.

Within the data-mining step, pattern-recognition contributions are distinguished from statistics by their emphasis on computational algorithms, more sophisticated data structures, and more search, both parametric and nonparametric. Given its strong ties to image analysis and problems in 2D signal processing, work in pattern recognition did not emphasize algorithms for dealing with symbolic and categorical data. Classification techniques applied to categorical data typically take the approach of mapping the data to a metric space (such as nearest-neighbor norms). Such a mapping is often not easy to formulate meaningfully: Is the distance between the values "square" and "circle" for the variable *shape* greater than the distance between "male" and "female" for the variable *sex*?

Techniques originating in AI have focused almost exclusively on dealing with data at the symbolic (categorical) level, with little attention paid to con-

tinuous variables. In machine learning and case-based reasoning, algorithms for classification and clustering have focused heavily on heuristic search and nonparametric models. Emphasis on mathematical rigor and analysis of results has not been as strong as in statistics or pattern recognition, with the exception of computational learning theory, which has focused on formal general worst-case bounds for a wide class of representations (a good starting point here is Computational Learning Theory). Machine learning work contributes mainly to the data – mining step of the process, with some contributions in the area of representation and selection of variables through significant search. In addition, the machine discovery community has focused on techniques for discovering structure in data as well as empirical laws to describe observations as in scientific discovery of laws.

AI techniques for reasoning, especially techniques form the Uncertainty in AI community and graphical models for Bayesian modeling and reasoning provide a powerful alternative to classical density estimation in statistics. These techniques have the advantage of allowing prior knowledge about the domain and data to be included in a relatively easy and natural framework. Other areas of AI, including knowledge-acquisition techniques, knowledge representation, and search, are relevant to the various steps in the process, including data mining, data transformation, data selection, and preprocessing.

Databases and Data Warehouses. The relevance of the field of databases to KDD is obvious from the name. Databases provide the necessary infrastructure to store, access, and manipulate the raw data. With parallel and distributed database management systems, they provide the essential layers to insulate the analysis for the extensive details of how the data is stored and retrieved. We focus here only on the aspects of database research relevant to the data-mining step. A strongly related term is on-line analytical processing, which mainly concerns providing new ways of manipulating and analyzing data using multidimensional methods. This has been primarily driven by the need to overcome limitations posed by SQL and relational DBMS schemes for storing and accessing data. The efficiencies achieved via relational structure and normalization can pose significant challenges to algorithms that require special access to the data; in data mining, one would need to collect statistics and counts based on various partitioning of the data, which would require excessive joins and new tables to be generated. Supporting operations form the data-mining perspective is an emerging research area in the database community. In the data-mining step itself, new approaches for functional dependency analysis and efficient methods for finding association rules directly from databases have emerged and are starting to appear as products. In addition, classical database techniques for query optimization and new object-oriented databases make the task of searching for patterns in databases much more tenable.

6.4 Summary

An emerging area in databases is data warehousing, which is concerned with schemes and methods of integrating legacy databases, on-line transaction databases, and various no homogeneous RDBMSs so that they can be accessed in a uniform and easily managed framework. Data warehousing primarily involves storage, data selection, data cleaning, and infrastructure for updating databases once new knowledge or representations are developed.

6.5 Review Questions

1. Define the term Knowledge Discovery Data mining Process (KDD)
2. Describe KDD process and the steps involved in the process with a neat sketches
3. Draw the block diagram of the overall KDD Process