9

Data Mining & KDD

Objectives:

- In this section, the ideas of Data Mining and KDD process are presented.
- The objective of data mining is to convert data and information into profitable knowledge for actionable strategic decisions.
- DMI: Data mining is traditional data analysis methodology updated with the most advanced analysis techniques applied to discovering previously unknown patterns.
- DMII: Data Mining is the activity of extracting hidden information (pattern and relationships) from large databases automatically: that is, without benefit of human intervention or initiative in the knowledge discovery process.
- DMIII: Data Mining is the step in the process of knowledge discovery in databases, that inputs predominantly cleaned, transformed data, searches the data using algorithms, and outputs patterns and relationships to the interpretation/evaluation step of the KDD process.
- KDDI: Knowledge discovery in databases in a process that requires hypothesis or model formulation, hypothesis or model testing, and derivatively all the data, techniques, and sub-processes necessary to bring hypothesis or model testing to a successful conclusion.
- KDDII: True knowledge discovery in databases is the process of automated data mining applied without benefit of human intervention or initiative.
- KDDIII: *"Knowledge discovery in databases* is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data."
- The historical roots of data mining come primarily from two different directions: from statistics and from artificial intelligence.
- The statistical culture in data mining emphasizes the role of predictive modeling (PM). The artificial intelligence culture emphasizes the role of knowledge discovery (KD).

S. Sumathi and S.N. Sivanandam: *Data Mining & KDD*, Studies in Computational Intelligence (SCI) **29**, 231–241 (2006)
www.springerlink.com

- We discuss these two data mining cultures and some key issues in data mining from this perspective.
- The PM tradition favors accuracy over understandability; the KD culture favors understandability over complexity.

Abstract. In this section, the ideas of Data Mining and KDD process and cultures of data mining taken from National Center for Data Mining, University of Illinosis at Chicago are presented.

The objective of data mining is to convert data and information into profitable knowledge for actionable strategic decisions. The process consists of three elements: data, mining and business issue. The data involved are large databases of geo-demographic, demographic, transactional and relationship type, obtainable from inside as well as outside the organization. They are both current/instantaneous and historical in nature. The mining process can go back to query systems and OLAP (on line analytical processing). However, we shall limit ourselves to the *knowledge discovery* part of data mining. Knowledge discovery is driven by advances in technologies, in particular, in Information Technologies, computing power, software performance, machine learning, and optimization. It allows us to dig deep into masses of data to detect hidden patterns, interesting behaviors useful trends or niche profiles that we had not suspected to exist. In order to detect these useful elements for particular decision, the business issues have to be addressed. The strategy has to be clearly articulated with well-defined business objectives. In a practical approach to data mining, other factors such as data mining tools, infrastructure, data management system, human expertise, solution architecture, and the use of data also contribute to the success of data mining endeavor. In this section, we also discuss these two data mining cultures and some key issues in data mining from this perspective.

9.1 Data Mining and KDD – Overview

Data mining is an idea based on a simple analogy. The growth of data warehousing has created mountains of data. The mountains represent a valuable resource to the enterprise. But to extract value from these data mountains, we must "mine" for high-grade "nuggets" of precious metal – the gold in data warehouses and data marts. The analogy to mining has proven seductive for business. Everywhere there are data warehouses, data mines are also being enthusiastically constructed, but not with the benefit of consensus about what data mining is, or what process it entails, or what exactly its outcomes (the "nuggets") are, or what tools one needs to do it right.

Data mining as a field is not yet through with the process of definition and conceptualization of the scope of the field. There are at least three distinct concepts of data mining that are being used by the practitioners and vendors.

DMI: Data mining is traditional data analysis methodology updated with the most advanced analysis techniques applied to discovering previously unknown patterns. A specific instance of this concept is stated more explicitly and with a more commercial orientation is provided by the SAS Institute [http://www.sas.com/feature/4qdm/whatisdm.html].

SAS defines data mining *as the process of selecting, exploring, and modeling large amounts of data to uncover previously unknown patterns for a business advantage.*

In specifying its notion of data mining further, SAS describes it as involving a five-step process: Sample, Explore, Modify, Model, and Assess, or the acronym SEMMA. The difference between SEMMA and traditional methodology used in statistical analysis is hard to see with the naked eye, though we emphasize that methodology and tools or techniques are different things, and we are certainly not saying that because SAS's SEMMA methodological approach is traditional, but it would not or could not incorporate the most advanced data mining tools.

DMII: Data mining is the activity of automatically extracting the hidden information (pattern and relationships) from large databases: that is, without benefit of human intervention or initiative in the knowledge discovery process. In this view, data mining is knowledge discovery in databases or at least it is automated knowledge discovery in databases.

DMII is the data-mining concept implicit in the advertising collateral of many specialized data mining companies. Customers are promised an automatic process of model development that requires little or no human interaction from sophisticated data analysts. The data mining package supplies the necessary high quality analysis, and business users are promised that they can achieve knowledge discovery and predictive success on their own with little investment of time or effort compared to what is necessary with nondata mining (often labeled as traditional statistical) techniques. We do not see this concept as much outside of the vendor literature, but it is either present or closely approached in many articles on data mining. It advocates drawing a sharp distinction between data-driven tools using automated discovery-based approaches and user- or verification-driven tools using hypothesis-testing approaches. The hypothesis-testing tools are seen as the ones limited by the skill and experience of humans, while the data mining tools are seen as free of human initiative or assumption, and empowered by pattern-matching algorithms. Most importantly, the hypothesis-testing tools are seen as "verifiers", while the data mining tools are seen as "discoverers".

DMIII: Data mining is the step in the process of knowledge discovery in databases, that inputs predominantly cleaned, transformed data, searches the data using algorithms, and outputs patterns and relationships to the interpretation/evaluation step of the KDD process. DMIII is the statement of the view of data mining emerging from the 1994 AAAI workshop on KDD, the KD Mine [http://www.kdnuggets.com], and S*I*FTWARE [http://kdnuggets.com/siftware.html] web sites and the recent Advances in Knowledge Discovery and Data Mining volume (U. Fayyad, G. Piatetsky-Shapiro, P, Smyth, and R. Uthurusamy (eds.), 1995). The definition clearly implies that what the data mining (in this view) discovers is the hypotheses about patterns and relationships. Those patterns

and relationships are then subject to interpretation and evaluation before they can be called knowledge.

The commitment to search algorithms in DMIII does not imply a completely automated data mining process. Data analyses must be algorithms with some degree of search autonomy to qualify as instances of data mining, but they will also use human initiative in the areas of background knowledge (a specification of which is required for applying some machine learning and case-based reasoning techniques), model selection and specification, input and output variable selection and specification, in constraining model parameters and in other ways. In short, the data mining process described by those adhering to DMIII is one in which automated search algorithms play a vital role in complex iterative "interaction, protracted over time, between a *human* and a database."

9.1.1 The Idea of Knowledge Discovery in Databases (KDD)

The three definitions of data mining are also closely associated with three apparently different concepts of KDD. DMI is associated with no explicit concept of KDD. But for the purposes of discussion, we will assume that supporters of DMI believe that KDD refers to a process that uses computer-based data analysis as a primary means of investigation, and that produces scientifically validated knowledge. Here are the three KDD concepts.

KDDI: Knowledge discovery in databases in a process that requires hypothesis or model formulation, hypothesis or model testing, and derivatively all the data, techniques, and sub-processes necessary to bring hypothesis or model testing to a successful conclusion. In this view, data analysis includes both exploratory and confirmatory data analysis, and the latter is necessary for hypothesis or model testing. That outcome of hypothesis or model testing is knowledge discovery, even if the knowledge discovered is a negative finding that some hypothesis or model is not knowledge.

KDDII: True knowledge discovery in databases is the process of automated data mining applied without benefit of human intervention or initiative. According to this view there is no distinction between data mining and KDD. The data mining does not just generate hypotheses. It produces valid knowledge that business can apply without fear of bad results.

KDDIII: *"Knowledge discovery in databases* is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." Further, this process includes five steps: data selection, data preprocessing, data transformation, data mining and interpreting, and evaluating mined patterns and relationships. This process is interactive and interaction with KDD users is heavily involved at every step. Many loops may occur between steps. There is no deterministic progression assumed from one step to another. Also, the interpretative and evaluative step can involve returns to any of the previous steps, any number of times.

9.1.2 How Data Mining Relates to KDD

The relationships between data mining and KDD are different for the three approaches generated by the DM/KDD pairs, and are implicit in the definitions already presented. We will consider these implications in the same order the contrasting DM and KDD concepts were presented.

Traditional Data Mining

DMI and KDDI, equate data mining with KDD and since they do not distinguish it from previous investigative processes, they essentially equate both with previous investigative processes and methodologies of analytical and statistical modeling. SAS's SEMMA data mining process could be used equally well to describe traditional processes of analysis followed for years by SAS users. Some of the tools, such as Neural Networks and Tree-based models may be different, but the patterns of investigation, and more important, of validation, are essentially the same. It is hard to escape the conclusion that for this approach, data mining is traditional modeling and analysis updated with the addition of some new technique and incorporated into the commercially relevant data warehouse framework.

In drawing this conclusion, we do not mean to be pejorative or to express criticism for traditional approaches. If the DMI/KDDI explication of data mining makes the most sense for further development of the field, so be it. But it is important to recognize the approach for what it is, and to refrain from claiming methodological novelty, when we are really talking about progress in software and hardware tools for data analysis.

In the DMI/KDDI approach, also, data mining is not restricted to the step of hypothesis formation. The SEMMA model assessment step is a validation step. That is why data mining and KDD can be so easily equated. But though data mining and KDD are equated, the data mining/KDD process is not viewed as fully automated. The traditional approach recognizes the vital role of human-initiated hypothesis and model formation, and computer-based, partly automated, exploratory and confirmatory data analysis in data mining.

Automated Data Mining

The DMII and KDDII definitions of data mining and knowledge discovery in databases also equate the two. But in contrast to DMI/KDDI, they imply that ***automated data mining,*** as expressed by DMII, ***includes both hypothesis formulation and scientific validation in the algorithmic process.*** That is why, it is logically consistent for proponents of DMII to claim that data mining and knowledge discovery are one and the same, and it is also the reason why they view the knowledge discovery process as one that excludes human intervention or initiative. But are there any commercial data mining products, or even any research efforts that fit this definition? The

answer is: it depends on what we mean by scientific validation of KDD, a matter on which there is far from universal agreement. All data mining products produce patterns and relationships from appropriate data input. So, all can produce hypotheses. But products may vary in the specific validation criteria they use empirical fit to sample data mining process. All data mining products seem to use empirical fit to sample data, as a general validity criterion, but the specific algorithms evaluating empirical fit may vary from product-to-product, and the criteria for measuring empirical fit may also vary. Also empirical fit of patterns to data is not the whole story, sometimes patterns can be over fit to data, and different products employ different tests of over fit.

Apart from empirical fit of patterns to data, there are many other validation criteria that can be applied to interpretation and evaluation of the results of data mining. To fully understand this, it is necessary to step back from the perspective of one or two products of a particular type, and to recognize that there are many data mining models out there, and they offer alternative theories of data. Which one is the right one for our data mine? If we compare the results of a number of alternatives, what criteria do we use to compare them? What if all fit the data well empirically, but the criteria of empirical fit are either not readily compared because of the use of different fit statistics, or a comparison is not meaningful because we cannot tell which model involves over fitting? What if the same data-mining product has generated alternative patterns based on slightly different input assumptions with number material difference in empirical fit? How do we choose then?

Recent KDD research has specified a number of criteria apart from empirical fit to sample data as relevant including: **predictive validity** on new data, **novelty** of the patterns discovered by the data-mining tool, **pragmatic utility** of the patterns as measured by some utility function, **ultimate understandability** of the patterns, and a composite of these called **"interestingness."** However, even though these criteria can be listed, research on applying them is not far advanced, and promises to be difficult to implement. Nor are these criteria in any sense exhaustive. Almost anyone in the KDD field today could specify additional criteria or alternatives for at least some of the criteria of validation and provide an equally plausible defense of these as reasonable validation criteria.

Validity criteria in KDD are a developing area of research, and there is no consensus yet on standards, procedures, or algorithms for measuring validity. Without such a consensus the DMII/KDDII concept of automated data mining is premature. The results of data-mining activities cannot now be validated by an algorithm or algorithms incorporating generally agreed upon validity standards. And there is no prospect that such validation will be available in the near future.

So, whatever the preferences of advocates of the DMII/KDDI position, the outcome of current automatic data mining investigations in the DMII/KDDII sense, must be viewed as highly hypothetical, exploratory in nature, and subject to a careful validation analysis before they are relied on for practical

applications considering the exploratory nature of results using the automated data mining perspective, we believe that vendors and consultants who are selling data mining on the basis of the DMII/KDDII position, are overselling data mining.

Data Mining as Part of KDD

The DMIII/KDDIII position is probably the one with the most current momentum. It attempts to distinguish data mining from traditional analyses by emphasizing the automated character of data mining in generating patterns and relationships, but it also clearly distinguishes the data mining from knowledge discovery, by emphasizing the much broader character of KDD as an overarching process, including an interpretation and evaluation step distinct from data mining and relying more heavily on human interaction. In a very real sense DMIII/KDDII is a middle way between the other two positions.

But if DMIII/KDDIII is a middle way, that does not necessarily, mean it is the right way. Sometimes compromises are just unstable platforms for methodological development. DMMII/KDDIII seems to postulate no difference from the traditional data mining approach in the area of validation or confirmatory analysis. The difference is in the area of exploratory data analysis where practitioners holding this position emphasize the use of automated methods to generate patterns, while practitioners of DMI/KDDI do not talk quite so much about automation, but talk more about using a variety of techniques including human initiative to guide exploratory analysis. But is this difference a real methodological difference between the two camps, or just a way of maintaining a distinct identify of placing old wine in new bottles?

Current studies by participants in the KDD group make overwhelmingly clear the exhaustive interaction between human and machine that is part of the data mining process in a real KDD project. The iterative process to prepare for data mining and to implement it follows the careful investigative pattern of traditional analysis. The algorithmic techniques are more powerful than they were ten years ago, but there is no methodological requirement that pattern generation be guided solely by automated data mining techniques. Instead, the requirement is a focus on techniques with a certain degree of search autonomy – a small difference from the viewpoint of traditional data mining at best.

9.1.3 The Data Mining Future

The data mining foreseeable future will involve an appreciable human component, whether we are taking the viewpoint of either DMI or DMIIII. The problems inherent in model and variable selection, in measurement and dynamic model construction, and in pattern validation methodology all guarantee that.

But, it is also true that we will continue to make progress in the area of adaptive intelligence that underlies data mining. DMII/KDDIII may be an

invalid construct now, but research on computational models of theory evaluation will eventually bring us much closer to having measurement models of theory evaluation will eventually brings us much closer to having measurement models of validity and to having agreement on both the models and the criteria they incorporate. Also, the new analysis techniques (Neural Networks, Genetic Algorithms, Machine Learning, Bayesian Belief Networks, Fuzzy Engineering, Chaotic Dynamics, etc.) that have come to prominence in the last 10 to 15 years, and that are now becoming fully commercialized, will continue to advance in power and sophistication and to become more fully integrated in analysis methodologies that we can partially automate. For now, the practical task at hand is to bring to bear the most powerful analytical techniques at our disposal to the problem of making private and public enterprises more adaptive. Practically speaking, this means analysis of enterprise performance in all its aspects through the use of the data in data warehouses and data marts. Exploratory analysis of this data is called data mining (DMI and DMII). Sometimes confirmatory data analysis is also included in data mining (DMI). The important thing is that, for the foreseeable future, good data mining cannot be done without significant human interaction between a human data miner and the computer-based software extension. That is because data mining is not automatic. And the dream of making it so is, at best, an ideal motivating long-term development.

9.2 Data Mining: The Two Cultures

Data mining is about finding patterns in data. The importance of data mining has grown dramatically as the amount of archived and warehoused digital data has grown. The historical roots of data mining come primarily from two different directions: from statistics and from artificial intelligence. The statistical culture in data mining emphasizes the role of predictive modeling (PM). The artificial intelligence culture emphasizes the role of knowledge discovery (KD). In this section, we discuss these two data mining cultures and some key issues in data mining from this perspective.

9.2.1 The Central Issue

The central issue is simple. To illustrate it, consider using data mining for fraud detection. In the PM tradition, given a credit card transaction, telephone call, or insurance claim x, the goal is to predict whether x is fraudulent as accurately as possible. This is usually considered to be a classification problem (0 means no fraud, 1 means fraud). A classifier examines the attributes of x (such as the number of transactions during the past hour) and returns as 0 or 1, indicating whether the transaction is fraudulent or not.

Generally, more accurate classifiers are more complex. For example, a good fraud classifier for a large data set using a tree-based classifier might contain

thousands of nodes. At best this is difficult to interpret. This is a basic trade-off. In the PM tradition, increased accuracy is traded for ease of interpretation.

On the other is the KD tradition, the goal is to extract useful from large data sets. To be useful, these facts must be easily interpretable and easily actionable. For example, an algorithm for extracting shallow trees might reveal that low dollar transactions at each machines outside of certain retail stores is highly correlated with fraud. The action here might be to put in place a rule, which defers subsequent transactions for certain types of high priced retail goods. This illustrates another basic trade-off. In the KD tradition, ease of interpretation and implementation is trade for accuracy.

9.2.2 What are Data Mining and the Data Mining Process?

Data mining is one step in the data mining process. The definition of data mining and of the data mining process differs somewhat between the two cultures.

A standard definition of data mining from the KD perspective is given by Fayyad, Piatesky-Shapiro, and Symth (1996): "Knowledge discovery in databases in the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data."

Here is the PM perspective: Data mining is the automatic discovery of associations, clusters, changes, patterns, anomalies, and other significant structures in large data sets and the exploitation of these discoveries to improve predictive modeling.

Recall that the data mining process consists of a sequence of steps, which are usually repeated, in an iterative fashion. The process typically includes, 1) data preparation and cleaning, 2) data warehousing, 3) identifying relevant predictive attributes, 4) computing derived attributes, 5) data reduction and attribute projection, 6) extracting patterns relevant to the predictive attributes using one or more mining algorithms, 7-pm) predictive modeling, 7-kd) knowledge extraction 8-pm) scoring of operational and warehoused data 8-kd) interactive data analysis and discovery 9) validation, report preparation, and related activities, and 10) repeating the process as necessary.

Here Steps 7 and 8 are slightly different in the two traditions and is indicated by using the suffixes pm or kd. Data mining projects usually incorporate aspects of both the PM and KD cultures. A common strategy is for groups of analysts and modelers in an organization to focus on the KD aspects of data mining and for IT staff with operational responsibility to focus on the PM aspects of data mining.

9.2.3 Machine Learning

The essence of data mining is machine learning and this is what occurs in Step 6 above. Both the PM and DM traditions, start with a space of learning

sets L. Each element of the space is a particular learning set, that is, some data set which is to be automatically analyzed.

The PM tradition requires a space of models M. In the fraud example above, M is the space of binary classifiers (0 for normal and 1 for fraud)

In the PM tradition, data mining can be thought of as a map from L to a space of models M:

L→ M (PM perspective)

The input is a data set and the outputs are (one or more) models: the goal is to produce as accurate a classifier as possible.

Two measures are relevant here: the detection rate of the model and the rate of false positive. In practice, increasing the detection rate of model usually is accompanied by increasing the false positive rate.

The KD tradition replaces the space of models M with a space of assertions of predicates P. To be more concrete, a simple type of assertion is a conditional: if X the Y. For example, if a credit transaction is for less than $2 and the transaction occurs at a gas station, then the transaction is fraudulent. Here X is the conjunction that the transaction is for less than $2 and occurs at a gas station and Y is claim that the transaction is fraudulent.

From the KD viewpoint, learning can be thought of as a map from L to a space of assertions or predicates P:

L→P (KD perspective)

The input is a data set and the outputs are (one or more) assertions. The goal is for the predicate or predicates discovered to be as relevant and useful as possible.

Two measures are relevant here. An assertion has confidence c% if c% of the transactions that contain X also contain Y. An assertion has support s% if s% of the transactions contain X or Y.

When someone talks about automatically extracting patterns or automatically discovering information, what they really mean is that there is an algorithm which takes a learning set and produces a model (in the PM tradition) or which takes a learning set and produce one ore more predicates (in the KD tradition). When there was less digital data and more analysts, automating this step was not important. Today, with so much more digital data than can ever be analyzed the automation of this step (Step 6 in the data mining process) is a key enabling technology for a variety of scientific, engineering, and business problems.

A deeper understanding of machine learning requires that the two maps above be understood in a probabilistic framework in the sense that one tries to understand the probability that the model or assertions produced is accurate.

9.2.4 Impact of Implementation

Here, we discuss some of the practical and organizational issues in data mining projects. An executive involved in a data mining project is responsible for making sure that the results of the project can be effectively exploited by the

organization. For projects with a KD focus this means that the modeling or analyst group understands the role of data mining for assisting them and are not threatened by a new technology and that reports summarizing discover is reach the relevant decision makers. For projects with a PM focus this means that the operational managers are included in the early discussions so that the predictive models produced can be easily exploited by the organization.

When designing and implementing data mining systems, the Data-Mining Administrator (DMA) must understand whether the primary goal is 1) to improve predictive modeling of an important business process (PM) or 2) to give analysts and modelers new knowledge and insights (KD). Those offering professional services involving data mining must also be aware of the same distinction.

9.3 Summary

This section described an overview of the data mining and KDD. Also it discussed the two data mining cultures. Both the predictive modeling (PM) culture and the knowledge discovery culture (KD) are essential to data mining. In some sense, data mining is about the interaction of these two cultures and the scaling up of traditional algorithms and systems from small data sets to the large datasets, which are common today. Machine learning is a step in the data mining process. In the PM culture, this step takes a learning set and produces one or more predictive models. In the KD culture, this step takes a learning set and produces one or more assertions (which are interpreted as discovered knowledge). The essence of data mining is that data mining automates this. With the amount of data growing so quickly it is simply number longer practical to develop all predictive models or assertions by hand.

The PM tradition favors accuracy over understandability; the KD culture favors understandability over complexity. Developing a good practical solution to a data mining problem understanding both the PM and KD perspective and implementing a solution by incorporating technique from both cultures. Some problems benefit from viewpoint emphasizing the PM perspective, others from a viewpoint emphasizing the KD perspective. It is important to understand both cultures and the expectations and objectives of the project team if an appropriate data mining solution is to be successfully developed.

9.4 Review Questions

1. Describe the five step process involved in SAS
2. Differentiate DMI, DMII and DMIII
3. Differentiate KDI, KDII and KDIII
4. What is traditional and automated data mining process?
5. What is the importance of machine learning in the discovery process?