Encyclopedia of Database Technologies and Applications

Knowledge Discovery from Databases

Jose Hernandez-Orallo
Dep. of Information Systems and Computation
Technical University of Valencia, Spain
jorallo@dsic.upv.es

INTRODUCTION

As databases are pervading every parcel of reality, they record what happens in and around organizations all over the world. Databases store the detailed history of organizations, institutions, governments and individuals. An efficient and agile analysis of the data recorded in a database can no longer be done manually.

Knowledge Discovery from Databases (KDD) is a collection of technologies which aim at extracting nontrivial, implicit, previously unknown, and potentially useful information (Fayyad, Piatetsky-Shapiro, Smyth & Uthurusamy, 1996) from raw data stored in databases. The extracted patterns, models or trends can be used to better understand the data, and hence the context of an organization, and to predict future behaviors in this context that could improve decision making. KDD can be used to answer questions such as: Is there a group of customers buying a special kind of products? Which sequence of financial products improves the chance of contracting a mortgage? Which telephone call patterns suggest a future *churn*? Are there relevant associations between risk factors in coronary diseases? How can I assess my e-mail messages more or less likely to be *spam* (junk mail)?

The previous questions cannot be answered by other tools usually associated to database technology such as OLAP tools, decision support systems, executive information systems, etc. The key difference is that KDD does not convert information into (more aggregated or interwoven) information, but generates inductive models (under the form of rules, equations or other kind of knowledge) that could be sufficiently consistent with the data. In other words, KDD is not a deductive process but an inductive one.

BACKGROUND

The more and more rapid evolution of the context of organizations forces the revision of their knowledge relentlessly (what used to work before does no longer work). This provisional character of knowledge helps understand why knowledge discovery from databases, still in its inception, is being so successful. The models and patterns that can be obtained by data mining methods are useful for virtually any area dealing with information: finance, insurance, banking, commerce, marketing, industry, private and public healthcare, medicine, bioengineering, telecommunications and many other areas (Berry & Linoff, 2004).

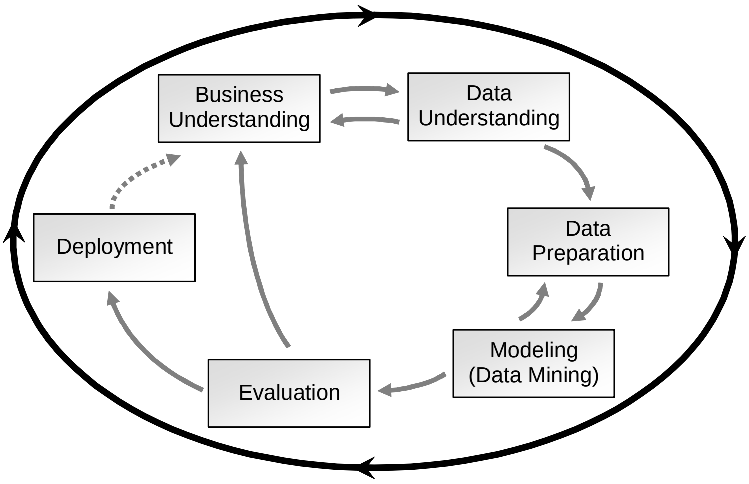
The name KDD dates back to the early nineties and is not a new “technology” itself; KDD is a heterogeneous area that integrates many techniques from several different fields without prejudices, incorporating tools from statistics, machine learning, databases, decision support systems, data visualization, WWW research, among others, in order to obtain novel, valid, and intelligible patterns from data (Han & Kamber 2001; Hand, Mannila & Smyth 2000; Berthold & Hand 2002; Dunham 2003).

THE PROCESS OF KNOWLEDGE DISCOVERY FROM DATABASES

The KDD process is a complex, elaborate process that comprises several stages (Han & Kamber, 2001, Dunham, 2003): data preparation (including data integration, selection, cleansing and transformation), data mining, model evaluation and deployment (including model interpretation, use, dissemination and monitoring).

CRISP-DM (http://www.crisp-dm.org) (CRoss-Industry Standard Process for Data Mining) is a standard reference that serves as a guide to carry out a knowledge discovery project and comprises all the stages mentioned above. Figure 1 shows these stages. According to this process, data mining is just a stage of the knowledge discovery process. The data mining stage is also called the “modeling” stage and converts the prepared data (usually in the form of a “minable view” with an assigned task) into one or more models. This stage is thereby the most characteristic of the whole process, and data mining is frequently used as a synonym for all the process. The process is cyclic, since after a complete cycle, the goals can be revised or extended, in order to start the whole process again.

Figure 1: Stages of the Knowledge Discovery Process



Many references on KDD choose the data preparation stage as the start of the knowledge discovery process (Han & Kamber, 2001, Dunham, 2003, Fayyad, 1996).

The CRISP-DM standard, however, emphasizes some issues that must be tackled before: business understanding and data understanding, which includes data integration. This is reasonable: the first thing before starting with a KDD project is to know what the business needs are, to establish the business goals according to the business context, to see whether there is (or we can get) enough data to solve them, and, in this case, to specify the "data mining objectives".

For example, consider a distribution company that has a problem of inadequate stocks, which are generating higher costs and, frequently, make some perishable products expire. From this problem, one business objective could be to "reduce the stock level of perishables". If there is an internal database with enough information about the orders performed by each customer in the last three years, and we can complement this with external information about the reference market sale prices of these product categories, this could be sufficient to start looking for models to address the business objective. This business objective could be translated into one or more data mining objectives, e.g., "predict how many perishable products by category the customer is to buy week by week, from the orders performed during the last three years and the current market sale price".

Table 1: Keys to Success in a Knowledge Discovery Project

- Business needs must drive the knowledge discovery project. The business problems and, from these, the **business objectives must be clearly stated**. These business objectives will help understand the data that will be needed and the scope of the project.
- Business objectives must be translated into specific **data mining objectives**, which must be relevant for the organization. The data mining objectives must be accompanied with a specification on the required quality of the models to be extracted in terms of a set of metrics and features: expected error, reliability, costs, relevance, comprehensibility, etc.
- The knowledge discovery project must be integrated with other plans in the organization and must have the unconditional **support from the organization executives**.
- **Data quality** is crucial. The integration of data from several sources (internal or external), its neatness and an adequate organization and availability (using a data warehouse, if necessary) is a *sinequanon* in knowledge discovery.
- The **use of integrated and friendly tools** is also decisive, helping the process on many issues, not only for the specific stages of the process but also on other issues: documentation, communication tools, workflow tools, etc.
- The **need of a heterogeneous team**, comprising not only professionals with a specific data mining training but also professionals from statistics, databases and business. A strong leadership among the group is also essential.
- In order to translate the positive results of knowledge discovery into positive results for the organization it is necessary to use a **holistic evaluation and an ambitious deployment of the models** to the know-how and daily operation of the organization.

Before starting the discussion about each specific KDD stage, it is important to highlight that many key issues on the success of a KDD project have to do with a good understanding of the goals and the feasibility of these goals, as well as a precise assessment of the resources needed (data, human, software, organizational). Table 1 shows the most important issues for success in a KDD project.

Additionally, a major reason for a possible failure in KDD projects may stem from an excessive focus on technology: implementing data mining because others do, without recognizing what the needs of the organization are and without understanding the resources and data necessary to cover them.

Data Integration and Preparation

Once the data mining goals are clear and the data that will be required to achieve them is located, it is necessary to get all the data and integrate them. Usually, the data can come from many sources, either internal or external. The typical internal data, and generally the main source of information, includes one or more transactional databases, possibly complemented with additional internal data, in the form of spreadsheets or other kind of formatted or unformatted data. External data includes all other data which can be useful to achieve the data mining objectives, such as demographic, geographic, climate, industrial, calendar or institutional information.

Some small-scale data mining projects can be done on a small set of tables or even on a couple of data files. But, generally, the size, different sources and formats of all this information require the integration into one single repository. Repositories of this kind are usually called data warehouses. Data warehouses and related technologies such as OLAP are covered by another entry in this encyclopedia. The reader is referred to this entry for more information.

It is important to note, however, that the data requirements for a data warehouse for OLAP analysis are not exactly the same that those for a data warehouse whose purpose is data mining. Data mining usually requires more fine-grain data (less aggregation) and must gather the data needed for the data mining objectives, not for performing complex reports or complex analytical queries. Even so, OLAP tools and data mining tools can work well together on the same data warehouse, which is designed taking into account both needs.

Along or after the task of determining the required data and integrating them, there is an even more thorny undertaking: data cleansing, selection and transformation. The quality of data is even more important for data mining than it is for transactional databases. Nevertheless, since the data used for data mining is generally historical, we cannot act on the sources of the data. In many situations the only possible solution is *cleansing* the data, by a heterogeneous set of techniques grouped around the term “data cleansing”, which might include exploratory data analysis, data visualization, and other techniques.

Data cleansing mostly focus on *inconsistent* and *missing* data. First, *inconsistent data* can only be detected by comparing the data with database constraints or context knowledge, or, more interestingly, by using some kind of statistical analysis for detecting *outliers*. Outliers might or might not be errors, but at least focus the attention of an expert or tool to check whether they are indeed an error. In this case, there are many possibilities: to remove the entire row, the entire column, to substitute the value by a null, by the mean or by an estimated value, among others. Secondly, *missing data* may seem easier to be detected, at least when the data come from databases which use null values, but can be more difficult if the data repository comes from several sources not using null values to represent unknown or non applicable data. The ways to handle missing values are analogous to those for handling wrong data. Data cleansing is essential since many data mining algorithms are highly sensitive to *missing* or *erroneous* values.

Having an integrated and clean data repository is not sufficient to perform data mining on them. Data mining tools may not able to handle the *overall* data repository, or this would be simply prohibitive. Consequently, it is important to select the relevant data from the data repository, de-normalize and transform them, in order to set up a view, which is called the "minable view". Data selection can be done horizontally (data sampling or instance selection) or vertically (feature selection). Data transformation includes a more varied collection of operations: creation of derived attributes: by aggregation, combination, numerization or discretization, by principal component analysis, de-normalization, pivoting, etc.

Finally, data integration and preparation must be done with a certain degree of understanding about the data itself, and frequently requires more than half of the time and resources of the whole knowledge discovery process. Furthermore, working on the integration and preparation of data, and especially through data visualization (Fayyad, Grinstein & Wierse, 2001), generates a better understanding of the data, and this might suggest new data mining objectives or could show that some of them are unfeasible.

Data Mining

The input of the data mining, modeling or learning stage, is usually a "minable view", jointly with a task. A data mining task specifies the kind of problem to be solved and the kind of model to be obtained. Table 2 shows the most important data mining tasks.

Data mining tasks can be divided into two main groups: predictive and descriptive, depending on the use of the patterns obtained. Predictive models aim at predict future values of unseen data, such as "a model that predicts the sales for next year". Descriptive models aim at describing or understanding the data, such as "there is a strong sequential association between contracting the free-mum-calls service and churn".

In order to solve a data mining task, we need to use a certain data mining technique, such as decision trees, neural networks, linear models, support vector machines, etc. A list of the most popular data mining techniques is shown in Table 3 (Han & Kamber, 2001, Hand, Mannila & Smyth, 2000, Berthold & Hand, 2002, Mitchell, 1997, Witten & Frank, 1999). There are some techniques which can be used for several tasks. For instance, decision trees are generally used for classification, but can also be used regression and for clustering. Other techniques are specific for solving one single task. For instance, the a priori algorithm is specific for association rules.

Table 2: Data Mining Tasks

- Predictive:
    - Classification: the model predicts a nominal output value (one from two or more categories) with one or more input variables. Related tasks are categorization, ranking, preference learning, class probability estimation, ...
    - Regression: the model predicts a numerical output value with one or more input variables. Related tasks are sequential prediction and interpolation.
- Descriptive:
    - Clustering: the model detects "natural" groups in the data. A related task is summarization.
    - Correlation and factorial analysis: the relation between two (bivariate) or more (multivariate) numerical variables is ascertained.
    - Association discovery (frequent itemsets): the relation between two or more nominal variables is determined. Associations can be undirected, directed, sequential, ...

Table 3: Data Mining Techniques

- Exploratory data analysis and other descriptive statistical techniques.
- Parametrical and non-parametrical statistical modeling: linear models, generalized linear models, discriminant analysis, Fisher linear discriminant functions, logistic regression, ...
- Frequent itemset techniques: Apriori algorithm and extensions, GRI, ...
- Bayesian techniques: Naïve Bayes, Bayesian networks, EM, ...
- Decision trees and rules: CART, ID3/C4.5/See5, CN2, ...
- Relational and structural methods: inductive logic programming, graph learning, ...
- Artificial neural networks: perceptron, multilayer perceptron with backpropagation, Radial Basis Functions, ...
- Support vector machines and other kernel-based methods: margin classifiers, soft margin classifiers, ...
- Evolutionary techniques, fuzzy logic approaches and other soft computing methods: genetic algorithms, evolutionary programming, genetic programming, simulated annealing, Wang & Mendel algorithm, ...
- Distance-based methods and case-based methods: nearest neighbors, hierarchical clustering (minimum spanning trees), kmeans, self-organizing maps, LVQ.

The existence of so many techniques for solving a few tasks is due to the fact that no technique can be better than the rest for all possible problems. They also differ in other features: some of them are numerical, others can be expressed in the form of rules, some are quick, others quite slow, ... Hence, it is useful to try an assortment of techniques for a particular problem, and to retain the one that gives the best model.

Model Evaluation, Interpretation and Deployment

Any data mining technique generates a tentative model, a hypothesis, which must be assessed before even thinking about using it. Furthermore, if we use several samples of the same data or use several algorithms for solving the same task, we will have several models, and we have to choose from them. There are many evaluation techniques and metrics. Metrics usually give a value of the *validity*, predictability or reliability of the model, in terms of the expectation of some defined error or cost. The appropriate metrics depend on the data mining task. For instance, a classification model can be evaluated with several metrics: *accuracy*, *cost*, precision/recall, area under the ROC curve, logloss, and many others. A regression model can be evaluated by other metrics: squared error, absolute error and others. Association rules are usually evaluated by confidence/support.

In order to estimate the appropriate metric with reliability it is well-known that the same data that was used for training should not be used for evaluation. Hence, there are several techniques to do this better: train and test evaluation, cross-validation, bootstrapping, ... Both the evaluation metric and the evaluation technique are necessary to avoid one typical problem of learned models: overfitting, i.e. lack of generality.

Once the model is evaluated as feasible, it is important to interpret and analyze its consequences. Is it comprehensible? Is it novel? Is it consistent with our previous knowledge? Is it useful? Can be put into practice? All these questions are related to the pristine goal of KDD, to obtain nontrivial, previously unknown, comprehensible and potential useful knowledge.

Next, if the model is novel and useful, we would like to apply or scatter it inside the organization. This deployment can be done manually or by embedding the obtained model into software applications. For this purpose, standards for exporting and importing data mining models, such as the Predictive Model Markup Language (PMML) standard, defined by the Data Mining Group (http://www.dmg.org/), can be of great help here.

Last, but not least, we must recall that our knowledge is always provisional. Therefore, we must monitor our models with the new incoming data and knowledge, and revise or re-generate them when they are no longer valid.

FUTURE TRENDS

Current and future areas of research are manifold. First, KDD can be specialized depending on the kind of data handled. Data mining has not only be applied to structured data in the form of, usually relational, databases, but it can also be applied to other heterogeneous kinds of data: semi-structured data, text, web, multimedia, geographical, ... This has given names to specific areas of KDD: web mining (Kosala & Blockeel, 2000), text mining (Berry 2003), multimedia mining (Simoff, Djeraba & Zaïane, 2002), ... Intensive research is being done in these areas.

Other more general areas of research include: dealing with specific, heterogeneous or multi-relational data, the latter known as relational data mining (Dzeroski & Lavrac, 2001), data mining query languages, data mining standards, comprehensibility, scalability, tool automation and user friendliness, integration with data warehouses, data preparation, ... (Smyth, 2001, Domingos & Hulten, 2001, Srikant, 2002, Roddick, 1998). The progress in all these areas will make KDD an even more ubiquitous and indispensable database technology than it is today.

CONCLUSION

This article has shown the main features of the process of Knowledge Discovery from Databases (KDD): goals, stages and techniques. KDD can be considered inside the group of other analytical, decision support and business intelligence tools that use the information contained in databases to support strategic decisions, such as OLAP, EIS, DSS or more classical reporting tools. However, we have shown that the most distinctive trait of KDD is its inductive character, which converts information into models, data into knowledge.

There are many other important issues around KDD not discussed so far. The appearance of data mining software constituted the final boost for KDD. During the nineties many companies and organizations released specific data mining algorithms and general purpose tools, known as “data mining suites”. While, initially, the tools focused on integrating an assortment of data mining techniques such as decision trees, neural networks, logistic regression, etc. nowadays the suites incorporate techniques for all the stages of the KDD process, including techniques for connecting to external data sources, data preparation tools and data evaluation and interpretation tools. Additionally, data mining suites, either vendor-specific or general, are being more and more integrated with DBMS and other decision support tools, and this trend will be reinforced in the future.

REFERENCES

Berry, M.W. (2003) *Survey of Text Mining: Clustering, Classification, and Retrieval*, Springer Verlag, 2003.

Berry, M.J.A., & Linoff, G.S. (2004) *Data Mining Techniques : For Marketing, Sales, and Customer Relationship Management*, 2nd Edition, Wiley Computing Publishers, 2004.

Berthold, M., & Hand, D.J. (ed.) (2002) *Intelligent Data Analysis. An Introduction*, Second Edition, Springer 2002.

Dzeroski, S., & Lavrac, N. (2001) *Relational Data Mining*, Springer 2001.

Domingos, P., & Hulten, G. (2001) "Catching Up with the Data: Research Issues in Mining Data Streams" Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD), 2001

Dunham, M. H. (2003) *Data Mining. Introductory and Advanced Topics*, Prentice Hall, 2003.

Fayyad U., Piatetsky-Shapiro G., Smyth P., & Uthurusamy R. (1996) *Advances in knowledge discovery and data mining*, Cambridge, M.A.: MIT Press, 1996.

Fayyad, U.M. Grinstein, G. & Wierse, A. (2001) *Information Visualization in Data Mining and Knowledge Discovery*, Morgan Kaufmann, Harcourt Intl., 2001.

Han, J., & Kamber, M. (2001) *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2001.

Hand, D.J., Mannila, H., & Smyth, P. (2000) *Principles of Data Mining*, The MIT Press, 2000.

Kosala, R., & Blockeel, H. (2000) Web Mining Research: A Survey, *ACM SIGKDD Explorations*, Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining , vol. 2, pp: 1-15, 2000.

Mitchell, T.M. (1997) *Machine Learning*, McGraw-Hill 1997.

Roddick, H, (1998) Data Warehousing and Data Mining: Are we working on the right things? *Advances in Database Technologies*. Berlin, Springer-Verlag. Lecture Notes in Computer Science. 1552. Kambayashi, Y., Lee, D. K., Lim, E.-P., Masunaga, Y. and Mohania, M., Eds. 141-144, 1998.

Simoff, S.J., Djeraba, C., & Zaïane O.R (2002) MDM/KDD 2002: Multimedia Data Mining between Promises and Problems, *SIGKDD Explorations* 4(2): 118-121, 2002.

Smyth. P. (2001) *Breaking Out of the Black-Box: Research Challenges in Data Mining*, Workshop on Research Issues in Data Mining and Knowledge Discovery DMKD 2001

Srikant, R. (2002) *New Directions in Data Mining*, Workshop on Research Issues in Data Mining and Knowledge Discovery, DMKD, 2002

Witten, I.H., & Frank, E. (1999) *Tools for Data Mining*, Morgan Kaufmann, 1999.

Terms and Definitions

Data Mining: It is a central stage in the process of knowledge discovering from databases. This stage transforms data, usually in the form of a “minable view”, into some kind of model (decision tree, neural network, linear or non-linear equation, etc.). Because of its the central stage in the KDD process and its catchy name, it is frequently used as a synonym for the whole KDD process.

Machine Learning: A discipline in computer science, generally considered a subpart of artificial intelligence, which develops paradigms and techniques for making computers learn autonomously. There are several types of learning: inductive, abductive, by analogy. Data mining integrates many techniques from *inductive* learning, devoted to learn general models from data.

Overfitting: A frequent phenomenon associated to learning, when models do not generalize sufficiently. A model can be overfitted to the training data, performing badly on fresh data. This means that the model has internalized not only the regularities (patterns) but also the irregularities of the training data (e.g. noise), which are useless for future data.

Minable View: This term is used to refer to the view constructed from the data repository which is passed to the data mining algorithm. This minable view must include all the relevant features for constructing the model.

Classification Model: A pattern or set of patterns that allows a new instance to be mapped to one or more classes. Classification models (also known as classifiers) are learned from data where a special attribute is selected as the "class". For instance, a model that classifies customers between likely to sign a mortgage and customers unlikely to do so, is a classification model. Classification models can be learned by many different techniques: decision trees, neural networks, support vector machines, linear and non-linear discriminants, nearest neighbors, logistic models, Bayesian, fuzzy, genetic techniques, etc.

Regression Model: A pattern or set of patterns that allows a new instance to be mapped to one numerical value. Regression models are learned from data where a special attribute is selected as the "output" or "dependent" value. For instance, a model that predicts the sales of the forthcoming year from the sales of the preceding years is a regression model. Regression models can be learned by many different techniques: linear regression, local linear regression, parametric and non-parametric regression, neural networks, etc.

Clustering Model: A pattern or set of patterns that allows examples to be separated into groups. All attributes are treated equally and no attribute is selected as "output". The goal is to find "clusters" such that elements in the same cluster are similar between them but are different to elements of other "clusters". For instance, a model that groups employees according to several features is a clustering model. Clustering models can be learned by many different techniques: *k*-means, minimum spanning trees (dendrograms), neural networks, Bayesian, fuzzy, genetic techniques, etc.

Association Rule: A rule showing the association between two or more nominal attributes. Associations can be directed or undirected. For instance, a rule of the form "if the customer buys French fries and hamburgers s/he buys ketchup" is a directed association rule. The techniques for learning association rules are specific, and many of them, such as the Apriori algorithm, are based on the idea of finding frequent itemsets in the data.