# Data Mining and Knowledge Discovery

Dr .T . SWARNALATHA,
Professor & HOD
Dept of MCA

V . Sireesha
Assistant Professor
Dept of MCA

Narayana Engineering College,Nellore

***Abstract:*** **- Data mining is a process that uses a variety of data analysis tools to discover patterns andrelationships in data that may be used to make valid predictions.The first and simplest analytical step in data mining is to describe the data - summarize its statisticalattributes (such as means and standard deviations), visually review it using charts and graphs, andlook for potentially meaningful links among variables. In the Data Mining Process, collecting, exploring and selecting the rightdata are critically important.Knowledge Discovery demonstrates intelligent computing at its best, and is the mostdesirable and interesting end-product of Information Technology. To be able to discoverand to extract knowledge from data is a task that many researchers and practitioners are endeavoring to accomplish. There is a lot of hidden knowledge waitingto be discovered – this is the challenge created by today's abundance of data.Knowledge Discovery in Databases (KDD) is the process of identifying valid,novel, useful, and understandable patterns from large datasets**.

*Keywords:Data mining, knowledge discovery, machine learning,datasets*

## I. INTRODUCTION:

Data Mining (DM) is the mathematical core of the KDD process, involving the inferring algorithms that explore the data, develop mathematical models and discover significant patterns
(implicit or explicit) which are the essence of useful knowledge. Advances in data gathering storage and distribution have created a need for computational tools and techniques to aid in data analysis. Data Mining and Knowledge Discovery in Databases is a rapidly growing area of research and application that builds on techniques and theories from many fields including statistics databases pattern recognition and learning data visualization uncertainty modeling data warehousing and OLAP optimization and high performance computing. KDD is concerned with issues of scalability, the multi-step knowledge discovery process for extracting useful patterns and models from raw data stores (including data cleaning and noise modeling) and issues of making discovered patterns understandable.

Knowledge Discovery includes: Theory and Foundational Issues: Data and knowledge representation; modeling of structured textual and multimedia data; uncertainty management; metrics of interestingness and utility of discovered knowledge; algorithmic complexity efficiency and scalability issues in data mining; statistics over massive data sets. Data Mining Methods: including classification clustering probabilistic modeling prediction and estimation dependency analysis search and optimization. Algorithms for data mining including spatial textual and multimedia data (e.g. the Web) scalability to large databases parallel and distributed data mining techniques and automated discovery agents.

## 2. THE KDD PROCESS:

The knowledge discovery process is iterative and interactive, consisting of several steps. The process starts with determining the KDD goals, and "ends" with the implementation of the discovered knowledge. As a result, changes would have to be made in the application domain (such as offering different features to mobile phone users in order to reduce churning). This closes the loop, and the effects are then measured on the new data repositories, and the KDD process is launched again. The following are the steps that are used:

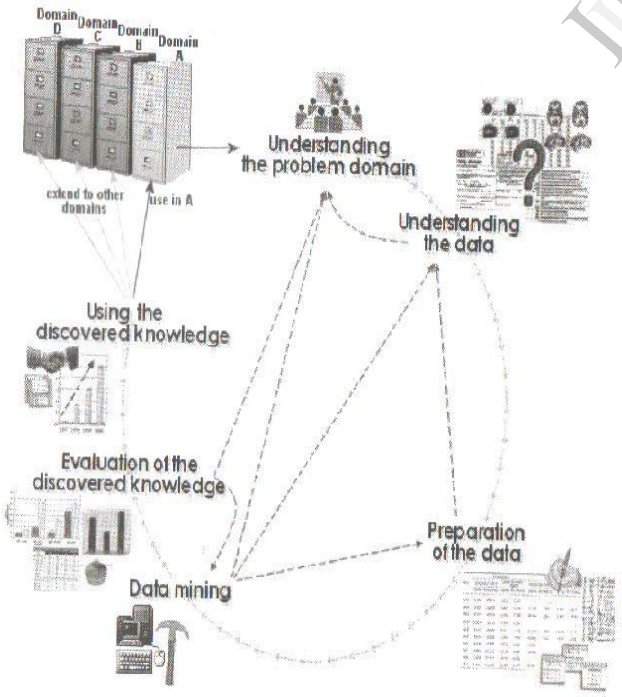
**2.1**. Developing an understanding of the application domain This is the initial preparatory step. It prepares the scene for understanding what should be done with the many decisions (about transformation, algorithms, representation, etc.). The people who are in charge of a KDD project need to understand and define the goals of the end-user and the environment in which the knowledge discovery process will take place (including relevant prior knowledge). As the KDD process proceeds, there may be even a revision and tuning of this step. Having understood the KDD goals, the preprocessing of the data starts, as

ISSN: 2278-0181

defined in the next three steps (note that some of the methods here are similar to Data Mining algorithms, but are used in the preprocessing context):

**2.2.** Selecting and creating a data set on which discovery will be performed. Having defined the goals, the data that will be used for the knowledge discovery should be determined. This includesfinding out what data is available, obtaining additional necessary data, and then integrating all the data for the knowledgediscovery into one data set, including the attributes that will be considered for the process. This process is very important because the Data Mining learns and discovers from the available data. This is the evidence base for constructing the models. If some important attributes are missing, then the entire study may fail.

From success of the process it is good to consider as many as possible attribute at this stage. On the other hand, to collect, organize and operate complex data repositories is expensive, and there is a tradeoff with the opportunity for best understanding the phenomena. This tradeoff represents an aspect where the interactive and iterative aspect of the KDD is taking place. It starts with the best available data set and later expands and observes the effect in terms of knowledge discovery and modeling.




**2.3.** Preprocessing and cleansing. In this stage, data reliability is enhanced. It includesdata clearing, such as handling missing values and removal of noise or outliers.Several methods are explained in the handbook, from doing nothing to becomingthe major part (in terms of time consumed) of a KDD process in certain projects.It may involve complex statistical methods, or using specific Data Mining algorithm

in this context. For example, if one suspects that a certain attribute isnot reliable enough or has too many missing data, then this attribute could becomethe goal of a data mining supervised algorithm. A prediction model forthis attribute will be developed, and then missing data can be predicted. The extensionto which one pays attention to this level depends on many factors. In anycase, studying these aspects is important and often revealing insight by itself,regarding enterprise information systems.

**2.4.** Data transformation. In this stage, the generation of better data for the data miningis prepared and developed. Methods here include dimension reduction (suchas feature selection and extraction, and record sampling), and attribute transformation(such as Discretization of numerical attributes and functional transformation).This step is often crucial for the success of the entire KDD project, but it isusually very project-specific. For example, in medical examinations, the quotientof attributes may often be the most important factor, and not each one by itself. Inmarketing, we may need to consider effects beyond our control as well as effortsand temporal issues (such as studying the effect of advertising accumulation).However, even if we do not use the right transformation at the beginning, wemay obtain a surprising effect that hints to us about the transformation needed(in the next iteration). Thus the KDD process reflects upon itself and leads toan understanding of the transformation needed (like a concise knowledge of anexpert in a certain field regarding key leading indicators). Having completed theabove four steps, the following four steps are related to the Data Mining part,where the focus is on the algorithmic aspects employed for each project.

**2.5.** Choosing the appropriate Data Mining task. We are now ready to decide onwhich type of Data Mining to use, for example, classification, regression, or clustering.This mostly depends on the KDD goals, and also on the previous steps.There are two major goals in Data Mining: prediction and description. Predictionis often referred to as supervised Data Mining, while descriptive Data Miningincludes the unsupervised and visualization aspects of Data Mining. Most datamining techniques are based on inductive learning, where a model is constructedexplicitly

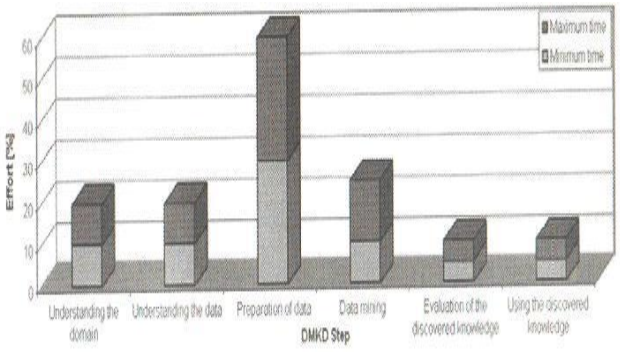
ISSN: 2278-0181

or implicitly by generalizing from a sufficient number of training examples.The underlying assumption of the inductive approach is that the trainedmodel is applicable to future cases. The strategy also takes into account the levelof meta-learning for the particular set of available data.

**2.6.** Choosing the Data Mining algorithm. Having the strategy, we now decide on thetactics. This stage includes selecting the specific method to be used for searchingpatterns (including multiple inducers). For example, in considering precisionversus understandability, the former is better with neural networks, whilethe latter is better with decision trees. For each strategy of meta-learning thereare several possibilities of how it can be accomplished. Meta-learning focuseson explaining what causes a Data Mining algorithm to be successful or not ina particular problem. Thus, this approach attempts to understand the conditionsunder which a Data Mining algorithm is most appropriate. Each algorithm hasparameters and tactics of learning (such as ten-fold cross-validation or anotherdivision for training and testing).

**2.7.** Employing the Data Mining algorithm. Finally the implementation of the DataMining algorithm is reached. In this step we might need to employ the algorithmseveral times until a satisfied result is obtained, for instance by tuning thealgorithm’s control parameters, such as the minimum number of instances in asingle leaf of a decision tree.

**2.8.** Evaluation. In this stage we evaluate and interpret the mined patterns (rules, reliabilityetc.), with respect to the goals defined in the first step. Here we consider1 Introduction to Knowledge Discovery and Data Mining 5the preprocessing steps with respect to their effect on the Data Mining algorithmresults (for example, adding features in Step 4, and repeating from there). Thisstep focuses on the comprehensibility and usefulness of the induced model. Inthis step the discovered knowledge is also documented for further usage. Thelast step is the usage and overall feedback on the patterns and discovery results obtained by the Data Mining.
The following figure presents a summary corresponding to the relative effort spent on each of the DMKD steps.




3. DATA MINING METHODOLOGY:

It should be clear from the above that data mining is not a singletechnique; any method that will help to get more information outof data is useful. Different methods serve different purposes, eachmethod offering its own advantages and disadvantages. However,most methods commonly used for data mining can be classifiedinto the following groups.

***Statistical Methods****:* Historically, statistical work has focusedmainly on testing of preconceived hypotheses and on fittingmodels to data. Statistical approaches usually rely on an explicitunderlying probability model. In addition, it is generally assumedthat these methods will be used by statisticians, and hence humanintervention is required for the generation of candidate hypothesesand models.

***Case-Based Reasoning:***Case-based reasoning (CBR) is atechnology that tries to solve a given problem by making directuse of past experiences and solutions. A case is usually a specific
problem that has been previously encountered and solved. Given aparticular new problem, case-based reasoning examines the set ofstored cases and finds similar ones. If similar cases exist, theirsolution is applied to the new problem, and the problem is addedto the case base for future reference.

***Neural Networks:***Neural networks (NN) are a class of systemsmodeled after the human brain. As the human brain consists ofmillions of neurons that are interconnected by synapses, neural
networks are formed from large numbers of simulated neurons,connected to each other in a manner similar to brain neurons. Likein the human brain, the strength of neuron interconnections maychange (or be changed by the learning algorithm) in response to apresented stimulus or an obtained output, which enables the
network to “learn”.

***Decision Trees:***A decision tree is a tree where each non-terminal node represents a test or decision on the considered data item.Depending on the outcome of the test, one chooses a certainbranch. To classify a particular data item, we start at the root nodeand follow the assertions down until we reach a terminal node (orleaf). When a terminal node is reached, a decision is made.Decision trees can also be interpreted as a special form of a ruleset, characterized by their hierarchical organization of rules.

***Rule Induction:***Rules state a statistical correlation between theoccurrence of certain attributes in a data item, or between certaindata items in a data set. The general form of an association rule is$Xl \wedge ... \wedge Xn => Y [C, S]$, meaning that the attributes $Xl,...,Xn$predict $Y$ with a confidence $C$ and a significance $S$.

***Bayesian Belief Networks:***Bayesian belief networks (BBN) aregraphical representations of probability distributions, derivedfrom co-occurrence counts in the set of data items. Specifically, aBBN is a directed, acyclic graph, where the nodes representattribute variables and the edges represent probabilisticdependencies between the attribute variables. Associated witheach node are conditional probability distributions that describethe relationships between the node and its parents.

***Genetic algorithms / Evolutionary Programming:***Genetic algorithms and evolutionary programming are algorithmicoptimization strategies that are inspired by the principles observed
in natural evolution. Of a collection of potential problem solutionsthat compete with each other, the best solutions are selected andcombined with each other. In doing so, one expects that the
overall goodness of the solution set will become better and better,similar to the process of evolution of a population of organisms.Genetic algorithms and evolutionary programming are used indata mining to formulate hypotheses about dependencies betweenvariables, in the form of association rules or some other internalformalism.

***Fuzzy Sets:***Fuzzy sets form a key methodology for representingand processing uncertainty. Uncertainty arises in many forms intoday's databases: imprecision, non-specificity, inconsistency,vagueness, etc. Fuzzy sets exploit uncertainty in an attempt tomake system complexity manageable. As such, fuzzy setsconstitute a powerful approach to deal not only with incomplete,noisy or imprecise data, but may also be helpful in developinguncertain models of the data that provide smarter and smootherperformance than traditional systems. Since fuzzy systems cantolerate uncertainty and can even utilize language-like vaguenessto smooth data lags, they may offer robust, noise tolerant modelsor predictions in situations where precise input is unavailable ortoo expensive.

***Rough Sets:*** A rough set is defined by a lower and upper bound ofa set. Every member of the lower bound is a certain member of theset. Every non-member of the upper bound is a certain nonmemberof the set. The upper bound of a rough set is the unionbetween the lower bound and the so-called boundary region. Amember of the boundary region is possibly (but not certainly) a
member of the set. Therefore, rough sets may be viewed as fuzzysets with a three-valued membership function (yes, no, perhaps).Like fuzzy sets, rough sets are a mathematical concept dealingwith uncertainty in data ([42]). Also like fuzzy sets, rough sets areseldom used as a stand-alone solution; they are usually combinedwith other methods such as rule induction, classification, orclustering methods.

## 4. INVESTIGATION OF KNOWLEDGEDISCOVERY AND DATA MINING TOOLSUSING A FEATURE CLASSIFICATIONSCHEME:

In this section we first provide a feature classification scheme tostudy knowledge discovery and data mining tools. We then applythis scheme to review existing tools that are currently available,either as a research prototype or as a commercial product.Although not exhaustive, we believe that the reviewed productsare representative for the current status of technology.As discussed in Section 2, knowledge discovery and data miningtools require a tight integration with database systems or datawarehouses for data selection, preprocessing, integrating,transformation, etc. Not all tools have the same databasecharacteristics in terms of data model, database size, queriessupported, etc. Different tools may perform different data miningtasks and employ different methods to achieve their goals. Somemay require or support more interaction with the user than theother. Some may work on a stand-alone architecture while theother may work on a client/server architecture. To capture allthese differences, we propose a feature classification scheme thatcan be used to study knowledge discovery and data mining tools.
In this scheme, the tools' features are classified into three groupscalled general characteristics, database connectivity, and data mining characteristics which are described below.

ISSN: 2278-0181

## 5. CONCLUSIONS AND FUTURERESEARCH

Knowledge discovery can be broadly defined as the automateddiscovery of novel and useful information from commercialdatabases. Data mining is one step at the core of the knowledge
discovery process, dealing with the extraction of patterns andrelationships from large amounts of data.Today, most enterprises are actively collecting and storing largedatabases. Many of them have recognized the potential value ofthese data as an information source for making business decisions.
The dramatically increasing demand for better decision support isanswered by an extending availability of knowledge discovery anddata mining products, in the form of research prototypes
developed at various universities as well as software productsfrom commercial vendors. In this paper, we provide an overviewof common knowledge discovery tasks, approaches to solve these
tasks, and available software tools employing these approaches.
However, despite its rapid growth, KDD is still an emerging field.The development of successful data mining applications stillremains a tedious process ([21]). The following is a (naturallyincomplete) list of issues that are unexplored or at least notsatisfactorily solved yet:

## REFERENCES

1. Arbel, R. and Rokach, L., Classifier evaluation under limited resources, Pattern RecognitionLetters, 27(14): 1619–1631, 2006,
2. Cohen S., Rokach L., Maimon O., Decision Tree Instance Space Decomposition withGrouped Gain-Ratio, Information Science, Volume 177, Issue 17, pp. 3592-3612, 2007.
3. Hastie, T. and Tibshirani, R. and Friedman, J. and Franklin, J., The elements of statisticallearning: data mining, inference and prediction, The Mathematical Intelligencer, 27(2): 83–85, 2005.
4. Han, J. and Kamber, M., Data mining: concepts and techniques, Morgan Kaufmann, 2006. H. Kriege, K. M. Borgwardt, P. Krger, A. Pryakhin, M. Schubert and Arthur Zimek, Futuretrends in data mining, Data Mining and Knowledge Discovery, 15(1):87-97, 2007.
5. Larose, D.T., Discovering knowledge in data: an introduction to data mining, JohnWiley andSons, 2005.Maimon O., and Rokach, L. Data Mining by Attribute Decomposition with semiconductorsmanufacturing case study, in Data Mining for Design and Manufacturing: Methods andApplications, D. Braha (ed.), Kluwer Academic Publishers, pp. 311–336, 2001.
6. Maimon O. and Rokach L., “Improving supervised learning by feature decomposition”, Proceedingsof the Second International Symposium on Foundations of Information andKnowledge Systems, Lecture Notes in Computer Science, Springer, pp. 178-196, 2002.

7.Maimon, O. and Rokach, L., Decomposition Methodology for Knowledge Discovery and Data Mining: Theory and Applications, Series in Machine Perception and Artificial Intelligence - Vol. 61, World Scientific Publishing, ISBN:981-256-079-3, 2005.