A Study on Basis of Data Mining & Techniques

M.Karpagam[1] R. Dharmarajan[2]
[1]Research Scholar [2]Assistant Professor
[1,2]Department of Computer Science & Engineering
[1,2]Thanthai Hans Roever College, Perambalur

Abstract— Data mining is the process of analyzing hidden patterns of data according to different perspectives for categorization into useful information, which is collected and assembled in common areas, such as data warehouses, for efficient analysis, data mining algorithms, facilitating business decision making and other information requirements to ultimately cut costs and increase revenue.

Key words: Data Mining, Clustering, Classification

I. Introduction

Data mining is also known as data discovery and knowledge discovery.

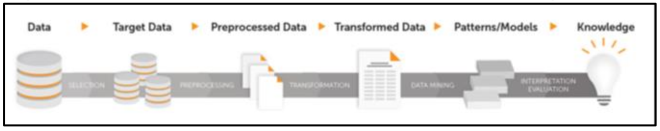
The major steps involved in a data mining process are:

- Extract, transform and load data into a data warehouse
- Store and manage data in a multidimensional databases
- Provide data access to business analysts using application software
- Present analyzed data in easily understandable forms, such as graphs

The first step in data mining is gathering relevant data critical for business. Company data is either transactional, non-operational or metadata. Transactional data deals with day-to-day operations like sales, inventory and cost etc. Non-operational data is normally forecast, while metadata is concerned with logical database design. Patterns and relationships among data elements render relevant information, which may increase organizational revenue. Organizations with a strong consumer focus deal with data mining techniques providing clear pictures of products sold, price, competition and customer demographics.

For instance, the retail giant Wal-Mart transmits all its relevant information to a data warehouse with terabytes of data. This data can easily be accessed by suppliers enabling them to identify customer buying patterns. They can generate patterns on shopping habits, most shopped days, most sought for products and other data utilizing data mining techniques.

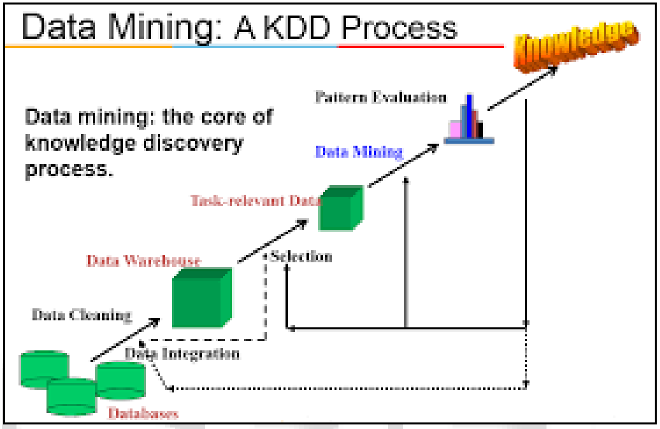
The second step in data mining is selecting a suitable algorithm - a mechanism producing a data mining model. The general working of the algorithm involves identifying trends in a set of data and using the output for parameter definition. The most popular algorithms used for data mining are classification algorithms and regression algorithms, which are used to identify relationships among data elements. Major database vendors like Oracle and SQL incorporate data mining algorithms, such as clustering and regression tress, to meet the demand for data mining.



A. KDD :

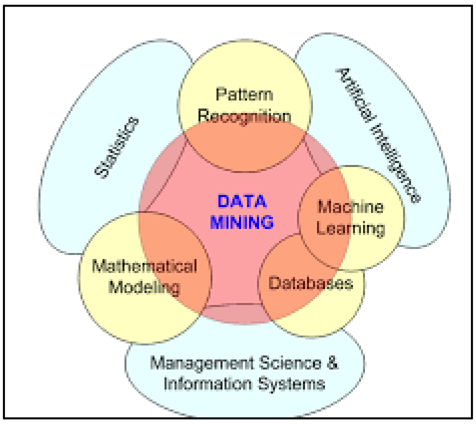
Knowledge discovery in databases (KDD) is the process of discovering useful knowledge from a collection of data. This widely used data mining technique is a process that includes data preparation and selection, data cleansing, incorporating prior knowledge on data sets and interpreting accurate solutions from the observed results.

Major KDD application areas include marketing, fraud detection, telecommunication and manufacturing.



Steps involved in the entire KDD process are:

1) Identify the goal of the KDD process from the customer’s perspective.
2) Understand application domains involved and the knowledge that's required
3) Select a target data set or subset of data samples on which discovery is be performed.
4) Cleanse and preprocess data by deciding strategies to handle missing fields and alter the data as per the requirements.
5) Simplify the data sets by removing unwanted variables. Then, analyze useful features that can be used to represent the data, depending on the goal or task.
6) Match KDD goals with data mining methods to suggest hidden patterns.
7) Choose data mining algorithms to discover hidden patterns. This process includes deciding which models and parameters might be appropriate for the overall KDD process.
8) Search for patterns of interest in a particular representational form, which include classification rules or trees, regression and clustering.
9) Interpret essential knowledge from the mined patterns.
10) Use the knowledge and incorporate it into another system for further action.
11) Document it and make reports for interested parties.



The tasks of data mining are twofold: create predictive power—using features to predict unknown or future values of the same or other feature—and create a descriptive power—find interesting, human-interpretable patterns that describe the data. In this post, we'll cover four data mining techniques:

- Regression (predictive)
- Association Rule Discovery (descriptive)
- Classification (predictive)
- Clustering (descriptive)
- Association Rule Discovery

Association rule discovery is an important descriptive method in data mining. It's a very simple method, but you'd be surprised how much intelligence and insight it can provide—the kind of information many businesses use on a daily basis to improve efficiency and generate revenue.

Our goal is to find all rules (X —> Y) that satisfy user-specified *minimum* support and confidenceconstraints, given a set of transactions, each of which is a set of items. Given a set of records—each of which contain some number of items from a given collection—we want to find dependency rules which will discover *occurrence* of an item based on *occurrences* of other items.

B. Classification

Classification is another important task you should handle before digging into the hardcore modeling phase of your analysis. Assume you have a set of records: each record contains a set of attributes, where one of the attributes is our *class* (think about letter grades). Our goal is to find a model for the*class* that will be able to *predict* unseen or unknown records (from external similar data sources)*accurately* as if the label of the class was seen or known, given all values of other attributes.

C. Clustering

Clustering is an important technique that aims to determine object groupings (think about different groups of consumers) such that objects within the same cluster are similar to each other, while objects in different groups are not. The Clustering problem in this sense is reduced to the following:

Given a set of data points, each having a set of attributes, and a similarity measure, find clusters such that:

1) Data points in one cluster are more similar to one another.
2) Data points in separate clusters are less similar to one another.

II. Conclusion:

This paper provide a broad idea of data mining, data techniques and data mining in various fields. The main objectives of data mining techniques are to discover the knowledge from active data. These applications use classification, Prediction, clustering, organization technique and so on. confidently in potential work we evaluation different classifications and clustering algorithm and its significance's.

Reference

[1] Yongjian Fu " data mining: task, techniques and application"

[2] Er. Rimmy Chuchra "Use of Data Mining Techniques for the Evaluation of Student Performance:A Case Study" International Journal of Computer Science and Management Research Vol 1 Issue 3 October 2012

[3] J. Han and M. Kamber. "Data Mining, Concepts and Techniques", Morgan Kaufmann, 2000.

[4] Aakanksha Bhatnagar, Shweta P. Jadye, Madan Mohan Nagar" Data Mining Techniques & Distinct Applications: A Literature Review" International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 9, November- 2012

[5] Brijesh Kumar Baradwaj, Saurabh Pal "Mining Educational Data to Analyze Students Performance" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011