# Data Mining Techniques Research and Evaluation

Garima Ahluwalia, garimaahluwaliya1996@gmail.com

Saraswati Institute of engineering and technology

# Introduction

Data mining is a rapidly growing field in a variety of disciplines. Finding data mining solutions that are appropriate for a certain investigation is becoming increasingly important. Data mining techniques have proven useful in a wide range of applications, including intrusion detection, manufacturing, process control, and fraud detection, marketing, and network administration. In recent years, data mining techniques have been applied to an increasing number of research projects to answer a variety of intrusion detection challenges.

## Introduction to Data Mining

Data mining, a relatively new discipline, is most typically used to extract usable knowledge from business data, but it is also beneficial in some scientific applications where a more empirical approach complements traditional data analysis. It is a necessary component of the broader process of Knowledge Discovery in Databases (KDD). The notion is that by mechanically sifting through enormous amounts of data, nuggets of insight should be extracted. Data mining has become popular not just in computer science (journals and conferences), but also in commercial IT. The rise is attributable to the expansion of data warehouses and the realization that this mass of operational data has the potential to be used as an extension of Business Intelligence.

## Objectives of data mining

Data mining techniques are used to achieve two main goals:

• To improve our understanding of the important components and their relationships, including the possible identification of non-obvious aspects in the data that may propose improved formulations of the physical models; and • to improve our understanding of the relevant factors and their relationships.

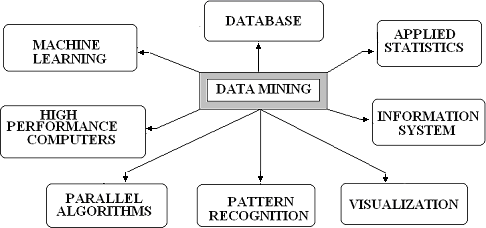
• To generate models simply from data, so that dynamical simulations may be compared to them and they can also be useful, providing (short-term) forecasting capability.

## Scope of Data Mining

The term "data mining" comes from the analogy between searching for important business information in a vast database and mining a mountain for a vein of gold ore. Both approaches need either sifting through massive amounts of material or intelligently probing it to determine where the value lies. Given sufficiently large and high-quality databases, data mining technology can provide new business prospects by enabling the following capabilities:

• Automated prediction of trends and behaviors’.

• Automated identification of previously unknown patterns.



\

**Figure 1:** Scope of data mining.

**The Characteristics of Data Mining**

'Is data mining as valuable in science as it is in commerce?' is a crucial issue to answer in order to understand the nature of data mining. Data mining in science has a lot in common with data mining in business. However, because there is a lot of existing scientific theory and knowledge, there is less chance of knowledge emerging purely from data. However, empirical results can be valuable in science (especially where it borders on engineering), such as in suggesting causality relationships or modeling complex phenomena. Another distinction is that in trade, regulations are soft, sociological or cultural in nature, and they assume consistent behavior.

**Data mining tools and stages**

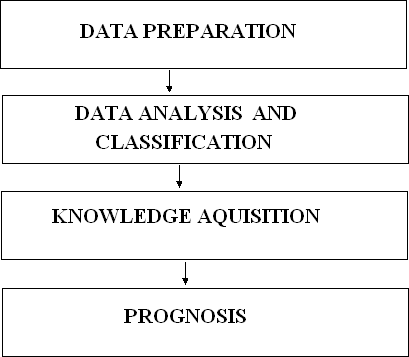
Because the tools that regulate the data mining process are not standardized, it is difficult to enumerate a precise list of data mining tool characteristics. They are not particular, and most of the time, different methodologies and technologies result in data mining, which causes families of problems. Despite the lack of precise standards, we may deduce that data mining is divided into four stages:

• Data preparation

• Data classification and analysis

• Acquiring knowledge

• Making a prognosis

They are followed one after the other in sequence, as seen in the diagram below.

**Figure:** Tools and phases of data mining

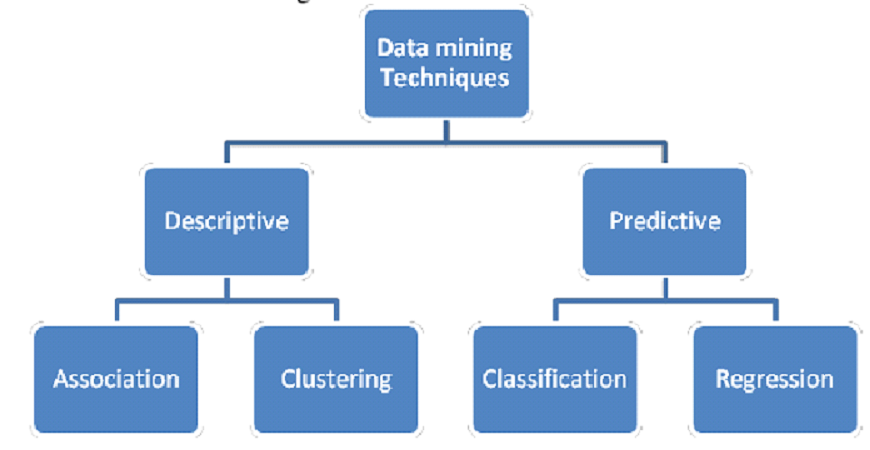
The key data sets to be used by the data mining operation are identified and cleansed of any contaminants as data in the data warehouse during the data preparation phase. Data warehouses are typically the target set for data mining operations because they are already integrated and filtered.

**Representations**

A decision tree, a neural net, a forecasting model, or a visual presenting interface can be used to depict the entire collection of findings, which can then be used to display future events or results. Many novel representation techniques are expected while data mining technology is still in its infancy.

**Mining Data Techniques**

The approaches used for data mining can be broadly classified into three groups, as shown in the picture below.



**Figure:** Techniques for data mining

**Association Regulations**

This method is typically used with transactional databases. These are databases in which each record represents a transaction, usually a sale. A transaction is defined formally as "given a set of items I, each transaction is a subset of the items". For example, the goods may represent all of a store's possible products, and the transaction specifies which of these products were purchased on a specific occasion by a specific consumer. A relation: A B, where A and B are both subsets of the items, is defined as an association rule. If the rule is valid, it means that all of the database entries that include the items of A also have the items of B. In our case, this may be translated as "all customers who purchased items #10 and #14 also purchased items #19 and #75."

**Modelling for prediction**

It is a type of supervised learning in which observations are used to learn and predict. It can also be of two types: classification and regression.

**Clustering: database segmentation**

The technique focuses on creating database segments in order to create clustering. Unsupervised learning is another name for it. It is the process of generating clusters or segments based on similar associations or traits. It is typically undertaken with the intention of capturing gestalt features of a cluster rather than focusing solely on commonalities.

Clustering is related to classification in certain ways. Again, the goal is to partition database records into similar homogenous groups, but this time the user does not know the classes prior to the analysis. The clustering algorithm must first determine the most natural way to group the records together before proceeding with the grouping. Clustering is most effective in spatial databases. These are databases in which each entry represents a point in space. The clustering algorithm locates all points that belong to the same cluster. For example, if our database represents an insurance company's customers, they will accordance to similar behaviours. Outliners are customers who exhibit uncommon behavior. This could conceal an attempt to defraud the company, necessitating further inquiry. This application of clustering is known as fraud detection.

**Methods for Clustering**

There are four fundamental methods to the clustering problem:

• Clustering based on partitions

• Clustering based on grids

• Clustering based on hierarchies

• Clustering based on density

**How Does Data Mining Work?**

'How exactly does data mining teach you significant stuff you didn't know or what will happen next?' is an important factor to consider. Modeling is the technique utilized to accomplish these accomplishments in data mining. Modeling is just the process of creating a model in one case where you know the answer and then applying it to another where you don't. For example, if you were seeking for a sunken galleon on the high seas, you may start by researching periods when wealth was discovered by others in the past. You may have noticed that these ships are frequently located off the coastlines, and that there are certain and that there are specific qualities to the ocean currents and paths that the ship's commanders in that era are likely to have taken. You take note of these similarities and construct a model that incorporates the properties that are shared by the locations of these buried treasures. With these models in hand, you set out in search of treasure where your model suggests it might be given a similar situation in the past. If you have a good model, you should be able to find your prize.

**References**

1. Mazda Salmanian, Julie H. Lefebvre, Steve Leonard and Scott Knight, “Intrusion Detection in 802.11 Wireless Local Area Networks”, Technical Memorandum, Defence R&D Canada & Ottawa , July 2004.
2. H.BELLAAJ, R.KETATA and A.HSINI, "Fuzzy approach for 802.11wireless intrusion detection", in proc. of 4th International Conference: Sciences of Electronic, Technologies of Information and Telecommunications, March 25- 29, Tunisia, 2007.
3. Zonghua Zhang and Hong Shen,” A Brief Observation-Centric Analysis on Anomaly-Based Intrusion Detection”, Springer-Verlag Berlin Heidelberg 2005
4. Mark Handley and Vern Paxson “Network Intrusion Detection: Evasion, Traffic Normalization and End-to-End Protocol Semantics”,
5. Martin Rehak, Michal pechoucek, karel Bartos, Martin Grill, Pavel celeda and vojtech krmick “An intrusion detection system for high-speed networks”, national institute of informatics, 2008
6. John Haggerty, Qi Shi and Madjid Merabti,” STATISTICAL SIGNATURES FOR EARLY DETECTION OF FLOODING DENIAL-OFSERVICE ATTACKS”, Springer Boston, 2006
7. Giovanni Vigna, Sumit Gwalani, Kavitha Srinivasan, Elizabeth M. Belding- Royer and Richard A. Kemmerer ,” An Intrusion Detection Tool for AODV- based Ad hoc Wireless Networks”, IEEE Computer Society Washington, DC, USA ,2004.
8. Shukor Abd Razak, Steven Furnell, Nathan Clarke, and Phillip Brooke,” A Two-Tier Intrusion Detection System for Mobile Ad Hoc Networks – A Friend Approach”, Springer-Verlag Berlin Heidelberg 2006
9. Eduardo Mosqueira-Rey, Amparo Alonso-Betanzos, Belen Baldonedo Del Rio, and Jesus Lago Pineiro, ” A Misuse Detection Agent for Intrusion Detection in a Multi-agent Architecture”. Springer-Verlag Berlin Heidelberg 2007
10. Magnus Almgren, Ulf Lindqvist, and Erland Jonsson,” A Multi-Sensor Model to Improve Automated Attack Detection”, Springer-Verlag Berlin Heidelberg 2008
11. Curtis A. Carver, Jr., Jeffrey W. Humphries, and Udo W. Pooch,”Adaptation Techniques for Intrusion Detection and Intrusion Response Systems”,
12. Naeimeh Laleh and Mohammad Abdollahi Azgomi,” A Taxonomy of Frauds and Fraud Detection Techniques”, Springer-Verlag Berlin Heidelberg 2009.
13. Jeyanthi Hall Michel Barbeau and Evangelos Kranakis, “Detecting Rouge Devices In Bluetooth Networks Using Radio Frequency Fingerprinting”

,School Of computer Science, Carleton University.

1. Jeyanthi Hall , Michel Barbeau and Evangelos Kranakis, “Enhancing Intrusion Detection In Wireless Networks Using Radio Frequency Fingerprinting”, School Of computer Science, Carleton University.
2. M. Charikar, E. Lehman, D. Liu, R. Panigrahy, M. Prabhakaran, A. Sahai, A. Shelat, “The smallest grammar problem,” IEEE Transactions on Information Theory, vol. 51, Issue 7, pp. 2554-2576, July 2005.Keesook J. Han and Joseph Giordano, “Intrusion Detection System Modeling,” Proceedings of the HPCMP UGC 2006, IEEE Computer Society, June 2006.