# Data Mining Techniques Research and Evaluation

Garima Ahlauwaliya, garimaahluwaliya1996@gmail.com

Guru Ramdas Khalsa Institute of Science & Technology, Jabalpur

**Abstract**

Data mining is one of the fastest-growing fields in many fields. Finding the right data mining solutions for a particular investigation is now more important than ever. Data mining technologies have been used in a wide variety of applications, from intrusion detection, to manufacturing process control, to fraud detection, to marketing, to network administration. More and more research projects have used data mining techniques to solve a range of intrusion detection issues in recent years.

## Introduction to Data Mining

Although data mining is most commonly used to extract valuable insights from business data, it is also useful for some scientific applications where an empirical approach is complementary to traditional data analysis Abstracts mining is a basic allotment of the KDD (Knowledge Discovery in Databases) process. The abstraction is that by mechanistically coursing through all-inclusive amounts of data, admired insights can be extracted [1]. Data mining has become increasingly popular not only in computer science (Journals and Conferences), but in commercial IT as well. The increase in popularity can be attributed to the growth of data warehouses as well as the realization that this vast mass of operational data can be used as a bridge to Business Intelligence [2].

## Objectives of data mining

The primary objectives of data mining are to: · Increase our knowledge of the critical elements and their relationship to each other, including the potential to identify non-observed elements in the data which may suggest better formulations of physical models [3]:

 • Improve our knowledge of the important factors and their relationship to one another.

 • Generate models directly from data so that they can be compared to dynamical simulations, which can also provide (short term) forecasting capability.

## Scope of Data Mining

The appellation “data mining” is acquired from the affinity amid mining a abundance for gold ore and mining a database for important business information. Both methods involve sifting through vast amounts of material or analyzing it intelligently to identify where the value is [4]. When databases are large enough and of sufficient quality, data mining technologies can open up new business opportunities by enabling:

 • Automatically predict trends and behaviors.

• Automatically identify previously unknown patterns.



\

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

 **Characteristics**

Is data mining in science as valuable as data mining in commerce? The answer to this question is two-fold. First, data mining in science is very similar to data mining in commerce. There are a lot of scientific theories and knowledge out there, so there’s less chance that knowledge will come directly from data. On the other hand, empirical results are valuable in science, especially where it comes down to engineering[5] . For example, they can be used to suggest causal relationships or to model complex phenomena. Second, data mining in commerce is very different from data mining in science. Regulations in trade are softer, sociological, or cultural, and assume consistent behavior.

1. **Data mining tools and stages**

Because the tools that control the data mining process aren’t standardized, it’s hard to give an exact list of the characteristics of a data mining tool. They’re not specific, and in most cases, different techniques and technologies lead to data mining, which creates families of issues. In the absence of exact standards, we can only assume that data mining is broken down into four phases[6] :

• Data Preparation

• Data Classification and Analysis

 • Learning

• Making a Prediction

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

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

During the data preparation stage, the key data sets that will be used in the data mining operations are identified and cleaned as data in a data warehouse[7] . Typically, a data warehouse is the data set to be used in a data mining operation as it has already been integrated and filtered.

1. **Representations**

You can use a decision tree, neural net, forecasting model, or visual presenting interface to represent the entire set of findings, which you can then use to visualize future events or outcomes. Many new representation technologies are anticipated while data mining technologies are still in their early stages[8].

1. **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

1. **Association Regulations**

This method is most commonly used in transactional databases. Transactional databases are databases where each record is a transaction, typically a sale. Traditionally, a transaction is defined as a subset of the set of items[9] I. For example, a set of goods represents all of a store’s possible products, and a transaction specifies which of those goods were purchased on a particular occasion by a particular consumer. An association rule is defined as A B, where “A” and “B” are subsets of “I.” If the rule is true, that means that all database entries that contain the items of A include the items of B as well. For example, in our example, all customers who bought items #10, #14, and #19 also bought items #75.

1. **Modelling for prediction**

It is a form of supervised learning where learning and prediction are based on observations[10]. Classification and regression can also be used.

1. **Clustering: database segmentation**

This technique is focused on building database segments to build clusters. Another name for this is unsupervised learning. This is the practice of building clusters or segments on the basis of similar associations or characteristics. It is usually done with the goal of capturing the “gestalt” of a cluster, rather than just the similarities [11] .

Clustering has some similarities to classification. Again, the objective is to group database records into homogeneous groups. However, this time, the user doesn’t know the classes before the analysis. The cluster algorithm must first identify the most natural way of grouping the records together before continuing with the grouping. Clustering works best in spatial databases. Spatial databases are databases where each record represents a point in the space. The cluster algorithm finds all points that are part of the same cluster[12] . For example, if we represent an insurance company’s customers in our database, they will behave similarly to each other. Outliners, on the other hand, are customers who exhibit unusual behavior. This could indicate an attempt to deceive the company and necessitate further investigation. This use case of clustering is fraud detection.

1. **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

1. **How Does Data Mining Work?**

How does data mining actually teach you important things you didn’t know or what’s going to happen next? Modeling is the process you use to achieve these successes in data mining. It’s just the process of building a model in one situation where you already know the answer, and then applying that model to another situation where you don’t[13] . For example, if you’re looking for sunken galleons on the high seas. You might start by researching when wealth was discovered in the past. These ships are often found off the coastlines. You might have noticed that there are certain ocean currents and certain ocean paths that the ship’s commanders in that era would have taken[14] . You take note of those similarities and build a model that includes the properties that are common to the locations of those buried treasures. With these models in place, you set out to find treasure where your model says it might have been found in the past. If you have a solid model, you should find your prize[15] .

**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.