Data Mining Technique and Issues

Amirali Barolia and Muhammad Nadeem
SZABIST
Karachi, Pakistan

Abstract:
With increased competition bearing down on all industries, the need of useful information to help in business decision-making has increased tremendously. Data mining, also known as Knowledge Discovery in Databases, or KDD, is a new research and applications area on the interface of computer science and statistics and aims at the discovery of useful and interesting information such as patterns, associations, changes and significant structures from large and complex data sets and repositories. It has attracted popular interest recently, due to the high demand for transforming huge amounts of data found in databases and other information repositories into useful knowledge. As data mining uses complex algorithms to generate patterns and extract valuable information those are previously hidden, the issues of efficiency, privacy, cost and scalability comes into consideration. This report focuses on all of the above referred topics certainly.

1. INTRODUCTION

Databases today can range in size into the terabytes — more than 1,000,000,000,000 bytes of data. Within these masses of data lies hidden information of strategic importance. But when there are so many trees, how do you draw meaningful conclusions about the forest? [1]

Data Mining is an idea based on a simple analogy. The growth of data warehousing has created mountains of data. The mountains represent a valuable resource to the enterprise. But to extract value from these data mountains, we must "mine" for high-grade "nuggets" of precious metal -- the gold in data warehouses and data marts. The analogy to mining has proven seductive for business. Everywhere there are data warehouses, data mines are also being enthusiastically constructed, but not with the benefit of consensus about what data mining is, or what process it entails, or what exactly its outcomes (the "nuggets") are, or what tools one needs to do it right. [2]

2. CONCEPTS OF DATA MINING

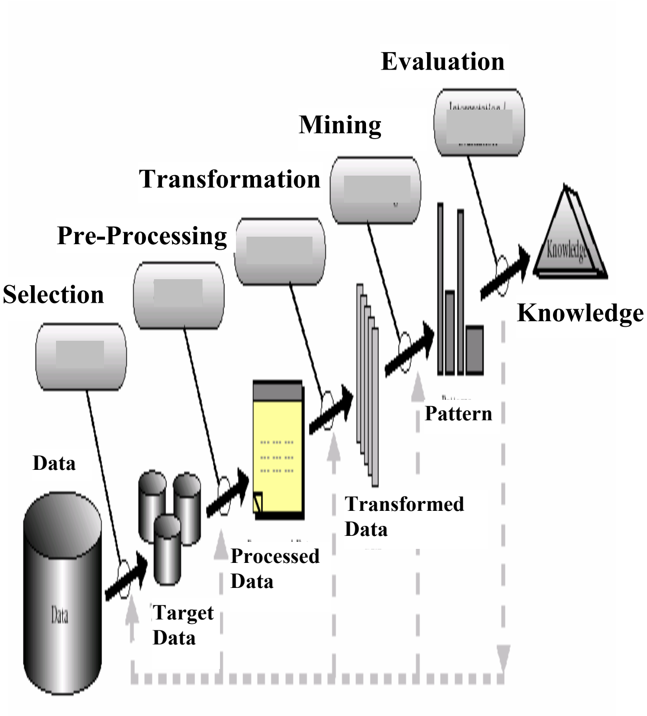
Data mining is traditional data analysis methodology updated with the most advanced analysis techniques applied to discovering previously unknown patterns. [2]

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. [3]

Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature: [3]

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

A typical data mining process can depicted as under: [4]

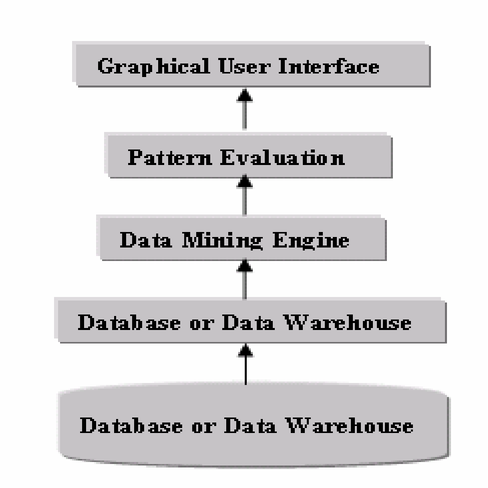


[Figure 1: A typical data mining process]

Data Mining is the activity of extracting hidden information (patterns and relationships) from large databases automatically: that is, without benefit of human intervention or initiative in the knowledge discovery process. [2]

Data mining is the process of selecting, exploring, and modeling large amounts of data to uncover previously unknown patterns for a business advantage. [5]

A typical data mining architecture can be expressed as follow[6].



[Figure 2: Data mining Architecture]

Why Data Mining?

Data mining is increasingly popular because of the substantial contribution it can make. It can be used to control costs as well as contribute to revenue increases.[1]

It also facilitates data exploration for problems that, due to high-dimensionality, would otherwise be very difficult to explore by humans, regardless of difficulty of use of, or efficiency issues with, SQL. [7]

Many organizations are using data mining to help manage all phases of the customer life cycle, including acquiring new customers, increasing revenue from existing customers, and retaining good customers. [1]

Data mining: What it can't do

Data mining is a tool, not a magic wand. It won't sit in your database watching what happens and send you e-mail to get your attention when it sees an interesting pattern. It doesn't eliminate the need to know your business, to understand your data, or to understand analytical methods. Data mining assists business analysts with finding patterns and relationships in the data — it does not tell you the value of the patterns to the organization. Furthermore, the patterns uncovered by data mining must be verified in the real world. [1]

To ensure meaningful results, it's vital that you understand your data. it is unwise to depend on a data mining product to make all the right decisions on its own. [1]

Answers to questions lie buried in your corporate data, but it takes powerful data mining tools to get at them, i.e. to dig user info for gold. [8] When users employ data mining tools to explore data, the tools perform the exploration. [9]

3. Data Preparation for Mining

Data preparation for mining is very necessary as dirty or noisy data would only produce un-reliable results.

Why preprocess the data?

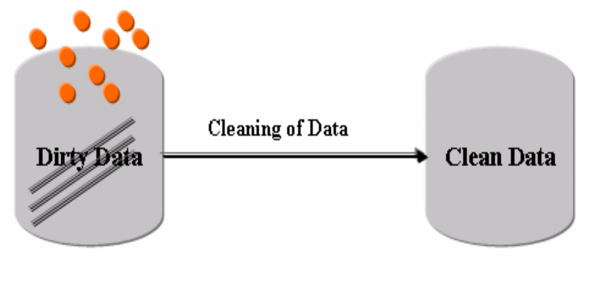
Data preprocessing plays an important role in mining. Data, which lies in transaction processing system, usually dirty. What dirty means? It may contain various errors, noisiness and inconsistencies due to different circumstances. [10]

Data is incomplete

Sometimes data is incomplete due to the circumstances when the data is collected. It may lack some attributes, necessary information, or may contain only aggregated information which may produce strange result in mining. [10]

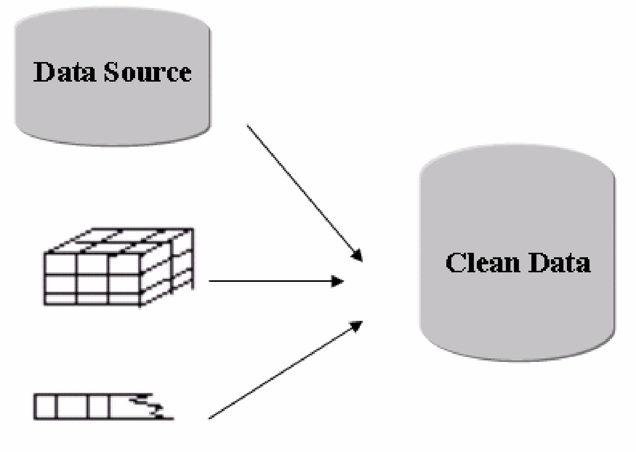
Major Tasks in Data Cleaning

It is extremely unlikely that the data you work with will be complete or free from errors. [11] GIGO (Garbage In, Garbage Out) is quite applicable to data mining, so if you want good models you need to have good data. A data quality assessment identifies characteristics of the data that will affect the model quality. [10] Essentially, you are trying to ensure not only the correctness and consistency of values but also that all the data you have is measuring the same thing in the same way. [1]



[Figure 3: Data cleaning process]

Data integration is the second step as the data you need may reside in a single database or in multiple databases. [10]

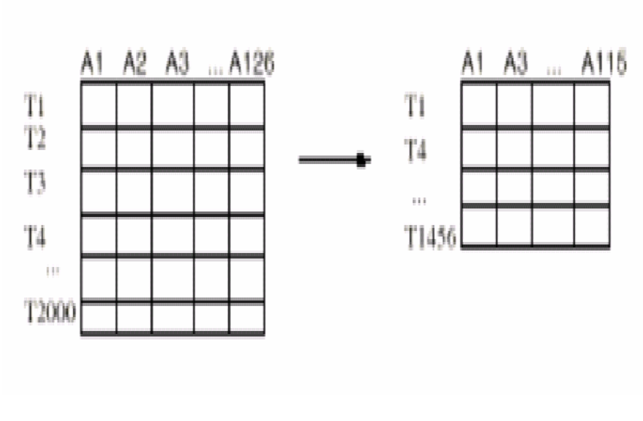


[Figure 4: Data Integration Process]

Data Transformation includes the following steps [10]

- Smoothing: remove noise from data[10]
- Aggregation: summarization, data cube construction
- Generalization: concept hierarchy climbing
- Normalization: scaled to fall within a small, specified range
- Attribute/feature construction i.e. new attributes constructed from the given ones

Obtains reduced representation in volume but produces the same or similar analytical results. [11] he term Data Reduction in the context of data mining is usually applied to projects where the goal is to aggregate or amalgamate the information contained in large datasets into manageable (smaller) information nuggets [12]



[Figure 5: Data reduction process]

4. DATA DESCRIPTION FOR DATA MINING

Clustering

Clustering of data is a method by which large sets of data is grouped into clusters of smaller sets of similar data. [13] Clustering is a division of data into groups of similar objects. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. [14]

[Figure 6: Un-clustered data]

The balls of same color are clustered into a group as shown below :

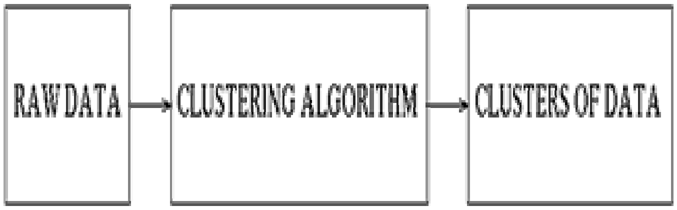
[Figure 7: Clustered data]

The goal of clustering is to find groups that are very different from each other, and whose members are very similar to each other. Unlike classification, you don't know what the clusters will be when you start, or by which attributes the data will be clustered. Consequently, someone who is knowledgeable in the business must interpret the clusters. [1]

Don't confuse clustering with segmentation. Segmentation refers to the general problem of identifying groups that have common characteristics. Clustering is a way to segment data into groups that are not previously defined, whereas classification is a way to segment data by assigning it to groups that are already defined. [1]

Clustering Algorithm

A clustering algorithm attempts to find natural groups of components (or data) based on some similarity. The clustering algorithm also finds the *centroid* of a group of data sets. **[13]**



[Figure 8: Clustering Algorithm Operation]

Types of Clustering Algorithms

The clustering algorithms operate on the raw data set. The various clustering concepts available can be grouped into two broad categories**:** [15]

- Hierarchical methods
- Nonhierarchical methods [15]

Nonhierarchical method initially takes the number of components of the population equal to the final required number of clusters [16] while hierarchical method starts by considering each component of the population to be a cluster. [17]

Association

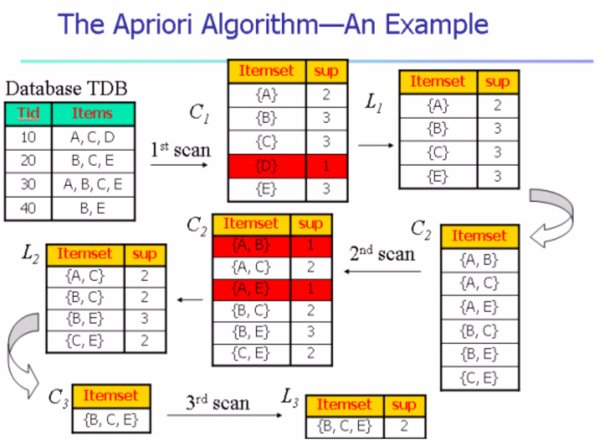
Association discovery finds rules about items that appear together in an event such as a purchase transaction. Market-basket analysis is a well-known example of association discovery. Sequence discovery is very similar, in that a sequence is an association related over time. [1]

Finding frequent patterns, associations, correlations, or causal structures among sets of items or objects in transaction databases, relational databases, etc is called association mining or discovery. [18]

Apriori: A Candidate Generation-and-test Approach for Association

This algorithm says that any subset of a frequent item set must be frequent if **{beer, diaper, nuts}** is frequent, so is **{beer, diaper}[18]**

Every transaction having {beer, diaper, nuts} also contains {beer, diaper} Apriori pruning principle: If there is any item set which is infrequent, its superset should not be generated/tested! [18]



The Apriori Algorithm—An Example

Database TDB

Tid	Items
10	A, C, D
20	B, C, E
30	A, B, C, E
40	B, E

1st scan

C_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{D}	1
{E}	3

L_1

Itemset	sup
{A}	2
{B}	3
{C}	3
{E}	3

C_2

Itemset
{A, B}
{A, C}
{A, E}
{B, C}
{B, E}
{C, E}

2nd scan

C_2

Itemset	sup
{A, B}	1
{A, C}	2
{A, E}	1
{B, C}	2
{B, E}	3
{C, E}	2

L_2

Itemset	sup
{A, C}	2
{B, C}	2
{B, E}	3
{C, E}	2

C_3

Itemset
{B, C, E}

3rd scan

L_3

Itemset	sup
{B, C, E}	2

[Figure 9: Apriori Algorithm Example]

5. Supervised Prediction & Models for Mining

There are three basic types of supervised predictions:

Classification

Classification problems aim to identify the characteristics that indicate the group to which each case belongs. This pattern can be used both to understand the existing data and to predict how new instances will behave. [1]

Data mining creates classification models by examining already classified data (cases) and inductively finding a predictive pattern. These existing cases may come from an historical database, such as people who have already undergone a particular medical treatment or moved to a new long distance service. They may come from an experiment in which a sample of the entire database is tested in the real world and the results used to create a classifier. [1]

Regression

Regression uses existing values to forecast what other values will be. In the simplest case, regression uses standard statistical techniques such as linear regression. [1]

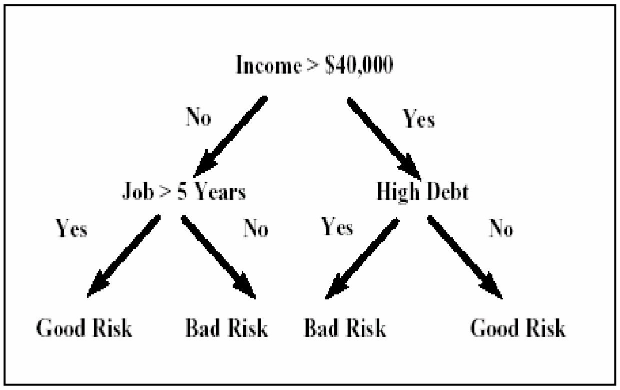
The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural nets too can create both classification and regression models. [1]

Time series

Time series forecasting predicts unknown future values based on a time-varying series of predictors. Like regression, it uses known results to guide its predictions. Models must take into account the distinctive properties of time, especially the hierarchy of periods (including such varied definitions as the five- or seven-day work week, the thirteen-“month” year, etc.), seasonality, calendar effects such as holidays, date arithmetic, and special considerations such as how much of the past is relevant. [1]

Decision trees Model

Decision trees are a way of representing a series of rules that lead to a class or value. For example, you may wish to classify loan applicants as good or bad credit risks. Figure shows a simple decision tree that solves this problem while illustrating all the basic components of a decision tree: the decision node, branches and leaves. [1]



[Figure 10: Decision Tree Example]

Decision trees which are used to predict categorical variables are called *classification trees* because they place instances in categories or classes. Decision trees used to predict continuous variables are called *regression trees* [1].

6. Data Mining Issues

Privacy

One of the key issues raised by data mining technology is not a business or technological one, but a social one. It is the issue of individual privacy. Data mining makes it possible to analyze routine business transactions and glean a significant amount of information about individuals buying habits and preferences. [19]

Recent developments in information technology have enabled collection and processing of vast amounts of personal data, such as criminal records, shopping habits, credit and medical history, and driving records. This information is undoubtedly very useful in many areas, including medical research, law enforcement and national security. However, there is an increasing public concern about the individuals' privacy. Privacy is commonly seen as the right of individuals to control information about him. The appearance of technology for Knowledge Discovery and Data Mining (KDDM) has revitalized concern about the following general privacy issues: secondary use of the personal information, handling misinformation, and granulated access to personal information. These issues demonstrate that existing privacy laws and policies are well behind the developments in technology, and no longer offer adequate protection. [19]

Handling Misinformation

Misinformation can cause serious and long-term damage, so individuals should be able to challenge the correctness of data about themselves. For example, District Cablevision in Washington fired James Russell Wiggings on the basis of information obtained from Equifax, Atlanta, about Wiggings' conviction for cocaine possession; the information was actually about James Ray Wiggings, and the case ended up in court. [19]

Stereotypes

General patterns may be used for guessing confidential properties. Also, they may lead to stereotypes and prejudices. If the patterns are based on properties such as race, gender or nationality, this issue can be very sensitive and controversial. Examples are debates over studies about intelligence across different races. The issue raises debate because KDDM tools may allow the application of different commercial standards based on race or ethnic group. Banks may use KDDM tools to find a different pattern of behavior between two racial or ethnic groups and then deny credit or apply a different policy based on this attribute. [20]

Scalability

How effective is the tool in dealing with large amounts of data — both rows and columns — and with sophisticated validation techniques? These challenges require the ability to take advantage of powerful hardware. What kinds of parallelism does the tool support? Is there parallel use of a parallel DBMS and are the algorithms themselves parallel? What kind of parallel computers does it support? How well does it scale as the number of processors increases? Does it support parallel data access? Data mining algorithms written for a uniprocessor machine won't automatically run faster on a parallel machine; they must be rewritten to take advantage of the parallel processors. [1]

Data Integrity

Another issue is that of data integrity. Clearly, data analysis can only be as good as the data that is being analyzed. A key implementation challenge is integrating conflicting or redundant data from different sources. For example, a bank may maintain credit cards accounts on several different databases. The addresses (or even the names) of a single cardholder may be different in each. Software must translate data from one system to another and select the address most recently entered. [18]

Cost

Finally, there is the issue of cost. While system hardware costs have dropped dramatically within the past five years, data mining and data warehousing tend to be self-reinforcing. The more powerful the data mining queries, the greater the utility of the information being gleaned from the data, and the greater the pressure to increase the amount of data being collected and maintained, which increases the pressure for faster, more powerful data mining queries. This increases pressure for larger, faster systems, which are more expensive. [18]

7. Conclusion

Data mining offers great promise in helping organizations uncover patterns hidden in their data that can be used to predict the behavior of customers, products and processes. Building models is only one step in knowledge discovery. It's vital to properly collect and prepare the data, and to check the models against the real world. The "best" model is often found after building models of several different types, or by trying different technologies or algorithms.

However, data mining tools must be guided by users who understand the business, the data, and the general nature of the analytical methods involved. Realistic expectations can yield rewarding results across a wide range of applications, from improving revenues to reducing costs.

Knowledge Discovery and Data Mining revitalizes some issues and posses new threats to privacy.

REFERENCES

[1]. Introduction to Data Mining and Discovery Knowledge, Third Edition, By Two Crows Corporation.

[2]. Data Mining and KDD: A Shifting Mosaic By Joseph M. Firestone, Ph.D. White Paper No. Two March 12, 1997

[3]. http://www.thearling.com/text/dmwhite/dmwhite.htm

[4]. From Data Mining to Knowledge Discovery in Databases, 1998, Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, American Association for Artificial Intelligence.

[5]. http://www.sas.com/feature/4qdm/whatisdm.html (SAS Institute)

[6]. Data Mining: Concepts and Techniques, Chapter 1, Jiawei Han and Micheline Kambe, Department of Computer Science ,University of Illinois at Urbana-Champaign (www.cs.uiuc.edu/~hanj)

[7]. Mining Databases: Towards Algorithms for Knowledge Discovery, Usama Fayyad, Microsoft Research

[8]. http://www.statoo.com/en/datamining/

[9]. http://www.db2mag.com/db_area/archives/1997/q1/9701edel.shtml#

[10]. Data Mining: Concepts and Techniques, Chapter 3, Jiawei Han and Micheline Kambe, Department of Computer Science ,University of Illinois at Urbana-Champaign (www.cs.uiuc.edu/~hanj)

[11]. Data Mining, Peter Ross, http://www.dcs.napier.ac.uk/~peter/vldb/dm/dm.html, 26th October, 2000

[12]. http://www.statsoftinc.com/textbook/stdatmin.html#concepts

[13]. http://cne.gmu.edu/modules/dau/stat/clustgalgs/clust1_frm.html

[14]. Survey of Clustering Data Mining Techniques, Pavel Berkhin, Accrue Software, Inc.

[15]. http://cne.gmu.edu/modules/dau/stat/clustgalgs/clust3_frm.html

[16]. http://cne.gmu.edu/modules/dau/stat/clustgalgs/clust5_bdy.html

[17]. http://cne.gmu.edu/modules/dau/stat/clustgalgs/clust4_frm.html

[18]. Data Mining: Concepts and Techniques, Chapter 6 Jiawei Han and Micheline Kambe, Department of Computer Science ,University of Illinois at Urbana-Champaign (www.cs.uiuc.edu/~hanj)

[19]. http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/issues.htm

[20]. Privacy Issues In Knowledge Discovery And Data Mining, Ljiljana Brankovic1 and Vladimir Estivill-Castro2, Australian Institute of Computer Ethics Conference, July, 1999, Lilydale