Database Report

Data Mining

SBE306

Submitted to:

Dr/ Ahmed Hesham

Team Number: 7

Team Members:

Mostafa Yehia

Galal Hossam

Nada Ashraf

Zeinab Walid

Mohamed Elsayed

Contents

Introduction

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information (with intelligent methods) from a data set and transform the information into a comprehensible structure for further use. Data mining is the analysis step of the "knowledge discovery in databases process" or KDD. Aside from the raw analysis step, it also involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

The term "data mining" is a misnomer, because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (*mining*) of data itself. It also is a buzzword and is frequently applied to any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) as well as any application of computer decision support system, including artificial intelligence (e.g., machine learning) and business intelligence. The book *Data mining: Practical machine learning tools and*

techniques with Java[8] (which covers mostly machine learning material) was originally to be named just *Practical machine learning*, and the term *data mining* was only added for marketing reasons.[9] Often the more general terms (*large scale*) *data analysis* and *analytics* – or, when referring to actual methods, *artificial intelligence* and *machine learning* – are more appropriate.

The actual data mining task is the semi-automatic or automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining, sequential pattern mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting are part of the data mining step, but do belong to the overall KDD process as additional steps.

The difference between data analysis and data mining is that data analysis is used to test models and hypotheses on the dataset, e.g., analyzing the effectiveness of a marketing campaign, regardless of the amount of data; in contrast, data mining uses machine learning and statistical models to uncover clandestine or hidden patterns in a large volume of data.

The related terms *data dredging*, *data fishing*, and *data snooping* refer to the use of data mining methods to sample parts of a larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These methods can, however, be used in creating new hypotheses to test against the larger data populations.

Data Mining History and Current Advances

The process of digging through data to discover hidden connections and

predict future trends has a long history. Sometimes referred to as "knowledge discovery in databases," the term "data mining" wasn't coined until the 1990s. But its foundation comprises three intertwined scientific disciplines: statistics (the numeric study of data relationships), **artificial intelligence** (human-like intelligence displayed by software and/or machines) and **machine learning** (algorithms that can learn from data to make predictions). What was old is new again, as data mining technology keeps evolving to keep pace with the limitless potential of **big data** and affordable computing power.

Over the last decade, advances in processing power and speed have enabled us to move beyond manual, tedious and time-consuming practices to quick, easy and automated data analysis. The more complex the data sets collected, the more potential there is to uncover relevant insights. Retailers, banks, manufacturers, telecommunications providers and insurers, among others, are using data mining to discover relationships among everything from **price optimization,** promotions and demographics to how the economy, risk, competition and social media are affecting their business models, revenues, operations and customer relationships.

Why is data mining important?

So why is data mining important? You've seen the staggering numbers – the volume of data produced is doubling every two years. Unstructured data alone makes up 90 percentage of the digital universe. But more information does not necessarily mean more knowledge.

Data mining allows you to:

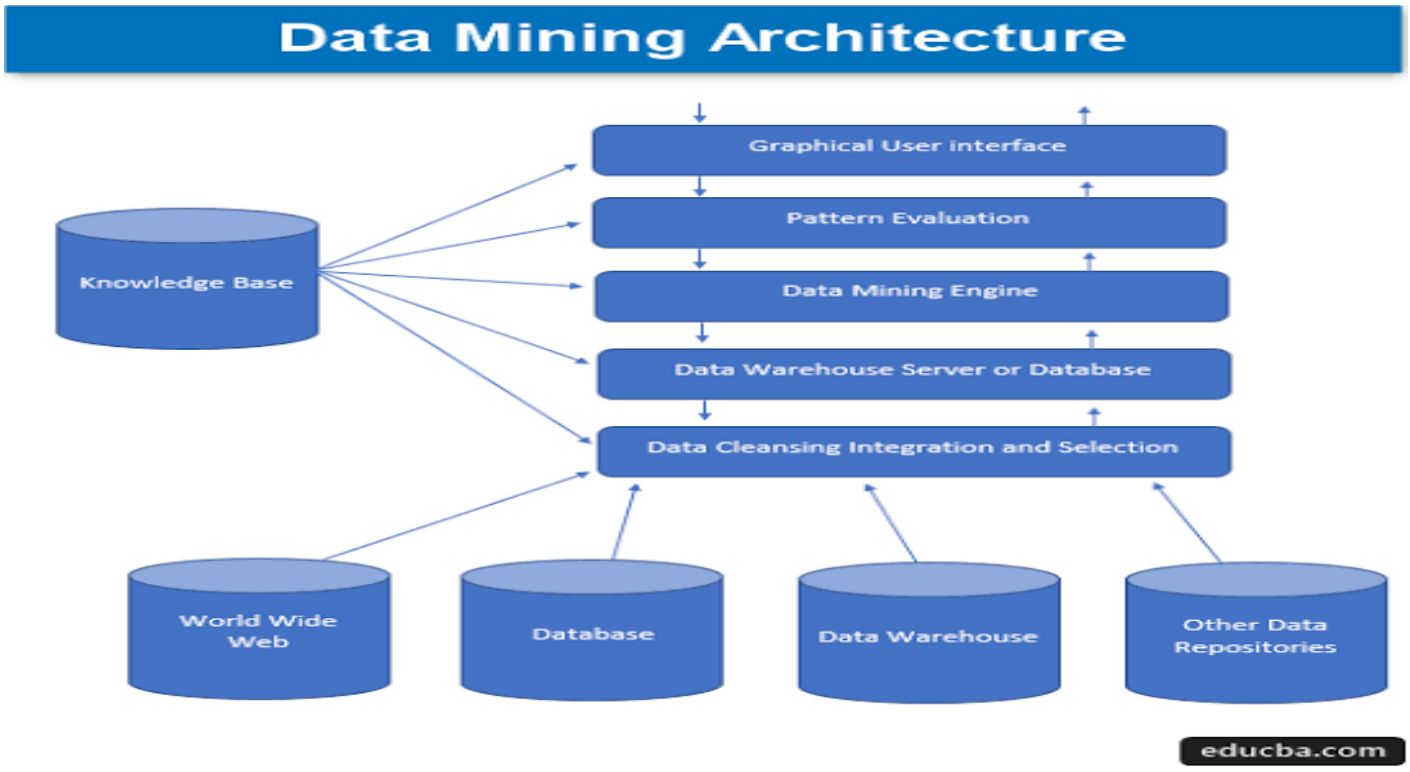
- Sift through all the chaotic and repetitive noise in your data.
- Understand what is relevant and then make good use of that

information to assess likely outcomes.

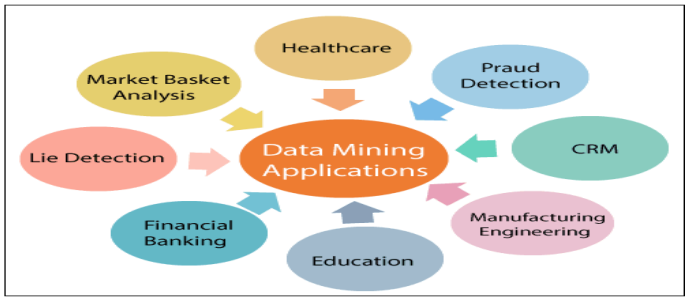
- Accelerate the pace of making informed decisions.

Data Mining Architecture

The significant components of data mining systems are a data source, data mining engine, data warehouse server, the pattern evaluation module, graphical user interface, and knowledge base.



Applications of data mining



- **Basket Analysis**

This term refers to either the real-world or virtual "shopping basket" that customers will use when purchasing items. The data analyst will look at customers' preferences and seek to predict future buying trends based on what has already happened. In addition to keeping track of products and services bought, basket analysis is also useful in monitoring payment options and rewards cards. For example, let's create a hypothetical shopper named "Sam." Sam goes to the grocery store and buys items that yield extra points on the store's rewards program. Sam then pays for his purchases with a credit card. The company can then correlate credit card use with reward program redemptions and stock levels to manage its inventory effectively.

- **Sales Forecasting**

Although this is a similar concept to basket analysis, it involves trying to guess when customers will buy certain items again in the future instead of trying to guess what they will buy. For example, if Sam buys a blender that should last three years based on performance reviews, the store from which Sam bought the blender would plan to release similar blenders three years from now so that Sam might buy one from them even if the new blenders aren't the same brand as Sam's old one. A store can also look at similar purchases customers make at other stores and tailor its approach to match that of the other stores in an effort to attract future customers.

- **Database Marketing**

By means of this **data mining** strategy, a company can create a line of products and services that sell themselves. Instead of looking at what Sam wants, the company will analyze, through data mining, what 100,000 "Sams" want. In this instance, let's say Sam is an overweight, 60-year-old Native American. The company will look at the preferences of 60-year-old, overweight Native Americans and create advertising programs that speak to that demographic. Even if one particular "Sam" doesn't respond to the advertising, the idea is that enough "Sams" will respond to it, which will make the marketing strategy worthwhile to the company.

- **Inventory Planning**

This use for data mining easy to understand. If a store has sold at least 2,500 doohickeys during every summer since 2003, then it stands to reason that the store can plan to sell at least 2,500 doohickeys next summer too. It's also fairly simple to assess increasing or decreasing sales trends on a month-to-month or even week-to-week basis and make well-reasoned decisions about which products to stock that people want to buy.

- **Customer Loyalty**

A company can look at data regarding its customers to see how price changes either attract them or send them scurrying to a slew of competitors. This data mining strategy ties in closely with a store's rewards program.

For example, if Sam has redeemed reward points for his blender, a rice cooker, a trip to Easter Island, and $1,000 in free groceries over the years, it's safe to say that he's a loyal customer. The store can plan its offerings around Sam's preferences and know he'll keep coming back. It's not all about price either. By checking out customer loyalty statistics, a company can determine what its customers consider valuable and work towardcreating extra value that falls in line with the customers' preferences irrespective of pricing.

As useful as all of these data mining techniques and the accompanying information being analyzed can be, it's essential for a business to handle them ethically. Fair use is one thing, but selling the gathered information to scam artists or fraudsters for a profit crosses the line. Besides, should the world at large find out a company has done such a thing, it won't be hard to track the buying trends of that company as they enter free fall. By being ethical and intelligent with their uses for data mining, a company can maintain its place in the world market.

- **Future Healthcare**

Data mining holds great potential to improve health systems. It uses data and analytics to identify best practices that improve care and reduce costs. Researchers use data mining approaches like multi-dimensional databases, machine learning, soft computing, data visualization and statistics. Mining can be used to predict the volume of patients in every category. Processes are developed that make sure that the patients receive appropriate care at the right place and at the right time. Data mining can also help healthcare insurers to detect fraud and abuse.

- **Education**

  There is a new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data originating from educational Environments. The goals of EDM are identified as predicting students' futurelearningbehaviour, studying the effects of educational support, and advancing scientific knowledge about learning. Data mining can be used by an institution to take accurate decisions and also to predict the results of the

student. With the results the institution can focus on what to teach and how to teach. Learning pattern of the students can be captured and used to develop techniques to teach them.

- **Manufacturing Engineering**

  Knowledge is the best asset a manufacturing enterprise would possess. Data mining tools can be very useful to discover patterns in complex manufacturing process. Data mining can be used in system-level designing to extract the relationships between product architecture, product portfolio, and customer needs data. It can also be used to predict the product development span time, cost, and dependencies among other tasks.

- **CRM**

  Customer Relationship Management is all about acquiring and retaining customers, also improving customers' loyalty and implementing customer focused strategies. To maintain a proper relationship with a customer a business need to collect data and analyze the information. This is where data mining plays its part. With data mining technologies the collected data can be used for analysis. Instead of being confused where to focus to retain customer, the seekers for the solution get filtered results.

- **Fraud Detection**

  Billions of dollars have been lost to the action of frauds. Traditional methods of fraud detection are time consuming and complex. Data mining aids in providing meaningful patterns and turning data into information. Any information that is valid and useful is knowledge. A perfect fraud detection system should protect information of all the users. A supervised method includes collection of sample records. These records are classified fraudulent or non-fraudulent. A model is built using this data and the algorithm is made to identify whether the record is fraudulent or not.

- **Intrusion Detection**

  Any action that will compromise the integrity and confidentiality of a

resource is an intrusion. The defensive measures to avoid an intrusion includes user authentication, avoid programming errors, and information protection. Data mining can help improve intrusion detection by adding a level of focus to anomaly detection. It helps an analyst to distinguish an activity from common everyday network activity. Data mining also helps extract data which is more relevant to the problem.

- **Lie Detection**

  Apprehending a criminal is easy whereas bringing out the truth from him is difficult. Law enforcement can use mining techniques to investigate crimes, monitor communication of suspected terrorists. This filed includes text mining also. This process seeks to find meaningful patterns in data which is usually unstructured text. The data sample collected from previous investigations are compared and a model for lie detection is created. With this model processes can be created according to the necessity.

- **Customer Segmentation**

  Traditional market research may help us to segment customers but data mining goes in deep and increases market effectiveness. Data mining aids in aligning the customers into a distinct segment and can tailor the needs according to the customers. Market is always about retaining the customers. Data mining allows finding a segment of customers based on vulnerability and the business could offer them with special offers and enhance satisfaction.

- **Financial Banking**

  With computerized banking everywhere huge amount of data is supposed to be generated with new transactions. Data mining can contribute to solving business problems in banking and finance by finding patterns, causalities, and correlations in business information and market prices that are not immediately apparent to managers because the volume data is too large or is generated too quickly to screen by experts. The managers may find these information for better segmenting, targeting, acquiring, retaining and maintaining a

profitable customer.

- **Corporate Surveillance**

  Corporate surveillance is the monitoring of a person or group's behaviour by a corporation. The data collected is most often used for marketing purposes or sold to other corporations, but is also regularly shared with government agencies. It can be used by the business to tailor their products desirable by their customers. The data can be used for direct marketing purposes, such as the targeted advertisements on Google and Yahoo, where ads are targeted to the user of the search engine by analyzing their search history and emails.

- **Research Analysis**

  History shows that we have witnessed revolutionary changes in research. Data mining is helpful in data cleaning, data pre-processing and integration of databases. The researchers can find any similar data from the database that might bring any change in the research. Identification of any co-occurring sequences and the correlation between any activities can be known. Data visualization and visual data mining provide us with a clear view of the data.

- **Criminal Investigation**

  Criminology is a process that aims to identify crime characteristics. Actually crime analysis includes exploring and detecting crimes and their relationships with criminals. The high volume of crime datasets and also the complexity of relationships between these kinds of data have made criminology an appropriate field for applying data mining techniques. Text based crime reports can be converted into word processing files. This information can be used to perform crime matching process.

- **Bio-informatics**

  Data Mining approaches seem ideally suited for Bioinformatics, since it is data-rich. Mining biological data helps to extract useful knowledge from massive datasets gathered in biology, and in other

related life sciences areas such as medicine and neuroscience. Applications of data mining to bioinformatics include gene finding, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing, and protein sub-cellular location prediction.

How Data Mining Works

Data mining involves exploring and analyzing large blocks of information to glean meaningful patterns and trends. It can be used in a variety of ways, such as database marketing, credit risk management, **fraud detection**, spam Email filtering, or even to discern the sentiment or opinion of users.

The data mining process breaks down into five steps. First, organizations collect data and load it into their data warehouses. Next, they store and manage the data, either on in-house servers or the cloud. Business analysts, management teams and information technology professionals access the data and determine how they want to organize it. Then, application software sorts the data based on the user's results, and finally, the end-user presents the data in an easy-to-share format, such as a graph or table.

Descriptive Modeling: It uncovers shared similarities or groupings in historical data to determine reasons behind success or failure, such as categorizing customers by product preferences or sentiment. Sample techniques include:

Clustering	Grouping similar records together.
Anomaly detection	Identifying multidimensional outliers.
Association rule learning	Detecting relationships between records.

Principal component analysis	Detecting relationships between variables.
Affinity grouping	Grouping people with common interests or similar goals (e.g., people who buy X often buy Y and possibly Z).

Predictive Modeling: This modeling goes deeper to classify events in the future or estimate unknown outcomes – for example, using credit scoring to determine an individual's likelihood of repaying a loan. Predictive modeling also helps uncover insights for things like customer churn, campaign response or credit defaults. Sample techniques include:

Regression	A measure of the strength of the relationship between one dependent variable and a series of independent variables.
Neural networks	Computer programs that detect patterns, make predictions and learn.
Decision trees	Tree-shaped diagrams in which each branch represents a probable occurrence.
Support vector machines	Supervised learning models with associated learning algorithms.

Prescriptive Modeling: With the growth in unstructured data from the web, comment fields, books, email, PDFs, audio and other text sources, the adoption of text mining as a related discipline to data mining has also

grown significantly. You need the ability to successfully parse, filter and transform unstructured data in order to include it in predictive models for improved prediction accuracy.

In the end, you should not look at data mining as a separate, standalone entity because pre-processing (data preparation, data exploration) and post-processing (model validation, scoring, model performance monitoring) are equally essential. Prescriptive modelling looks at internal and external variables and constraints to recommend one or more courses of action – for example, determining the best marketing offer to send to each customer. Sample techniques include:

Predictive analytics plus rules	Developing if/then rules from patterns and predicting outcomes.
Marketing optimization	Simulating the most advantageous media mix in real time for the highest possible ROI.

Data Warehousing and Mining Software

Data miningprograms analyze relationships and patterns in data based on what users request. For example, a company can use data mining software to create classes of information. To illustrate, imagine a restaurant wants to use data mining to determine when it should offer certain specials. It looks at the information it has collected and creates classes based on when customers visit and what they order.

In other cases, data miners find clusters of information based on logical relationships or look at associations and sequential patterns to draw conclusions about trends in consumer behavior.

Warehousing is an important aspect of data mining. Warehousing is when companies centralize their data into one database or program. With a data

warehouse, an organization may spin off segments of the data for specific users to analyze and use.

However, in other cases, analysts may start with the data they want and create a **data warehouse** based on those specs. Regardless of how businesses and other entities organize their data, they use it to support management's decision-making processes.

Classification of data mining systems

Data Mining is considered as an interdisciplinary field. It includes a set of various disciplines such as statistics, database systems, machine learning, visualization and information sciences.Classification of the data mining system helps users to understand the system and match their requirements with such systems.
Data mining systems can be categorized according to various criteria, as follows:

1. **Classification according to the application adapted:**
   This involves domain-specific application. For example, the data mining systems can be tailored accordingly for telecommunications, finance, stock markets, E-mails and so on.
2. **Classification according to the type of techniques utilized:**
   This technique involves the degree of user interaction or the technique of data analysis involved.For example, machine learning, visualization, pattern recognition, neural networks, database-oriented or data-warehouse oriented techniques.
3. **Classification according to the types of knowledge mined:**
   This is based on functionalities such as characterization, association, discrimination and correlation, prediction etc.
4. **Classification according to types of databases mined:**
   A database system can be classified as a ‘type of data’ or ‘use of data’ model or ‘application of data’.

Etymology

In the 1960s, statisticians and economists used terms like *data fishing* or *data dredging* to refer to what they considered the bad practice of analyzing data without a priori hypothesis. The term "data mining" was used in a similarly critical way by economist Michael Lovell in an article published in the *Review of Economic Studies* in 1983. Lovell indicates that the practice "masquerades under a variety of aliases, ranging from "experimentation" (positive) to "fishing" or "snooping" (negative).

The term *data mining* appeared around 1990 in the database community, generally with positive connotations. For a short time in 1980s, a phrase "database mining"™ , was used, but since it was trademarked by HNC, a San Diego-based company, to pitch their Database Mining Workstation; researchers consequently turned to *data mining*. Other terms used include *data archaeology*, *information harvesting*, *information discovery*, *knowledge extraction*, etc. Gregory Piatetsky-Shapiro coined the term "knowledge discovery in databases" for the first workshop on the same topic (KDD-1989) and this term became more popular in AI and machine learning community. However, the term data mining became more popular in the business and press communities.[14] Currently, the terms *data mining* and *knowledge discovery* are used interchangeably.

In the academic community, the major forums for research started in 1995 when the First International Conference on Data Mining and Knowledge Discovery (KDD-95) was started in Montreal under AAAI sponsorship. It was co-chaired by Usama Fayyad and Ramasamy Uthurusamy. A year later, in 1996, Usama Fayyad launched the journal by Kluwer called Data Mining and Knowledge Discovery as its founding editor-in-chief. Later he started the SIGKDD Newsletter SIGKDD Explorations.[15] The KDD International conference became the primary highest quality conference in data mining with an acceptance rate of research paper submissions below 18% . The journal *Data Mining and Knowledge Discovery* is the primary research journal of the field.

Background

The manual extraction of patterns from data has occurred for centuries. Early methods of identifying patterns in data include Bayes'

theorem (1700s) and regression analysis (1800s). The proliferation, ubiquity and increasing power of computer technology have dramatically increased data collection, storage, and manipulation ability. As data sets have grown in size and complexity, direct "hands-on" data analysis has increasingly been augmented with indirect, automated data processing, aided by other discoveries in computer science, such as neural networks, cluster analysis, genetic algorithms (1950s), decision trees and decision rules (1960s), and support vector machines (1990s). Data mining is the process of applying these methods with the intention of uncovering hidden patternsin large data sets. It bridges the gap from applied statistics and artificial intelligence (which usually provide the mathematical background) to database management by exploiting the way data is stored and indexed in databases to execute the actual learning and discovery algorithms more efficiently, allowing such methods to be applied to ever-larger data sets.

Example of Data Mining

Grocery stores are well-known users of data mining techniques. Many supermarkets offer free loyalty cards to customers that give them access to reduced prices not available to non-members. The cards make it easy for stores to track who is buying what, when they are buying it and at what price. After analyzing the data, stores can then use this data to offer customers coupons targeted to their buying habits and decide when to put items on sale or when to sell them at full price.

Data mining can be a cause for concern when a company uses only selected information, which is not representative of the overall sample group, to prove a certain hypothesis.

Process

The *knowledge discovery in databases (KDD) process* is commonly defined with the stages:

1. Selection
2. Pre-processing
3. Transformation
4. *Data mining*
5. Interpretation/evaluation.

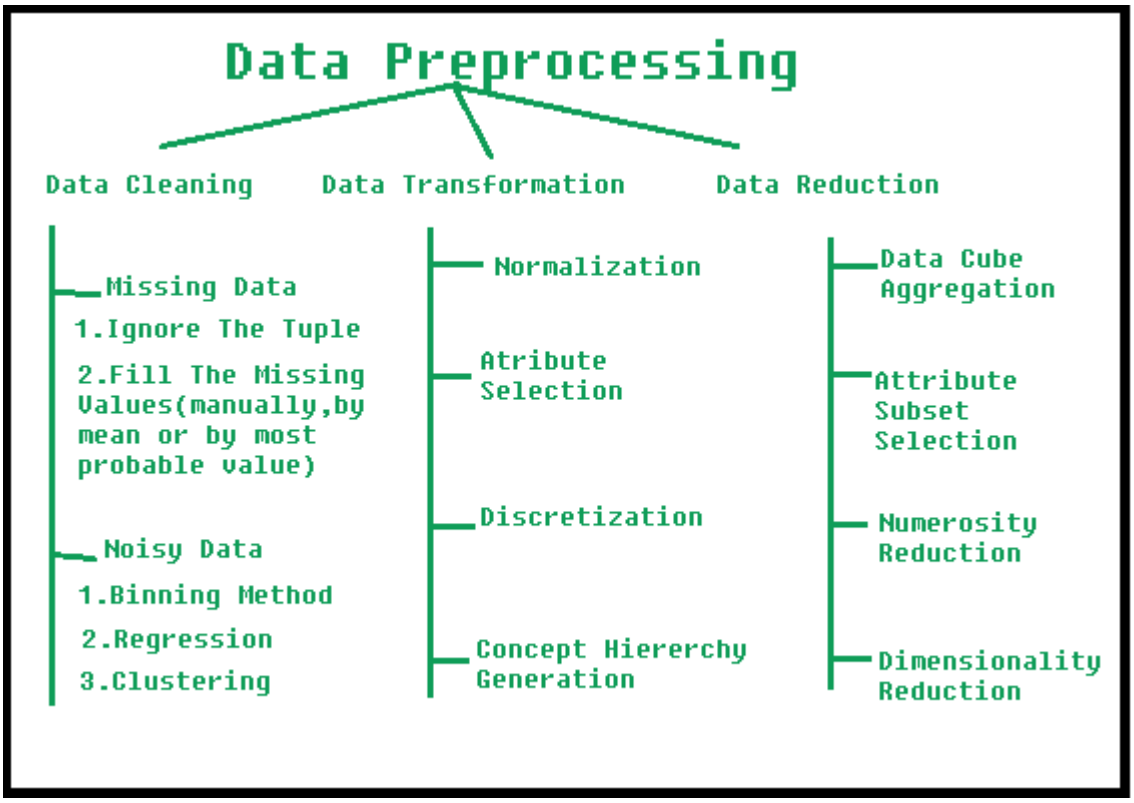
It exists, however, in many variations on this theme, such as the Cross-industry standard process for data mining (CRISP-DM) which defines six phases:

1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment

or a simplified process such as (1) Pre-processing, (2) Data Mining, and (3) Results Validation.

Polls conducted in 2002, 2004, 2007 and 2014 show that the CRISP-DM methodology is the leading methodology used by data miners.[17] The only other data mining standard named in these polls was SEMMA. However, 3–4 times as many people reported using CRISP-DM. Several teams of researchers have published reviews of data mining process models,[18] and Azevedo and Santos conducted a comparison of CRISP-DM and SEMMA in 2008.

Pre-processing



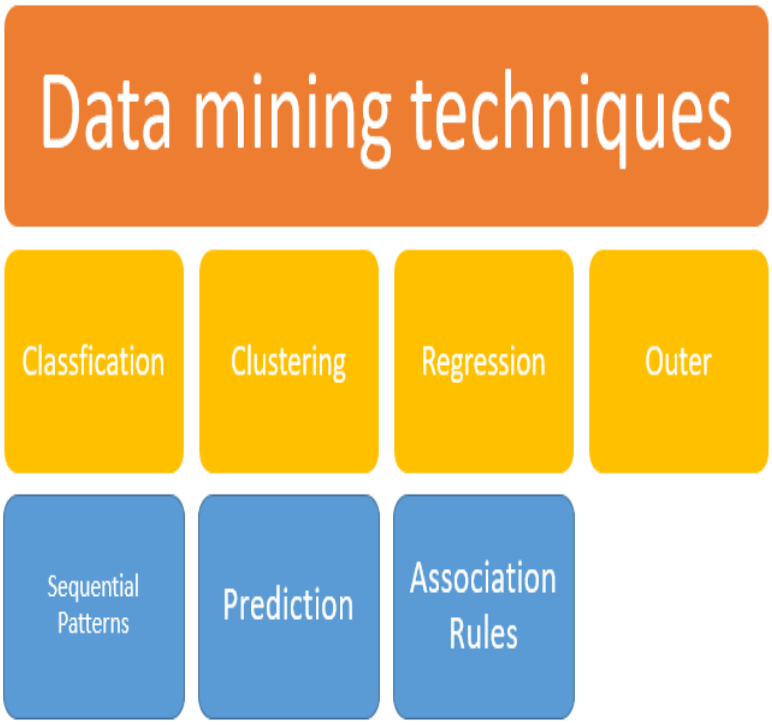
Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.

Before data mining algorithms can be used, a target data set must be assembled. As data mining can only uncover patterns actually present in the data, the target data set must be large enough to contain these patterns while remaining concise enough to be mined within an acceptable time limit. A common source for data is a data mart or data warehouse. Pre-processing is essential to analyze the multivariate data sets before data mining. The target set is then cleaned. Data cleaning removes the observations containing noise and those with missing data.

Steps Involved in Data Preprocessing:

1. **Data Cleaning:**

   The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.



- **(a). Missing Data:**
  This situation arises when some data is missing in the data. It can be handled in various ways.
  Some of them are:
  1. **Ignore the tuples:**
     This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.
  2. **Fill the Missing values:**
     There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.
- **(b). Noisy Data:**
  Noisy data is a meaningless data that can't be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :
  1. **Binning Method:**
     This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.
  2. **Regression:**
     Here data can be made smooth by fitting it to a regression function. The regression used may be linear (having one independent

variable) or multiple (having multiple independent variables).

3. **Clustering:**
   This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

2. Data Transformation:

This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

1. **Normalization:**
   It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0).
2. **Attribute Selection:**
   In this strategy, new attributes are constructed from the given set of attributes to help the mining process.
3. **Discretization:**
   This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.
4. **Concept Hierarchy Generation:**
   Here attributes are converted from level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

3. Data Reduction:

Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we use data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

The various steps to data reduction are:

1. **Data Cube Aggregation:**
   Aggregation operation is applied to data for the construction of the data cube.
2. **Attribute Subset Selection:**
   The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use level of significance and p- value of the attribute. the attribute having p-value greater than significance level can be discarded.
3. **Numerosity Reduction:**
   This enable to store the model of data instead of whole data, for

example: Regression Models.

4. **Dimensionality Reduction:**
   This reduce the size of data by encoding mechanisms.It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction. The two effective methods of dimensionality reduction are: Wavelet transforms and PCA (Principal Componenet Analysis).

Data mining

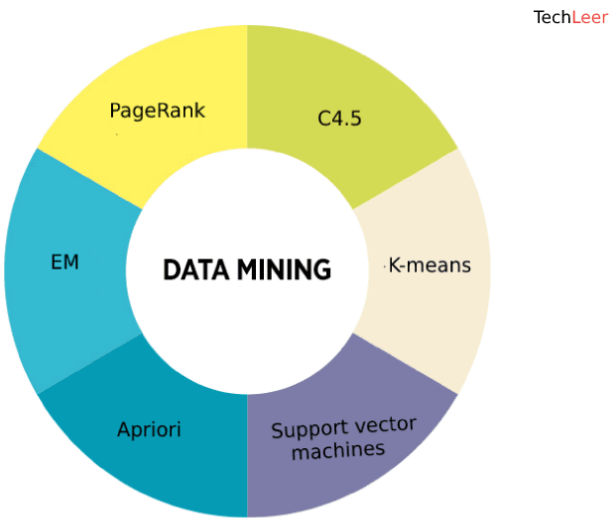
Data mining involves six common classes of tasks:

- Anomaly detection (outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.
- Association rule learning (dependency modeling) – Searches for relationships between variables. For example, a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.
- Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".
- Regression – attempts to find a function that models the data with the least error that is, for estimating the relationships among data or datasets.
- Summarization – providing a more compact representation of the data

set, including visualization and report generation.

Data mining algorithms

Data mining is known as an interdisciplinary subfield of computer science and basically is a computing process of discovering patterns in large data sets. It is considered as an essential process where intelligent methods are applied in order to extract data patterns.



Top Data Mining Algorithms

Given below is a list of Top Data Mining Algorithms:

1. C4.5:

C4.5 is an algorithm that is used to generate a classifier in the form of a decision tree and has been developed by **Ross Quinlan**. And in order to do the same, **C4.5** is given a set of data that represent things that have already been classified.

C4.5 that is often referred to as a statistical classifier is basically an extension of Quinlan's ID3 algorithm. The decision trees that are generated by C4.5 can be further used for classification. The C4.5 algorithm has also been described as "a landmark decision tree program that is probably the machine learning workhorse most widely used in practice to date" by the authors of the Weka machine learning software.

2. k-means:

k-means clustering that is also known as nearest centroid classifier or The Rocchio algorithm is a method of vector quantization, that is considerably popular for cluster analysis in data mining.

k-means is used to create k groups from a set of objects just so that the members of a group are more similar. It's a well known popular cluster analysis technique used for exploring a dataset.

3. Support vector machines:

When it comes to **machine learning**, support vector machines that are also known as support vector networks are basically supervised learning models that come with associated learning algorithms which then analyze data that are used for the analysis of regression and classification.

An **SVM** model is created that is a representation of the examples as points in space, that are further mapped so that the examples of the separate categories are then divided by a clear gap that is ought to be as wide as possible.

4. Apriori:

Apriori is an algorithm that is used for frequent itemset mining and association rule learning overall transactional databases. The algorithm is proceeded by the identification of the individual items that are frequent in the database and then extending them to larger itemsets as long as sufficiently those item sets appear often enough in the database. These

frequent itemsets that are determined by Apriori can be used for the determination of association rules which then highlight general trends.

5. EM (Expectation-Maximization):

An **expectation–maximization (EM) algorithm**, when it comes to statistics is an iterative method that is used to find maximum a posteriori(MAP) or maximum likelihood estimates of parameters in statistical models, that basically depends on unobserved latent variables.

6. PageRank (PR):

PageRank (PR) that was named after Larry Page who is one of the founders of Google is an algorithm that is used by Google Search to rank the websites in their search engine results. PageRank, that is the first algorithm that was used by the company is not the only algorithm that is being used by Google to order search engine results, but it is the best-known way of measuring the importance of website pages.

7. AdaBoost:

Adaptive Boosting or **AdaBoost**,that has been formulated by Yoav Freund and Robert Schapire is a machine learning meta-algorithm, that won the founders the 2003 Godel Prize for the same. The algorithm can be used in composition with many other types of learning algorithms in order to improve performance. AdaBoost is sensitive to noisy data as well as outliers.

8. kNN:

The **k-nearest neighbours algorithm (k-NN)** is a type of lazy learning or instance-based learning and is considered as a non-parametric method that is used for classification and regression.\In both the mentioned cases, the input consists of the k closest training examples in the feature space and the output depends on whether the algorithm is being used for classification or regression. This **kNN** Algorithm is considered and is also among the simplest of all **machine learning algorithms.**

9. Naive Bayes:

When it comes to machine learning**, Naive Bayes classifiers** that are considered to be highly scalable are known to be a family of simple probabilistic classifiers that are based on the application of Bayes' theorem with the help of strong independent assumptions between the features.

10. CART:

CART is an algorithm that basically stands for classification and regression trees. It is a decision tree learning technique that either outputs classification or regression trees and similarly like C4.5, CART is also a classifier.

Many of the reasons that a user would use C4.5 for also apply to that of CART, since both of them are decision tree learning techniques and features like ease of interpretation and explanation are also applied to CART as well.

Results validation

Data mining can unintentionally be misused, and can then produce results that appear to be significant; but which do not actually predict future behavior and cannot be reproduced on a new sample of data and bear little use. Often this results from investigating too many hypotheses and not performing proper statistical hypothesis testing. A simple version of this problem in machine learning is known as overfitting, but the same problem can arise at different phases of the process and thus a train/test split—when applicable at all—may not be sufficient to prevent this from happening.[20]

The final step of knowledge discovery from data is to verify that the patterns produced by the data mining algorithms occur in the wider data set. Not all patterns found by data mining algorithms are necessarily valid. It is common for data mining algorithms to find patterns in the training set which are not present in the general data set. This is called overfitting. To overcome this, the evaluation uses a test set of data on which the data mining algorithm was not trained. The learned patterns are applied to this test set, and the resulting output is compared to the desired output. For example, a data mining algorithm trying to distinguish "spam" from "legitimate" emails would be trained on a training set of sample e-mails. Once trained, the learned patterns would be applied to the test set of e-mails on which it had *not* been trained. The accuracy of the patterns can then be measured from how many e-mails they correctly classify. Several statistical methods may be used to evaluate the algorithm, such as ROC curves.

If the learned patterns do not meet the desired standards, subsequently it is necessary to re-evaluate and change the pre-processing and data mining steps. If the learned patterns do meet the desired standards, then the final step is to interpret the learned patterns and turn them into knowledge.

Conclusion

The ultimate goal of data mining is the prediction of human behavior, and is by far the most common business application; however this can easily be modeled to meet the objective of detection and deterrence of criminals. These and many more application have demonstrated that, rather than requiring a human to attempt to deal with hundreds of descriptive attributes, data mining allows the automatic analysis of databases and the recognition of important trends and behavioral patterns. Sophisticated data mining and artificial intelligence tools are now available to the law enforcement communities. These tools are extremely powerful, fast, and relatively easy to use. Data mining supports enhanced decision making and analysis, and is a powerful tool that can be used to address the large volume of crime information currently facing all agencies. Data mining tools increase not only the speed of analysis, but the depth of its approach. By mining the essential nuggets of information, crime analysts are able to fully explore existing datasets and identify actionable patterns and trends. The examples listed in this paper represent only a small fraction of the potential for this approach in the public safety and intelligence arena.

References

- https://link.springer.com/book/10.1007%2F978-0-387-09823-4
- https://www.investopedia.com/terms/d/datamining.asp
- https://www.computersciencedegreehub.com/lists/5-uses-for-data-mining/
- https://bigdata-madesimple.com/14-useful-applications-of-data-mining/
- https://www.geeksforgeeks.org/classification-of-data-mining-systems/
- https://www.sas.com/en_us/insights/analytics/data-mining.html
- https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/

- HTTPS://WWW.TECHLEER.COM/ARTICLES/438-A-LIST-OF-TOP-DATA-MINING-ALGORITHMS/