2

Introduction to Data Mining

Chapter Objectives

- ✓ To learn about the concepts of data mining.
- ✓ To understand the need for, and the applications of data mining
- ✓ To differentiate between data mining and machine learning
- ✓ To understand the process of data mining.
- ✓ To understand the difference between data mining and machine learning.

2.1 Introduction to Data Mining

In the age of information, an enormous amount of data is available in different industries and organizations. The availability of this massive data is of no use unless it is transformed into valuable information. Otherwise, we are sinking in data, but starving for knowledge. The solution to this problem is data mining which is the extraction of useful information from the huge amount of data that is available.

Data mining is defined as follows:

'Data mining is a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful and understandable patterns in large databases. The patterns must be actionable so they may be used in an enterprise's decision making.'

From this definition, the important take aways are:

- Data mining is a process of automated discovery of previously unknown patterns in large volumes of data.
- This large volume of data is usually the historical data of an organization known as the data warehouse.
- Data mining deals with large volumes of data, in Gigabytes or Terabytes of data and sometimes as much as Zetabytes of data (in case of big data).

- Patterns must be valid, novel, useful and understandable.
- Data mining allows businesses to determine historical patterns to predict future behaviour.
- Although data mining is possible with smaller amounts of data, the bigger the data the better the accuracy in prediction.
- There is considerable hype about data mining at present, and the Gartner Group has listed data mining as one of the top ten technologies to watch.

2.2 Need of Data Mining

Data mining is a recent buzz word in the field of Computer Science. It is a computing process that uses intelligent mathematical algorithms to extract the relevant data and computes the probability of future actions. It is also known as Knowledge Discovery in Data (KDD).

Although data mining has existed for 50 years, it has become a very important technology only recently due to the Internet boom. Database systems are being used since the 1960s in the Western countries (perhaps, since the 1980s in India). These systems have generated mountains of data. There is a broad agreement among all sources that the size of the digital universe will double every two years at least. Therefore, there will be a 50-fold growth from 2010 to 2020.

Every minute, huge data is generated over Internet as depicted in Figure 2.1. The analysis of this huge amount of data is an important task and it is performed with the help of data mining. This analysis of data has significant potential to improve business outcomes for an organization.

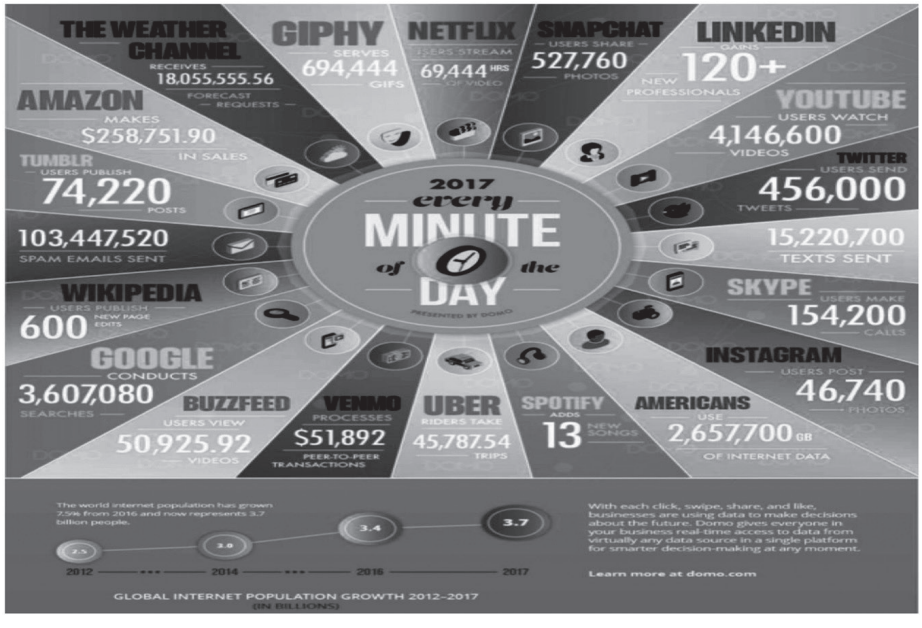

Figure 2.1 Per minute generation of data over the Internet according to a 2017 report

[*Credit*: http://www.iflscience.com/technology/how-much-data-does-the-world-generate-every-minute/]

Some facts about this rapid growth of data are as follows.

- 48 hours of new videos are uploaded by YouTube users every minute of the day.
- 571 new websites are created every minute of the day.
- 90% of world's data has been produced in the last two years.
- The production of data will be 44 times greater in 2020 than it was in 2010.
- Brands and organizations on Facebook receive 34,722 'Likes' every minute of the day.
- 100 terabytes of data is uploaded daily to Facebook.
- In early 2012, there were more than 465 million accounts and 175 million tweets were shared by people on Twitter every day, according to Twitter's own research.
- Every month, 30 billion pieces of content are shared by people on Facebook.
- By late 2011, approximately 1.8 zetta byte of data had been created in that year alone, according the report published by IDC Digital Universe.

The analysis of such huge amounts of data is a challenging task that requires suitable analytics to sieve it for the bits that will be useful for business purposes. From the start, data mining was designed to find data patterns from data warehouses and data mining algorithms are tuned so that they can handle large volumes of data. The boom in big data volumes has led to great interest in data mining all over the world.

The other allied reasons for popularity of data mining are as follows.

- Growth in generation and storage of corporate data
- Need for sophisticated decision making
- Evolution of technology
- Availability of much cheaper storage, easier data collection and better database management for data analysis and understanding
- Point of sale terminals and bar codes on many products, railway bookings, educational institutions, mobile phones, electronic gadgets, e-commerce, etc., all generate data.
- Great volumes of data generated with the recent prevalence of Internet banking, ATMs, credit and debit cards; medical data, hospitals; automatic toll collection on toll roads, growing air travel; passports, visas, etc.
- Decline in the costs of hard drives
- Growth in worldwide disk capacities

Thus, the need for analyzing and synthesizing information is growing in the fiercely competitive business environment of today.

2.3 What Can Data Mining Do and Not Do?

Data mining is a powerful tool that helps to determine the relationships and patterns within data. However, it does not work by itself and does not eliminate the requirement for understanding data, analytical methods and to know business. Data mining extracts hidden information from the data, but it is not able to assess the value of the information.

One should know the important patterns and relationships to work with data over time. In addition to discovering new patterns, data mining can also highlight other empirical observations that are not instantly visible through simple observation.

It is important to note that the relationships or patterns predicted through data mining are not necessarily causes for an action or behavior. For example, data mining may help in determining that males with income between Rs 20,000 and Rs 75,000 who contribute to certain journals or magazines, may be expected to purchase such-and-such a product. This information can be helpful in developing a marketing strategy. However, it is not necessary that the population identified through data mining will purchase the product simply because they belong to this category.

2.4 Data Mining Applications

The applications of data mining exist in almost every field. Some of the important applications of data mining are in finance, telecom, insurance and retail sectors include loan/credit card approval, fraud detection, market segmentation, trend analysis, better marketing, market basket analysis, customer churn and web site design and promotion. These are discussed below.

Loan/Credit card approvals

Banks are able to assess the credit worthiness of their customers by mining a customer's historical records of business transactions. So, credit agencies and banks collect a lot of customer' behavioural data from many sources. This information is used to predict the chances of a customer paying back a loan.

Market segmentation

A huge amount of data about customers is available from purchase records. This data is very useful to segment customers on the basis of their purchase history. Let us suppose a mega store sells multiple items ranging from grocery to books, electronics, clothing et al. That store now plans to launch a sale on books. Instead of sending SMS texts to all the customers it is logical and more efficient to send the SMS only to those customers who have demonstrated interest in buying books, i.e., those who had earlier purchased books from the store. In this case, the segmentation of customers based on their historical purchases will help to send the message to those who may find it relevant. It will also give a list of people who are prospects for the product.

Fraud detection

Fraud detection is a very challenging task because it's difficult to define the characteristics for detection of fraud. Fraud detection can be performed by analyzing the patterns or relationships that deviate from an expected norm. With the help of data mining, we can mine the data of an organization to know about outliers as they may be possible locations for fraud.

Better marketing

Usually all online sale web portals provide recommendations to their users based on their previous purchase choices, and purchases made by customers of similar profile. Such recommendations are generated through data mining and help to achieve more sales with better marketing of their products. For example, amazon.com uses associations and provides recommendations to customers

on the basis of past purchases and what other customers are purchasing. To take another example, a shoe store can use data mining to identify the right shoes to stock in the right store on the basis of shoe sizes of the customers in the region.

Trend analysis

In a large company, not all trends are always visible to the management. It is then useful to use data mining software that will help identify trends. Trends may be long term trends, cyclic trends or seasonal trends.

Market basket analysis

Market basket analysis is useful in designing store layouts or in deciding which items to put on sale. It aims to find what the customers buy and what they buy together.

Customer churn

If an organization knows its customers better then it can retain them longer. In businesses like telecommunications, companies very hard to retain their good customers and to perhaps persuade good customers of their competitors to switch to them. In such an environment, businesses want to rate customers (whether they are worthwhile to be retained), why customers switch over and what makes customers loyal. With the help of this information some businesses may wish to get rid of customers that cost more than they are worth, e.g., credit card holders that don't use the card; bank customers with very small amounts of money in their accounts.

Website design

A web site is effective only if the visitors easily find what they are looking for. Data mining can help discover affinity of visitors to pages and the site layout may be modified based on data generated from their web clicks.

Corporate analysis and risk management

Data mining can be used to perform cash flow analysis to plan finance and estimate assets. The analysis of data can be performed by comparing and summarizing the spending and planning of resources. It helps to analyze market directions and monitor competitors to analyze the competition.

Thus, one can conclude that the applications of data mining are numerous and it is worthwhile for an organization to invest in data mining for generating better business revenue. This is the reason that many organizations are investing in this field and there is a huge demand for data mining experts in the world.

2.5 Data Mining Process

Data mining process consists of six phases, as illustrated in Figure 2.2.

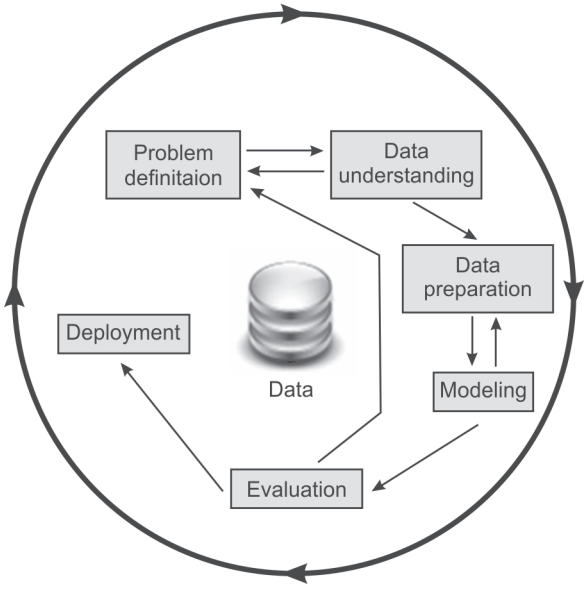

Figure 2.2 Data mining process

A brief description about these phases is as follows.

Problem definition phase

The main focus of the first phase of a data mining process is to understand the requirements and objectives of such a project. Once the project has been specified, it can be formulated as a data mining problem. After this, a preliminary implementation plan can be developed.

Let us consider a business problem such as 'How can I sell more of my product to customers?' This business problem can be translated into a data mining problem such as 'Which customers are most likely to buy the product?' A model that predicts the frequent customers of a product must be built on the previous records of customers' data. Before building the model, the data must be assembled that consists of relationships between customers who have purchased the product and customers who have not purchased the product. The attributes of customers might include age, years of residence, number of children, owners/renters, and so on.

Data understanding phase

The next phase of the data mining process starts with the data collection. In this phase, data is collected from the available sources and in order to make data collection proper, some important activities such as data loading and data integration are performed. After this, the data is analyzed closely to determine whether the data will address the business problem or not. Therefore, additional data can be added or removed to solve the problem effectively. At this stage missing data is also identified. For example, if we require the AGE attribute for a record then column DATE_OF_BIRTH can be changed to AGE. We can also consider another example in which average income

can be inserted if value of column INCOME is null. Moreover, new computed attributes can be added in the data in order to obtain better focused information. For example, a new attribute such as 'Number of Times Amount Purchased Exceeds Rs. 500 in a 12 month time period.' can be created instead of using the purchase amount.

Data preparation phase

This phase generally consumes about 90% of the time of a project. Once available data sources are identified, they need to be selected, cleaned, constructed and formatted into the desired form for further processing.

Modeling phase

In this phase, different data mining algorithms are applied to build models. Appropriate data mining algorithms are selected and applied on given data to achieve the objectives of proposed solution.

Evaluation phase

In the evaluation phase, the model results are evaluated to determine whether it satisfies the originally stated business goal or not. For this the given data is divided into training and testing datasets. The models are trained on training data and tested on testing data. If the accuracy of models on testing data is not adequate then one goes back to the previous phases to fine tune those areas that may be the reasons for low accuracy. Having achieved a satisfactory level of accuracy, the process shifts to the deployment phase.

Deployment phase

In the deployment phase, insights and valuable information derived from data need to be presented in such a way that stakeholders can use it when they want to. On the basis of requirements of the project, the deployment phase can be simple (just creating a report) or complex (requiring further iterative data mining processing). In this phase, Dashboards or Graphical User Interfaces are built to solve all the requirements of stakeholders.

2.6 Data Mining Techniques

Data mining can be classified into four major techniques as given below.

- Predictive modeling
- Database segmentation
- Link analysis
- Deviation detection

To briefly discuss each technique:

2.6.1 Predictive modeling

Predictive modeling is based on predicting the outcome of an event. It is designed on a pattern similar to the human learning experience in using observations to form a model of the important characteristics of some task. It is developed using a supervised learning approach, where we have some labeled data and we use this data to predict the outcome of unknown instances. It can be of two types, i.e., classification or regression as discussed in Chapter 1.

Some of the applications of predictive modeling are: predicting the outcome of an event, predicting the sale price of a property, predicting placement of students, predicting the score of any team during a cricket match and so on.

2.6.2 Database segmentation

Database segmentation is based on the concept of clustering of data and it falls under unsupervised learning, where data is not labeled. This data is segmented into groups or clusters based on its features or attributes. Segmentation is creating a group of similar records that share a number of properties.

Applications of database segmentation include customer segmentation, customer churn, direct marketing, and cross-selling.

2.6.3 Link analysis

Link analysis aims to establish links, called associations, between the individual record, or sets of records, in a database. There are three specialisations of link analysis.

- Associations discovery
- Sequential pattern discovery
- Similar time sequence discovery

Associations discovery locates items that imply the presence of other items in the same event. There are association rules which are used to define association. For example, 'when a customer rents property for more than two years and is more than 25 years old, in 40% of cases, the customer will buy a property. This association happens in 35% of all customers who rent properties'.

Sequential pattern discovery finds patterns between events such that the presence of one set of items is followed by another set of items in a database of events over a period of time. For example, this approach can be used to understand long-term customer buying behavior.

Time sequence discovery is used to determine whether links exist between two sets of data that are time-dependent. For example, within three months of buying property, new home owners will purchase goods such as cookers, freezers, and washing machines.

Applications of link analysis include market basket analysis, recommendation system, direct marketing, and stock price movement.

2.6.4 Deviation detection

Deviation detection is a relatively new technique in terms of commercially available data mining tools. It is based on identifying the outliers in the database, which indicates deviation from some

previously known expectation and norm. This operation can be performed using statistics and visualization techniques.

Applications of deviation detection include fraud detection in the use of credit cards and insurance claims, quality control, and defects tracing.

2.7 Difference between Data Mining and Machine Learning

Data mining refers to extracting knowledge from a large amount of data, and it is a process to discover various types of patterns that are inherent in the data and which are accurate, new and useful. It is an iterative process and is used to uncover previously unknown trends and patterns in vast amount of data in order to support decision making.

Data mining is the subset of business analytics; it is similar to experimental research. The origins of data mining are databases and statistics. Two components are required to implement data mining techniques: the first is the database and the second is machine learning.

Data mining as a process includes data understanding, data preparation and data modeling; while machine learning takes the processed data as input and performs predictions by applying algorithms. Thus, data mining requires involvement of human beings to clean and prepare the data and to understand the patterns. While in machine learning human effort is involved only to define an algorithm, after which the algorithm takes over operations. Tabular comparison of data mining and machine learning is given in Table 2.1.

Table 2.1 Tabular comparison of data mining and machine learning

Basic for comparison	*Data mining*	*Machine learning*
Meaning	It involves extracting useful knowledge from a large amount of data.	It introduces new algorithm from data as well as past experience.
History	Introduced in 1930 it was initially called knowledge discovery in databases.	It was introduced in 1959.
Responsibility	Data mining is used to examine patterns in existing data. This can then be used to set rules.	Machine learning teaches the computer to learn and understand the given rules.
Nature	It involves human involvement and intervention.	It is automated, once designed it is self-implementing and no or very little human effort is required.

In conclusion the analysis of huge amounts of data generated over Internet or by traditional database management systems is an important task. This analysis of data has the huge potential to improve business returns. Data mining and machine learning play vital roles in understanding and analyzing data. Thus, if data is considered as oil, data mining and machine learning are equivalent to modern day oil refineries that make this data more useful and insightful.

Remind Me

- Data mining is a process of automated discovery of previously unknown patterns in large volumes of data.
- Data mining process consists of six phases, namely problem definition, data understanding, data preparation, modeling, evaluation and deployment.
- The important applications of data mining are in finance, telecom, insurance and retail sectors; and include loan/credit card approval, fraud detection, market segmentation, trend analysis, focused marketing, market basket analysis, customer churn and web-site design.
- There are four main operations associated with data mining techniques, namely, predictive modeling, database segmentation, link analysis and deviation Detection.

Point Me (Books)

- Han, Jiawei, Micheline Kamber, and Jian Pei. 2011. *Data Mining: Concepts and Techniques*, 3rd ed. Amsterdam: Elsevier.
- Gupta, G. K. 2014. *Introduction to Data Mining with Case Studies*. Delhi: PHI Learning Pvt. Ltd.
- Mitchell, Tom M. 2017. *Machine Learning*. Chennai: McGraw Hill Education.
- Witten, Ian H., Eibe Frank, Mark A. Hall, and Christopher Pal. 2016. *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Burlington: Morgan Kaufmann.

Connect Me (Internet Resources)

- https://www.kdnuggets.com/data_mining_course/x1-intro-to-data-mining-notes.html
- https://www.cse.iitb.ac.in/infolab/Data/Talks/datamining-intro-IEP.ppt
- https://tutorials.datasciencedojo.com/data-mining-fundamentals-part-1-3/

Test Me

1. What is the full form of KDD .....
2. Which learning approach is used by Database Segmentation?
   (a) Supervised Learning (b) Unsupervised Learning
3. Links between the individual record, or sets of records in a database is called .....
4. What are the two types of predictive modeling?
   (a) ..... (b) .....
5. Deviation detection can be performed by using ..... and ..... techniques.
6. Predictive Modeling is developed using a supervised learning approach.
   (a) True (b) False
7. Data mining is .....
   (a) The process of automated discovery of previously unknown patterns in large volumes of data.
   (b) The stage of selecting the right data for a KDD process
   (c) A subject-oriented integrated time variant non-volatile collection of data in support of management
   (d) None of these
8. Value prediction uses the traditional statistical techniques of ..... and ..... .
9. Classification falls under which technique of data mining?
   (a) Predictive modeling (b) Database segmentation
   (c) Link analysis (d) Deviation detection.

10. Regression falls under which technique of data mining?
    (a) Predictive modeling (b) Database segmentation
    (c) Link analysis (d) Deviation detection.
11. Clustering falls under which technique of data mining?
    (a) Predictive modeling (b) Database segmentation
    (c) Link analysis (d) Deviation detection.
12. Visualization is core part for which of following data mining technique?
    (a) Predictive modeling (b) Database segmentation
    (c) Link analysis (d) Deviation detection.
13. Association mining falls under which technique of data mining?
    (a) Predictive modeling (b) Database segmentation
    (c) Link analysis (d) Deviation detection.

Answer Keys:

1. Knowledge Discovery in Database; 2. (b); 3. Associations;
4. (a) Classification (b) Value prediction 5. Statistics, Visualization
6. (a); 7. (a); 8. Linear regression, Nonlinear regression;
9. (a); 10. (a); 11. (b); 12. (d);
13. (c).