AN OVERVIEW OF DATA MINING

Abstract

Now-a-days, data is very important and it is the basic and important thing before proceeding with any task of data mining. Data mining is a field of combination of statistics and computer science that is used to find useful patterns in the information. This chapter includes different types of data, requirements for data mining, and types of data mining. Finally, this chapter also includes steps involved in each type of data mining, process involved in data mining and different data mining techniques.

Keywords: Data mining, Computer science, Quantitative Data, Data storage

Authors

Someswari Perla
Department of CSE (AI&ML)
GMRIT
Rajam, Andhra Pradesh
India
someswari.p@gmrit.edu.in

A Vineela
Department of CSE
GMRIT
Rajam, Andhra Pradesh
India
alladavineela4@gmail.com

I. DATA

It is a collection of raw facts, figures, or information that can be in different forms, including images, numbers, text, video, audio or any other representation of facts used for analysis to discover patterns, relationships, and knowledge.

Types of Data

Following figure shows different types of data:

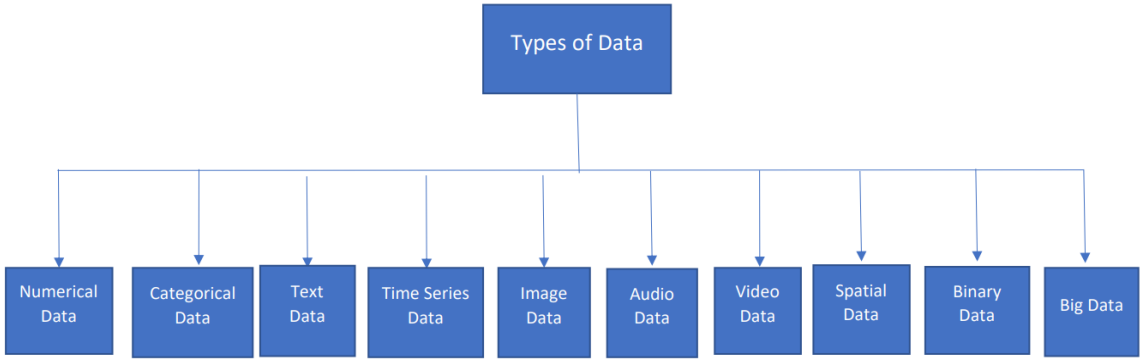

Figure 1

1. **Numerical Data (Quantitative Data):** Numerical data consists of quantitative values represented as numbers. It can be further categorized into two subtypes:

   - **Discrete Data:** Consists of whole numbers and represents countable values. Examples are number of subjects taught by a teacher in a class or number of things purchased by a customer.
   - **Continuous Data:** Consists of real numbers and represents measurable values. For example, temperature, height, weight, or time.

2. **Categorical Data (Qualitative Data):** Categorical data represents distinct categories or labels and cannot be measured numerically. It can be further divided into two subtypes:

   - **Nominal Data:** Represents categories with no meaningful order or ranking. For example, gender, colors, or types of animals.
   - **Ordinal Data:** Represents categories with a meaningful order or ranking. For example, financial status of a person, satisfaction levels of a customer.

3. **Text Data:** Text data consists of words, sentences, paragraphs, or any textual information. It is commonly found in documents, emails, social media posts, and web pages.

4. **Time Series Data:** Time series data is collected at regular intervals over time. It is used to study trends, patterns, and changes over specific periods. Examples include weather data, stock price.

5. **Image Data:** It represents visual information captured in pictures or graphics. It is widely used in fields like computer vision, medical imaging, and satellite imagery analysis.

6. **Audio Data:** Audio data comprises sound recordings, such as music files, voice recordings, or speech samples.

7. **Video Data:** Video data consists of a sequence of images and audio, typically capturing moving scenes and actions.

8. **Spatial Data:** Spatial data represents geographic information and is associated with specific locations on the Earth's surface.

9. **Binary Data:** Binary data represents information in a binary format, using only two possible values, typically 0 and 1. It is commonly used in computer systems and digital communication.

10. **Big Data:** It refers to very large datasets that cannot be easily process using conventional data processing techniques. It often involved in high volume, velocity, and variety of data.

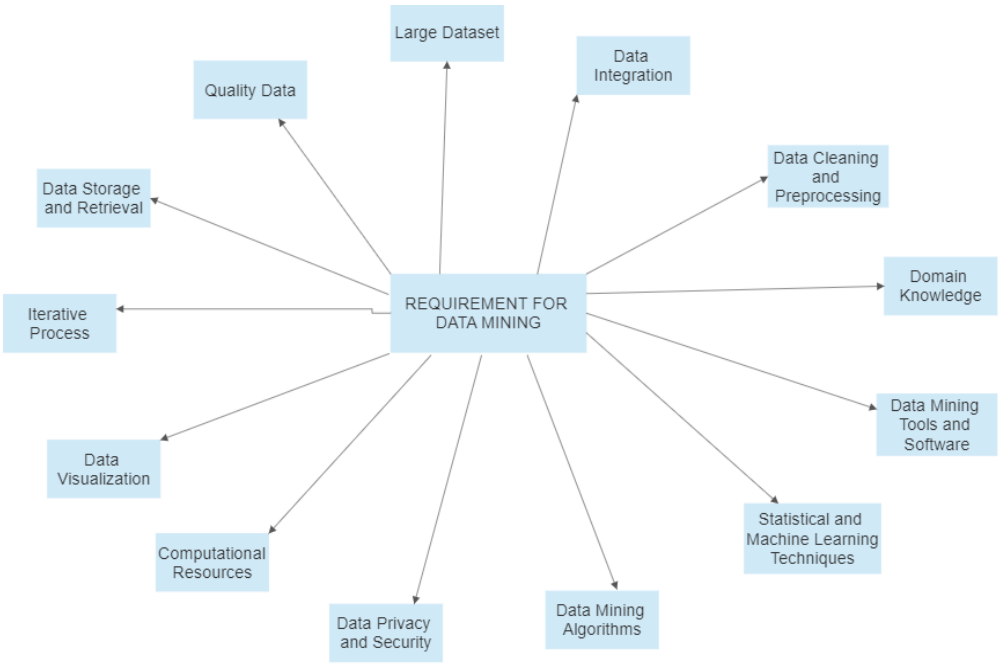

Figure 2

Above figure shows Requirements for Data Mining:

II. REQUIREMENTS FOR DATA MINING

1. **Quality Data:** Data mining heavily depends on quality of data. Data used for analysis should be accurate, reliable, and relevant to the business problem or research question at hand. Poor-quality data leads to inaccurate or misleading results.

2. **Large Dataset:** Data mining is particularly effective when working with large datasets. A substantial volume of data increases the likelihood of finding meaningful patterns and trends within the data.

3. **Data Integration:** Data mining often involves combining data from various sources. It is essential to integrate the data properly to ensure that the analysis is comprehensive and unbiased.

4. **Data Cleaning and Preprocessing:** Raw data may contain noise, inconsistencies, missing values, and outliers. Data cleaning and data preprocessing steps are crucial to remove these issues and prepare the data for analysis.

5. **Domain Knowledge:** Subject matter expertise or domain knowledge is essential in data mining. Understanding the data and its context helps in formulating relevant questions, selecting appropriate algorithms, and interpreting the results correctly.

6. **Data Mining Tools and Software:** There are various data mining tools and software available that facilitate the analysis process. These tools often include algorithms for data mining, visualization capabilities, and data manipulation functionalities.

7. **Machine Learning and Statistical Techniques:** Data mining involves applying various machine learning and statistical techniques to analyze data. Knowledge of these methods and when to use them is crucial for successful data mining.

8. **Data Mining Algorithms:** Understanding and selecting the appropriate data mining algorithms (e.g., decision trees, clustering, regression, association rules) for the specific task at hand is crucial for effective analysis.

9. **Security and Data Privacy:** Ensuring security and privacy of sensitive data is of utmost importance in data mining. Encryption and access controls are some of the methods used to protect data during analysis.

10. **Computational Resources:** Data mining can be computationally demanding, especially with big datasets and difficult algorithms. Sufficient computational resources, such as processing power and memory, are necessary to carry out the analysis efficiently.

11. **Data Visualization:** Presenting the results of data mining in a visually attractive and understandable manner is essential for efficient communication of insights to stakeholders.

12. **Iterative Process:** Data mining is often an iterative process, where analysts may need to refine their approaches, adjust parameters, and try different algorithms to get meaningful results.

13. **Data Storage and Retrieval:** An efficient and scalable data storage system is necessary to handle large datasets. Quick retrieval of relevant data is essential for smooth data mining operations.

The specific requirements for data mining can vary depending on kind of data, objectives of the analysis, tools and techniques used.

III. DATA MINING

It is the method of finding valuable relationships, patterns, and insights from huge and complex datasets. It involved in using various techniques, algorithms, statistical methods to dig out meaningful information from the data, make predictions about future trends or behaviors, and uncover hidden patterns. Data mining aims to turn raw data into actionable knowledge, providing valuable insights for decision-making, problem-solving, and business optimization.

Types of Data Mining

Data mining is of 2 types and is shown in figure:

1. Predictive Data Mining
2. Descriptive Data Mining

Figure 3

1. **Predictive Data Mining:** It is also known as predictive analytics where it is a subset of data mining. It focuses on using historical information to make prediction about future outcomes or events. It involves in application of various statistical and machine learning techniques to identify patterns and trends , analyze historical data, and build predictive models that can be used to predict future events. Here's a detailed overview of predictive data mining:

    - **Data Collection:** The first step in predictive data mining is to collect relevant historical data from a variety of sources. This data can come from databases, data warehouses, spreadsheet, web sources, or any other data repositories.
    - **Data Preprocessing and Cleaning:** Once data is collected, it requires be cleaning and preprocessing. It involve in removing duplicate records, handling misplaced values, dealing with outliers, transforming data into an appropriate format for analysis.
    - **Exploratory Data Analysis (EDA):** Before building predictive models, analysts often perform EDA to gain insights into data. EDA involves using various visualization and statistical techniques to understand the distribution of data, identify correlations, and detect any interesting patterns that might be useful for prediction.
    - **Feature Selection:** In predictive data mining, not all features (variables) in the dataset may be relevant for making predictions. Feature selection involves identifying the most important and informative features that will be used as input to the predictive models. This process helps to reduce noise and improve the model's efficiency and accuracy.
    - **Model Selection**: The next step is to choose the appropriate predictive modeling techniques based on nature of problem and data. Usual predictive modeling techniques include decision trees, linear regression, support vector machines, random forests, and neural networks.
    - **Model Training:** With selected predictive model, historical data is divided into 2 sets namely training and testing sets. Training set is used to train the model by giving it with known input-output pairs, allows it to learn underlying relationships and patterns within the data.
    - **Model Evaluation:** After training the model, testing set is used to evaluate it. Performance of the model is assessed by comparing its predictions against the actual outcomes in the testing data. Different metrics like precision, accuracy, F1 score, recall and ROC curves are used to evaluate the performance of model.
    - **Model Tuning:** If the model's performance is not satisfactory, it may be necessary to fine-tune the model by adjusting its parameters or trying different algorithms. This iterative process continues until a satisfactory predictive model is achieved.
    - **Deployment and Prediction:** Once a reliable predictive model is obtained, it can be deployed to make prediction on unseen and new data. When new data becomes available, the model uses the selected features to generate predictions about future events or outcomes.
    - **Monitoring and Maintenance:** Predictive models require monitoring and maintenance to ensure their accuracy and relevance over time. As new data is collected, the model may need to be retrained or updated to reflect changing patterns or business conditions.

Predictive data mining has numerous applications across various industries, such as finance, marketing, healthcare, manufacturing, and more. It helps organizations make data-driven decisions, identify potential risks and opportunities, optimize processes, and gain a competitive advantage in the market.

2. **Descriptive Data Mining:** It is the one that involves in summarizing and exploring historical data to obtain insights and understand trends or patterns within the data. It focuses on providing a comprehensive. Understandable summary of the data, enabling analysts and stakeholders to make informed decisions and identify meaningful patterns. Steps involved in descriptive data mining:

- **Data Collection:** Its very first step is to collect relevant historical data from different sources, like data warehouses, databases, spreadsheets, logs and also from other repositories of data. The data may be structured or unstructured.
- **Cleaning and Preprocessing of Data:** Once the data is gathered, it need to be cleaned and preprocessed to handle different issues such as outliers, missing values, and duplicates. Data cleaning ensures that analysis is based on reliable and accurate information.
- **Data Exploration and Visualization:** Data exploration is a critical phase in descriptive data mining. Analysts use various techniques to understand the structure and characteristics of the dataset. This involves summarizing the data using statistical measures (like median, mean and standard deviation), visualizing the data through charts and graphs (e.g., bar charts, histograms, scatter plots), and identifying any initial patterns or trends.
- **Summary Statistics:** Descriptive data mining involves calculating summary statistics to provide a brief overview of distribution of data and central tendencies. Various common summary statistics measures include median, mean, mode, standard deviation, minimum, maximum, and percentiles.
- **Data Clustering:** It is a method used to group similar data points together based on certain features or characteristics. It helps identify natural patterns or segments within the data. Various clustering algorithms, such as hierarchical and K-means clustering can be employed.
- **Data Segmentation:** Data segmentation involves dividing the dataset into meaningful segments or subsets based on specific criteria. This process allows analysts to focus on specific groups or categories of interest and gain deeper insights into each segment.
- **Association Rule Mining:** Another technique used in descriptive data mining is association rule mining. It identifies interesting relationships or associations between different variables in the data. For example, it might reveal that customers who purchase Product A are likely to buy Product B as well.
- **Pattern Recognition:** Descriptive data mining involves recognizing and understanding recurring patterns within the data. This may include periodic trends, seasonal patterns, cyclic behaviors, or sudden shifts.
- **Data Summarization:** Summarization techniques like data aggregation, dimensionality reduction are employed to provide a concise and informative representation of the data, especially when dealing with large datasets. Dimensionality reduction method like Principal Component Analysis (PCA) can be useful in reducing the number of features while preserving essential information.

- **Insights and Reporting:** The final step in descriptive data mining is to interpret results and communicate the insights to relevant stakeholders. Clear and understandable reports and visualizations are generated to convey the findings efficiently. Visualization tools like Tableau or matplotlib in Python can be used for creating insightful charts and graphs.

Descriptive data mining is widely used in various fields, including business intelligence, marketing analytics, healthcare, and scientific research. It helps organizations understand their data better, discover hidden patterns, identify opportunities, and make data-driven decisions based on historical trends and patterns. By exploring and summarizing historical data, descriptive data mining provides valuable insights that can serve as the foundation for further analysis, such as predictive and prescriptive data mining.

IV. DATA MINING AN ITS STEPS

The process of data mining is as shown in figure:

Figure 4

1. **Problem Definition:** In this initial step, data mining process starts by understanding the business problem and defining the objectives of the data mining project. It is crucial to clearly state the goals and requirements to ensure that analysis is aligned with the business needs. The problem definition helps in selecting appropriate data mining techniques and evaluating the success of analysis.

2. **Data Collection:** It involves gathering relevant data from various sources, such as databases, data warehouses, APIs, sensors, or web scraping. The data should cover all necessary variables and attributes related to the problem at hand. It is essential to ensure that data collected is representative of population or domain being studied and has sufficient volume to support accurate analysis.

3. **Data Cleaning:** Data cleaning is a critical step to ensure reliability and data quality. It involves in handling missing values, correcting errors, and resolving inconsistencies in the dataset. Missing data can be imputed by using various methods, such as mean imputation or regression imputation. Errors or inconsistencies can be addressed based on domain knowledge or statistical techniques. Clean and accurate data is essential for obtaining meaningful and accurate results in the subsequent steps.

4. **Data Exploration:** Here, an exploratory data analysis (EDA) is conducted to gain insights into data distribution, identify patterns, and detect outliers. Visualization tools, such as charts, histograms, scatter plots, and heat maps, are often used to aid in understanding the data. EDA helps data scientists to detect any anomalies or unusual observations that might require further investigation.

5. **Feature Selection:** It is the process of choosing a large amount of relevant and informative features (variables) from dataset. This step is crucial in reducing computational complexity and avoiding over fitting. Techniques such as correlation analysis, information gain, or feature importance ranking are used to identify the most influential features that contribute extensively to the analysis. Reducing number of features also helps to improve model performance and interpretability.

6. **Model Building:** Model building is a key step where the selected data mining technique is applied to construct a predictive or descriptive model based on the prepared data. The choice of model depends on the nature of problem and type of data. Common data mining techniques used in this step include Regression, Classification, Clustering, Time Series Analysis an Association Rule Mining. The model is trained using a piece of the data known as the training set. The choice of training set can impact the model's performance, and techniques like cross-validation are often used to evaluate how well the model generalizes to new, unseen data.

7. **Model Evaluation and Deployment:** After building the model, it needs to be evaluated to assess its accuracy and effectiveness. The model is tested on a separate portion of the data known as the test set, which was not used during training. The performance of the model is measured using various evaluation metrics, depending on the type of data mining task:

    - For classification models, evaluation metrics include accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC).
    - For regression models, evaluation metrics often include mean squared error (MSE), mean absolute error (MAE), and R-squared (R2).
    - For clustering models, internal evaluation metrics like silhouette score and external evaluation metrics like adjusted Rand index (ARI) can be used.

- For association rule mining, metrics like support, confidence, and lift are used to assess the quality of discovered rules.
- Model deployment involves integrating the trained model into the business processes or applications to support decision-making and operations. For instance, a predictive model for customer churn can be integrated into a customer relationship management (CRM) system to identify high-risk customers and take appropriate retention measures.

V. DATA MINING TECHNIQUES

Following are different techniques in data mining and is as shown in figure:

- Association
- Classification
- Prediction
- Regression
- Clustering
- Artificial Neural Networks (ANN)
- Outlier detection

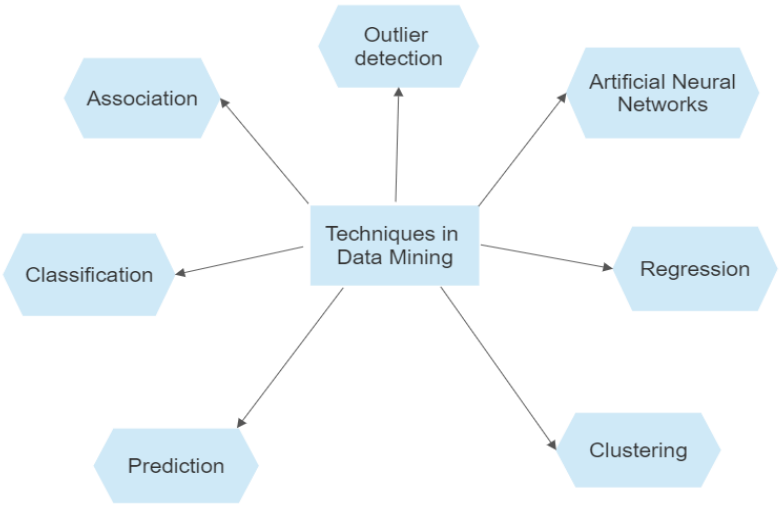

Figure 5

1. **Association:** This analysis is the finding of association rules showing attribute-value conditions that occur frequently together in a given set of data. It is widely used for a market basket or transaction data analysis. The Apriori algorithm is one of the most well-known and widely used techniques for association. Support, Confidence and Lift are the three evaluation metrics used in association rule.

   Example: If a customer buys butter, he most likely can also buy bread, eggs, or milk, so these products are stored within a shelf or mostly nearby in a Super bazaar which is shown in figure.

Figure 6

2. **Classification:** It is a popular supervised learning data mining technique. It involves the process of categorizing data instances into predefined classes or categories based on past observations. Sample example for classification is shown in figure that includes different objects belonging to two classes – class A and class B. It is widely for various tasks, including pattern recognition, image and speech recognition, sentiment w, fraud detection, and medical diagnosis. Decision Tree, K-NN Classifier and Support Vector Machine (SVM) are some of the classification algorithms.

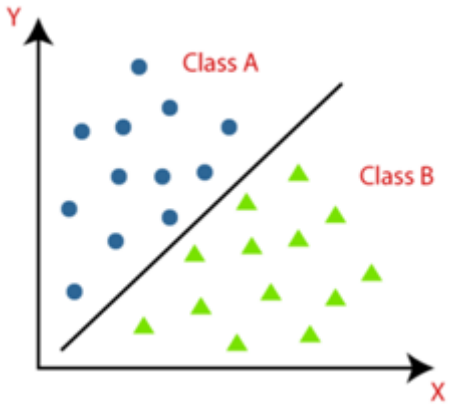

Figure 7

3. **Prediction:** It is a fundamental aspect of data mining and it is a supervised learning. It involves using historical data with known outcomes to build a model that can make predictions on new, unseen data instances. The goal of prediction is to estimate the value of a target variable or outcome based on values of other variables or features in the dataset. Prediction is extensively used in various fields, including healthcare, finance, marketing and manufacturing, to forecast future trends, make informed decisions, and gain insights from data. Regression analysis is most often used for numeric prediction.

   Example: Suppose the marketing manager needs to predict how much a given customer will spend during a sale at his company. In this case, a model or a predictor will be constructed that predicts a continuous-valued-function or ordered value.

4. **Clustering:** It is a data mining technique that falls under the category of unsupervised learning. It involves in the process of grouping similar data points together based on their similarities, without any predefined class labels. The goal of clustering is to discover

natural structures or segments within the data and identify patterns or relationships that might not be apparent initially. Clustering is widely used in various applications, including customer segmentation, anomaly detection, image segmentation, and document grouping.

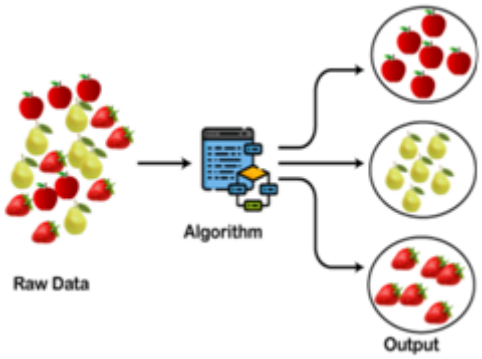

Figure 8

5. **Regression:** It is a data mining technique that falls under the category of supervised learning. It involves the process of modeling the relationship between a dependent variable (target) and one or more independent variables (features) based on historical data with known outcomes. The goal of regression is to build a predictive model that can estimate the value of the dependent variable for new, unseen data instances. Regression is widely used for various tasks, including forecasting, prediction, and trend analysis. Figure shows a graphical representation for linear regression.

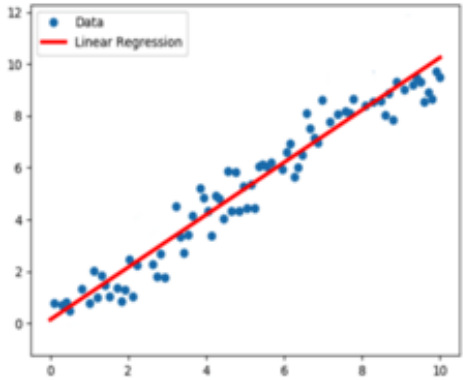

Figure 9

6. **Artificial Neural Networks (ANN)**: This is a powerful data mining technique that falls under the category of machine learning, specifically supervised learning. They are inspired by the structure and function of the human brain, consisting of interconnected nodes (neurons) organized in layers which are shown in figure. ANNs can be used for both classification and regression tasks and are known for their ability to learn complex patterns and relationships in data.

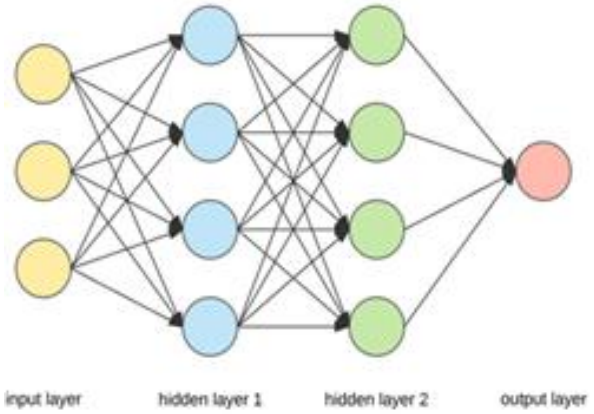

Figure 10

7. **Outlier Detection:** This is also known as anomaly detection, is a data mining technique that involves identifying data points or instances that deviate significantly from the majority of the data. These data points are considered outliers as they do not conform to the regular patterns or a behavior observed in the dataset and is shown in figure. Outlier detection is important in various domains, including fraud detection, fault detection, network intrusion detection, and quality control, as it helps identify unusual or suspicious observations that may indicate critical issues or interesting patterns in the data.