UNIT-5

Data Mining Trends and Research Frontiers : Mining Complex Data Types - Other Methodologies - Data Mining Applications - Data Mining and Society – Data Mining Trends – Real world applications – Data Mining Tool study.

Complex Data Types in Data Mining:

The Complex data types require advanced data mining techniques. Some of the Complex data types are sequence Data which includes the Time-Series, Symbolic Sequences, and Biological Sequences. The additional preprocessing steps are needed for data mining of these complex data types.

1. Time-Series Data Mining:

In time-series data, data is measured as the long series of the numerical or textual data at equal time intervals per minute, per hour, or per day. Time-series data mining is performed on the data obtained from the stock markets, scientific data, and medical data. In time series mining it is not possible to find the data that exactly matches the given query. We employ the similarity search method that finds the data sequences that are similar to the given query string. In the similarity search method, subsequence matching is performed to find the subsequences that are similar to a given query string. In order to perform the similarity search, dimensionality reduction of complex data to transform the time-series data into numerical data.

2. Sequential Pattern Mining in Symbolic Sequences:

Symbolic sequences are composed of long nominal data sequences, which dynamically change their behaviour over time intervals. Examples of the Symbolic Sequences include online customer shopping sequences as well as sequences of events of experiments. Mining of Symbolic Sequences is called Sequential Mining. A sequential pattern is a subsequence that exists more frequently in a set of sequences. so it finds the most frequent subsequence in a set of sequences to perform the mining. Many scalable algorithms have been built to find out the frequent subsequence. There are also algorithms to mine the multidimensional and multilevel sequential patterns.

3. Data mining of Biological Sequences:

Biological sequences are the long sequences of nucleotides and data mining of biological sequences is required to find the features of the DNA of humans. Biological sequence analysis is the first step of data mining to compare the alignment of the biological sequences. Two species are similar to each other only if their nucleotide (DNA, RNA) and protein sequences are close and similar. During the data mining of Biological Sequences, the degree of similarity between nucleotide sequences is measured. The degree of similarity obtained by sequence alignment of nucleotides is essential in determining the homology between two sequences.

There can be the situation of alignment of two or more input biological sequences by identifying similar sequences with long sub-sequences. The amino acids also called proteins sequences are also compared and aligned.

4. Graph Pattern Mining:

Graph Pattern Mining can be done by using Apriori-based and pattern growth-based approaches. We can mine the subgraphs of the graph and the set of closed graphs. A closed graph g is the graph that doesn't have a super graph that carries the same support count as g. Graph Pattern Mining is applied to different types of graphs such as frequent graphs, coherent graphs, and dense graphs. We can also improve the mining efficiency by applying the user constraints on the graph patterns. Graph patterns are two types. Homogeneous graphs where nodes or links of the graph are of the same type by having similar features. In Heterogeneous graph patterns, the nodes and links are of different types.

5. Statistical Modeling of Networks:

A network is a collection of nodes where each node represents the data and the nodes are linked through edges, representing relationships between data objects. If all the nodes and links connecting the nodes are of the same type, then the network is homogeneous such as a friend network or a web page network. If the nodes and links connecting the nodes are of different types, then the network is heterogeneous such as health-care networks (linking the different parameters such as doctors, nurses, patients, diseases together in the network). Graph Pattern Mining can be further applied to the network to derive the knowledge and useful patterns from the network.

6. Mining Spatial Data:

Spatial data is the geo space-related data that is stored in large data repositories. The spatial data is represented in "vector" format and geo-referenced multimedia format. A spatial database is constructed from large geographic data warehouses by integrating geographical data of multiple sources of areas. we can construct spatial data cubes that contain information about the spatial dimensions and measures. It is possible to perform the OLAP operations on the spatial data for spatial data analysis. Spatial data mining is performed on spatial data warehouses, spatial databases, and other geospatial data repositories. Spatial Data mining discovers knowledge about the geographic areas. The preprocessing of spatial data involves several operations like spatial clustering, spatial classification, spatial modeling, and outlier detection in spatial data.

7. Mining Cyber-Physical System Data:

Cyber-Physical System Data can be mined by constructing a graph or network of data. A cyber-physical system (CPS) is a heterogeneous network that consists of a large number of interconnected nodes that store patients or medical information. The links in the CPS network represent the relationship between the nodes . cyber-physical systems store dynamic, inconsistent, and interdependent data that contains spatiotemporal information. Mining cyber-physical data links the situation as a query to access the data from a large information database and it involves real-time calculations and analysis to prompt responses from the CPS system. CPS analysis requires rare-event detection and anomaly analysis in cyber-physical data streams, in cyber-physical networks, and the processing of Cyber-Physical Data involves the integration of stream data with real-time automated control processes.

8. Mining Multimedia Data:

Multimedia data objects include image data, video data, audio data, website hyperlinks, and linkages. Multimedia data mining tries to find out interesting patterns from multimedia databases. This includes the processing of the digital data and performs tasks like image processing, image classification, video, and audio data mining, and pattern recognition. Multimedia Data mining is becoming the most interesting research area because most of the social media platforms like Twitter, Facebook data can be analyzed through this and derive interesting trends and patterns.

9. Mining Web Data:

Web mining is essential to discover crucial patterns and knowledge from the Web. Web content mining analyzes data of several websites which includes the web pages and the multimedia data such as images in the web pages. Web mining is done to understand the content of web pages, unique users of the website, unique hypertext links, web page relevance and ranking, web page content summaries, time that the users spent on the particular website, and understand user search patterns. Web mining also finds out the best search engine and determines the search algorithm used by it. So it helps improve search efficiency and finds the best search engine for the users.

10. Mining Text Data:

Text mining is the subfield of data mining, machine learning, Natural Language processing, and statistics. Most of the information in our daily life is stored as text such as news articles, technical papers, books, email messages, blogs. Text Mining helps us to retrieve high-quality information from text such as sentiment analysis, document summarization, text categorization, text clustering. We apply machine learning models and NLP techniques to derive useful information from the text. This is done by finding out the hidden patterns and trends by means such as statistical pattern learning and statistical language modeling. In order to perform text mining, we need to preprocess the text by applying the techniques of stemming and lemmatization in order to convert the textual data into data vectors.

11. Mining Spatiotemporal Data:

The data that is related to both space and time is Spatiotemporal data. Spatiotemporal data mining retrieves interesting patterns and knowledge from spatiotemporal data. Spatiotemporal Data mining helps us to find the value of the lands, the age of the rocks and precious stones, predict the weather patterns. Spatiotemporal data mining has many practical applications like GPS in mobile phones, timers, Internet-based map services, weather services, satellite, RFID, sensor.

12. Mining Data Streams:

Stream data is the data that can change dynamically and it is noisy, inconsistent which contain multidimensional features of different data types. So this data is stored in NoSql database systems. The volume of the stream data is very high and this is the challenge for the effective mining of stream data. While mining the Data Streams we need to perform the tasks such as clustering, outlier analysis, and the online detection of rare events in data streams.

Data Mining and Society

Data Mining is the process of collecting data and then processing them to find useful patterns with the help of statistics and machine learning processes. By finding the relationship between the database, the peculiarities can be easily identified. Aggregation of useful datasets from a heap of data in the database help in the growth of many industries we depend in our daily life and enhance customer service. We can’t deny the fact that we live in a world of data. From the local grocery store to detecting network frauds, data mining plays a significant role. Beyond the benefits, data mining has negative impacts on society like privacy breaches and security problems. This article shows both the positive and negative effects of data mining on society.

Positive effects of data mining on society

Data mining has influenced our lives whether we feel it or not. Its applications are widely used in many fields to reduce strain and time. It has also supplemented the life of humans. Let’s see some of the examples.

- **Customer relationship management:** By using the techniques of data mining the company provides customized and preferred services for customers which provides a pleasant experience while using the services. By aggregating and grouping the data, the company can create advertisements only when needed and it can reach the right people who require the service. By targeting the customer, unwanted promotional activities can be avoided which saves a lot of money for the company. The customer also doesn’t get annoyed when heaps of junk mails and messages are not sent. Data mining can also help in saving time and provide satisfaction to the customers.

- **Personalized search engines:** In the world of data and networks, our lives become intertwined with web browsers. They had obtained an inevitable place in our lifestyle, knowledge and so on. With the help of data mining algorithms, the suggestions and the order of websites are tailored according to the gathered information by summarizing it. Ranking the page according to the content, the no of visits also help the web browser to provide necessary results for the query given by the user. By giving a personalized environment, spam and misleading advertisements can be avoided. By data mining, frequent spam accounts can be identified and they are automatically moved into the spam folder. For e.g. Gmail has a spam folder where unwanted and frequent junk messages are placed instead of heaping the inbox. Web-wide tracking is a process in which the system keeps track of every website a user visits. By incorporating the DoubleClick mechanism in these websites, they can note the websites that have been visited. And personalized lifestyle, educational ads are made visible in that sites relevant to that user.

- **Mining in the health sector:** Data mining helps in maintaining the health and welfare of a human. The layers of data mining embedded in pharmaceutical industries help to analyze data, to establish relationships while creating and improving drugs. It also helps in analyzing the effects of drugs on patients, the side effects and outcomes. They also help in tracking the number of chronically diseased patients, ICU patients which help in reducing the overflow of admissions in hospitals. Some medicines can also cause side effects or other benefits regardless of what disease it treats. In such cases, data mining can largely influence the growth of the health sector.

- **E-shopping:** E- retail platforms are one of the fastest-growing major industries in the world. From books, movies, groceries, lifestyles everything is listed on online e-retail platforms. This cannot run successfully without the help of data mining and predictive analysis. By these techniques, cross-selling and holding onto regular customers have become possible. Data mining helps in announcing offers and discounts to keep the customers intact and to increase sales. By using the algorithms of data science, the e-commerce website can largely influence the customers using targeted ad campaigns which will surge the number of users as well as it provide satisfactory results to customers.

- **Crime prevention:** Data mining plays a huge role in the prevention of crimes and reducing fraud rates. In telecommunication industries, it helps in identifying subscription theft and super-imposed frauds. It also helps in the identification of fraudulent calls. By doing this, user security can be ensured and prevent the company from facing a huge loss. It also plays an important role in police departments for identifying key patterns in crime and predicting them. It also helps in identifying the unsolved crimes committed by the same criminal by establishing a relationship between previous and present datasets in the crime database. By extracting and aggregating data, the police department can identify future crimes and prevent them. It also helps in identifying the cause of crime and the criminal behind that. This application largely supports the safety of people.

Negative effects of data mining on society

- **The exploitation of data and discrimination:** By agreeing on the terms and conditions provided by a company, the company gets access to collect data of the customers. From age groups to economical status, the company profiles the customers. By customer profiling, they get to know the datasets of rich, poor, elder, or younger. Some unethical or devious companies offer low credits or inferior deals to the customer in an area where fewer sales rate is noted. For. eg. An unethical company decreasing the credit scores in the loyalty card of people connected to a branch whose transactions are less. Sometimes while profiling customers, wrongly accusing a customer happens. Though he is faultless, his needs and comfort are denied. Even though the company declares the customer faultless after investigations, still the wrongly accused customer struggles mentally and this incident will negatively impact his life. Certain companies don't take the responsibility of securing the data of customers which makes the data vulnerable and causes privacy breaches.
- **Health-related ethical problem:** Using data mining techniques, the companies can extract data about the health problems of the employees. They can also relate the summarized dataset with the datasets from the past history of previous employees. By discovering the pattern of diseases and frequencies, the company chooses the specific insurance plans accordingly. But, there is a chance that the company uses this data while hiring new employees. Hence, they avoid hiring people with a higher frequency of sickness. Insurance companies collect this data so they can avoid policies with companies with a high risk of health issues.
- **Privacy breach:** Every single piece of data we enter into the database of the internet is indirectly under the control of data miners. When used for unethical purposes, the privacy of an individual is invaded. Certain companies use this data to filter the latent people but with the potential to become customers of that company. In this way, the company sends targeted advertisements and increase customer traffic. For e.g. In telecommunication industries, the call details of the customer are created to enhance business growth and to maintain low customer churn. But, the company uses the data selfishly for its growth and it leads to the exploitation of privacy. Thus every single piece of data given to the network stands for greater risk under the influence of data mining.
- **Manipulation of data and unethical problems:** There are circumstances when normal data provided by a customer or user becomes manipulative data. That is when a customer makes a promotion on social media, which means he has a good financial status in his growing business. Using such information, miners can obtain unethical data to gain profits or access. Spreading of false information through social media and erroneous opinions can mislead people because when data miners collect this information, they become facts and which leads to a scam. Also, by using predictive

analytics and machine learning algorithms, the outcome of an event is predicted by the government which may fail sometimes and that will create a disastrous effect on the public. When the prediction is based on unprompted unsafe sources, those predictions lead to severe losses and the company may fail.

- **Invasive marketing:** The junk advertisements that heap your mobile while using social media or other social platforms are the result of data mining. Targeted ads benefit both seller and customer and save time but when it gets intense and unethical, wrong products are forced through advertisements that may negatively influence the life of the user. From the browser histories to previously purchased items, the data is extracted and used to influence the user to buy other products sometimes which may be harmful. This aggressive technique will cause undesirable effects on the user. Every discovery or field has its own merits and demerits. A part of that application may help the human and a part may degrade the values and ethics of society. As a part of the society we live in, it is our duty to use the applications of technology following the rules and maintaining ethics. Industries, companies and marketing agents should respect the privacy of individual humans and should provide the space they need. When every single person out there takes responsibility for the proper handling of data, data mining would be a gift of technology that could build and ease our life in so many ways.

Applications of Data Mining

Data is a set of discrete objective facts about an event or a process that have little use by themselves unless converted into information. We have been collecting numerous data, from simple numerical measurements and text documents to more complex information such as spatial data, multimedia channels, and hypertext documents.

Nowadays, large quantities of data are being accumulated. The amount of data collected is said to be almost doubled every year. An extracting data or seeking knowledge from this massive data, data mining techniques are used. Data mining is used in almost all places where a large amount of data is stored and processed. For example, banks typically use 'data mining' to find out their prospective customers who could be interested in credit cards, personal loans, or insurance as well. Since banks have the transaction details and detailed profiles of their customers, they analyze all this data and try to find out patterns that help them predict that certain customers could be interested in personal loans, etc.

Basically, the motive behind mining data, whether commercial or scientific, is the same – the need to find useful information in data to enable better decision-making or a better understanding of the world around us.

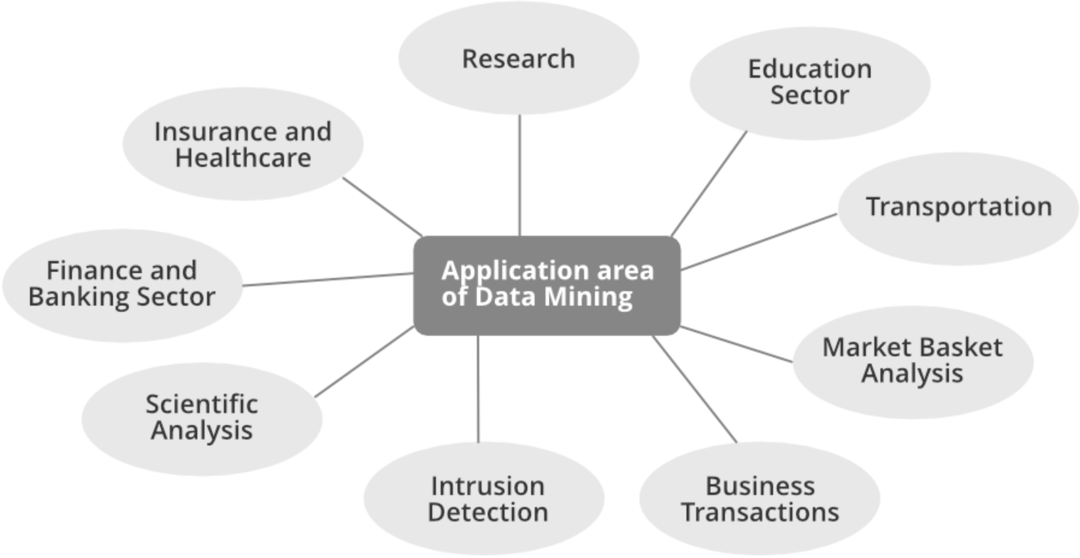
"Extraction of interesting information or patterns from data in large databases is known as data mining."

According to William J.Frawley "Data mining or KDD(Knowledge Discovery in Databases) as it is also known, is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data."

Technically, data mining is the computational process of analyzing data from different perspectives, dimensions, angles and categorizing/summarizing it into meaningful information. Data Mining can be applied to any type of data e.g. Data Warehouses, Transactional Databases, Relational Databases, Multimedia Databases, Spatial Databases, Time-series Databases, World Wide Web.

Data mining provides competitive advantages in the knowledge economy. It does this by providing the maximum knowledge needed to rapidly make valuable business decisions despite the enormous amounts of available data.

There are many measurable benefits that have been achieved in different application areas from data mining. So, let's discuss different applications of Data Mining:

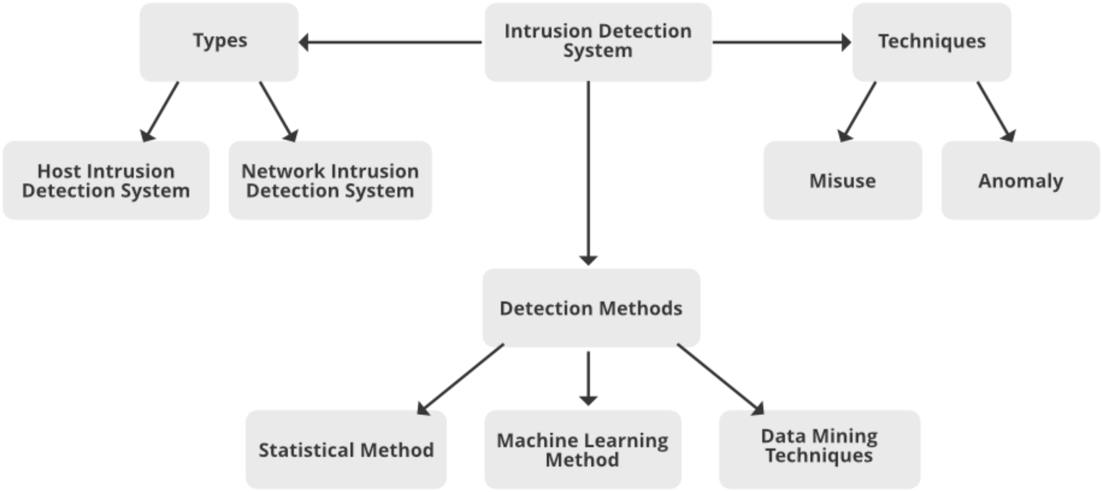


Scientific Analysis: Scientific simulations are generating bulks of data every day. This includes data collected from nuclear laboratories, data about human psychology, etc. Data mining techniques are capable of the analysis of these data. Now we can capture and store more new data faster than we can analyze the old data already accumulated. Example of scientific analysis:

- Sequence analysis in bioinformatics
- Classification of astronomical objects
- Medical decision support.

Intrusion Detection: A network intrusion refers to any unauthorized activity on a digital network. Network intrusions often involve stealing valuable network resources. Data mining technique plays a vital role in searching intrusion detection, network attacks, and anomalies. These techniques help in selecting and refining useful and relevant information from large data sets. Data mining technique helps in classify relevant data for Intrusion Detection System. Intrusion Detection system generates alarms for the network traffic about the foreign invasions in the system. For example:

- Detect security violations
- Misuse Detection
- Anomaly Detection



Business Transactions: Every business industry is memorized for perpetuity. Such transactions are usually time-related and can be inter-business deals or intra-business operations. The effective and in-time use of the data in a reasonable time frame for competitive decision-making is definitely the most important problem to solve for businesses that struggle to survive in a highly competitive world. Data mining helps to analyze these business transactions and identify marketing approaches and decision-making. Example :

- Direct mail targeting
- Stock trading
- Customer segmentation
- Churn prediction (Churn prediction is one of the most popular Big Data use cases in business)

Market Basket Analysis: Market Basket Analysis is a technique that gives the careful study of purchases done by a customer in a supermarket. This concept identifies the pattern of frequent purchase items by customers. This analysis can help to promote deals, offers, sale by the companies and data mining techniques helps to achieve this analysis task. Example:

- Data mining concepts are in use for Sales and marketing to provide better customer service, to improve cross-selling opportunities, to increase direct mail response rates.
- Customer Retention in the form of pattern identification and prediction of likely defections is possible by Data mining.
- Risk Assessment and Fraud area also use the data-mining concept for identifying inappropriate or unusual behavior etc.

Education: For analyzing the education sector, data mining uses Educational Data Mining (EDM) method. This method generates patterns that can be used both by learners and educators. By using data mining EDM we can perform some educational task:

- Predicting students admission in higher education
- Predicting students profiling
- Predicting student performance
- Teachers teaching performance
- Curriculum development
- Predicting student placement opportunities

Research: A data mining technique can perform predictions, classification, clustering, associations, and grouping of data with perfection in the research area. Rules generated by data mining are unique to find results. In most of the technical research in data mining, we create a training model and testing model. The training/testing model is a strategy to measure the precision of the proposed model. It is called Train/Test because we split the data set into two sets: a training data set and a testing data set. A training data set used to design the training model whereas testing data set is used in the testing model. Example:

- Classification of uncertain data.
- Information-based clustering.
- Decision support system
- Web Mining
- Domain-driven data mining
- IoT (Internet of Things)and Cybersecurity
- Smart farming IoT(Internet of Things)

Healthcare and Insurance: A Pharmaceutical sector can examine its new deals force activity and their outcomes to improve the focusing of high-value physicians and figure out which promoting activities will have the best effect in the following upcoming months, Whereas the Insurance sector, data mining can help to predict which customers will buy new policies, identify behavior patterns of risky customers and identify fraudulent behavior of customers.

- Claims analysis i.e which medical procedures are claimed together.
- Identify successful medical therapies for different illnesses.
- Characterizes patient behavior to predict office visits.

Transportation: A diversified transportation company with a large direct sales force can apply data mining to identify the best prospects for its services. A large consumer merchandise organization can apply information mining to improve its business cycle to retailers.

- Determine the distribution schedules among outlets.
- Analyze loading patterns.

Financial/Banking Sector: A credit card company can leverage its vast warehouse of customer transaction data to identify customers most likely to be interested in a new credit product.

- Credit card fraud detection.
- Identify ‘Loyal’ customers.
- Extraction of information related to customers.
- Determine credit card spending by customer groups.

Data Mining - Applications & Trends:

Data mining is widely used in diverse areas. There are a number of commercial data mining system available today and yet there are many challenges in this field. In this tutorial, we will discuss the applications and the trend of data mining.

Data Mining Applications

Here is the list of areas where data mining is widely used −

- Financial Data Analysis
- Retail Industry
- Telecommunication Industry
- Biological Data Analysis
- Other Scientific Applications

- Intrusion Detection

Financial Data Analysis

The financial data in banking and financial industry is generally reliable and of high quality which facilitates systematic data analysis and data mining. Some of the typical cases are as follows −

- Design and construction of data warehouses for multidimensional data analysis and data mining.
- Loan payment prediction and customer credit policy analysis.
- Classification and clustering of customers for targeted marketing.
- Detection of money laundering and other financial crimes.

Retail Industry

Data Mining has its great application in Retail Industry because it collects large amount of data from on sales, customer purchasing history, goods transportation, consumption and services. It is natural that the quantity of data collected will continue to expand rapidly because of the increasing ease, availability and popularity of the web.

Data mining in retail industry helps in identifying customer buying patterns and trends that lead to improved quality of customer service and good customer retention and satisfaction. Here is the list of examples of data mining in the retail industry −

- Design and Construction of data warehouses based on the benefits of data mining.
- Multidimensional analysis of sales, customers, products, time and region.
- Analysis of effectiveness of sales campaigns.
- Customer Retention.
- Product recommendation and cross-referencing of items.

Telecommunication Industry

Today the telecommunication industry is one of the most emerging industries providing various services such as fax, pager, cellular phone, internet messenger, images, e-mail, web data transmission, etc. Due to the development of new computer and communication technologies, the telecommunication industry is rapidly expanding. This is the reason why data mining is become very important to help and understand the business.

Data mining in telecommunication industry helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and improve quality of service. Here is the list of examples for which data mining improves telecommunication services −

- Multidimensional Analysis of Telecommunication data.
- Fraudulent pattern analysis.
- Identification of unusual patterns.
- Multidimensional association and sequential patterns analysis.
- Mobile Telecommunication services.
- Use of visualization tools in telecommunication data analysis.

Biological Data Analysis

In recent times, we have seen a tremendous growth in the field of biology such as genomics, proteomics, functional Genomics and biomedical research. Biological data mining is a very important part of Bioinformatics. Following are the aspects in which data mining contributes for biological data analysis −

- Semantic integration of heterogeneous, distributed genomic and proteomic databases.
- Alignment, indexing, similarity search and comparative analysis multiple nucleotide sequences.
- Discovery of structural patterns and analysis of genetic networks and protein pathways.
- Association and path analysis.
- Visualization tools in genetic data analysis.

Other Scientific Applications

The applications discussed above tend to handle relatively small and homogeneous data sets for which the statistical techniques are appropriate. Huge amount of data have been collected from scientific domains such as geosciences, astronomy, etc. A large amount of data sets is being generated because of the fast numerical simulations in various fields such as climate and ecosystem modeling, chemical engineering, fluid dynamics, etc. Following are the applications of data mining in the field of Scientific Applications −

- Data Warehouses and data preprocessing.
- Graph-based mining.
- Visualization and domain specific knowledge.

Intrusion Detection

Intrusion refers to any kind of action that threatens integrity, confidentiality, or the availability of network resources. In this world of connectivity, security has become the major issue. With increased usage of internet and availability of the tools and tricks for intruding and attacking network prompted intrusion detection to become a critical component of network administration. Here is the list of areas in which data mining technology may be applied for intrusion detection −

- Development of data mining algorithm for intrusion detection.
- Association and correlation analysis, aggregation to help select and build discriminating attributes.
- Analysis of Stream data.
- Distributed data mining.
- Visualization and query tools.

Data Mining System Products

There are many data mining system products and domain specific data mining applications. The new data mining systems and applications are being added to the previous systems. Also, efforts are being made to standardize data mining languages.

Choosing a Data Mining System

The selection of a data mining system depends on the following features −

- **Data Types** − The data mining system may handle formatted text, record-based data, and relational data. The data could also be in ASCII text, relational database data or data warehouse data. Therefore, we should check what exact format the data mining system can handle.
- **System Issues** − We must consider the compatibility of a data mining system with different operating systems. One data mining system may run on only one operating system or on several. There are also data mining systems that provide web-based user interfaces and allow XML data as input.
- **Data Sources** − Data sources refer to the data formats in which data mining system will operate. Some data mining system may work only on ASCII text files while others on multiple relational sources. Data mining system should also support ODBC connections or OLE DB for ODBC connections.

- **Data Mining functions and methodologies** − There are some data mining systems that provide only one data mining function such as classification while some provides multiple data mining functions such as concept description, discovery-driven OLAP analysis, association mining, linkage analysis, statistical analysis, classification, prediction, clustering, outlier analysis, similarity search, etc.
- **Coupling data mining with databases or data warehouse systems** − Data mining systems need to be coupled with a database or a data warehouse system. The coupled components are integrated into a uniform information processing environment. Here are the types of coupling listed below −
  - No coupling
  - Loose Coupling
  - Semi tight Coupling
  - Tight Coupling
- **Scalability** − There are two scalability issues in data mining −
  - **Row (Database size) Scalability** − A data mining system is considered as row scalable when the number or rows are enlarged 10 times. It takes no more than 10 times to execute a query.
  - **Column (Dimension) Salability** − A data mining system is considered as column scalable if the mining query execution time increases linearly with the number of columns.
- **Visualization Tools** − Visualization in data mining can be categorized as follows −
  - Data Visualization
  - Mining Results Visualization
  - Mining process visualization
  - Visual data mining
- **Data Mining query language and graphical user interface** − An easy-to-use graphical user interface is important to promote user-guided, interactive data mining. Unlike relational database systems, data mining systems do not share underlying data mining query language.

Trends in Data Mining

Data mining concepts are still evolving and here are the latest trends that we get to see in this field −

- Application Exploration.

- Scalable and interactive data mining methods.
- Integration of data mining with database systems, data warehouse systems and web database systems.
- Standardization of data mining query language.
- Visual data mining.
- New methods for mining complex types of data.
- Biological data mining.
- Data mining and software engineering.
- Web mining.
- Distributed data mining.
- Real time data mining.
- Multi database data mining.
- Privacy protection and information security in data mining.

Data Mining Trends and Research Frontiers:

Data mining is the process of analyzing a large size of information to find out the patterns, trends. It can be used for corporations to find out about customers' choices, make a good relationship with customers, increase the revenue, reduce risks. Data mining is based on complex algorithms that allow data segmentation to discover numerous trends and patterns, detect deviations, and estimate the likelihood of certain occurrences occurring. Raw data can be in both analog and digital formats, and it is essentially dependent on the data's source. Companies must keep up with the latest data mining trends and stay current in order to succeed in the industry and beat out the competition.

Data mining is one of the most widely used methods to extract data from different sources and organize them for better usage. Despite having different commercial systems for data mining, many challenges come up when they are actually implemented. With the rapid evolution in the field of data mining, companies are expected to stay abreast with all the new developments.

Complex algorithms form the basis for data mining as they allow data segmentation to identify trends and patterns, detect variations, and predict the probabilities of various events. The raw data may come in both analog and digital formats and is inherently based on the source of the data. Companies need to keep track of the latest data mining trends and stay updated to do well in the industry and overcome challenging competition.

Corporations can use data mining to discover customers' choices, make a good relationship with customers, increase revenue, and reduce risks. Data mining is based on complex algorithms that allow data segmentation to discover numerous trends and patterns, detect deviations, and estimate the likelihood of certain occurrences occurring. Raw data can be in both analog and digital formats, and it is essentially dependent on the data's source. Companies must keep up with the latest data mining trends and stay current to succeed in the industry and beat out the competition.

Types of Mining Sequence in Data Mining:

- Mining time series
- Mining symbolic sequence
- Mining biological sequence

1. Mining Time Series

A specified number of data points are recorded at a specific time or events obtained over repeated measurements of time in a mining time series. The values or data are typically measured in equal time intervals like- hourly, weekly, daily. In time-series data is also recorded regular intervals or characteristic time-series components are trend, seasonal, cycle, irregular.

Application of Time Series:

- Financial: Stock market analysis
- Industry: Power consumption
- Scientific: Experiment result
- Meteorological: Precipitation

Time Series Analysis Methods:

- **Trend Analysis:** Categories of Time Series movements:
  - **Long-term or Trend Movements:** General direction in which a time series is moving over a long interval of time.
  - **Cyclic Movements:** Long-term oscillation about a trend line or curve.

    - **Seasonal Movements:** A time series appears to follow substantially identical patterns during the corresponding months of subsequent years.
    - **Irregular or Random Movements:** It changes that occur randomly due to unplanned events.

- **Similarity Search:**
    - Data Reduction
    - Indexing Methods
    - Similarity Search Methods
    - Query Languages

2. Mining Symbolic Sequence

A symbolic sequence is made up of an ordered list of elements that can be recorded with or without a sense of time. This sequence can be used in a variety of ways, including consumer shopping sequences, web clickstreams, software execution sequences, biological sequences, and so on.

Mining of sequential patterns entails identifying the subsequences that appear frequently in one or more sequences. As a result of substantial research in this area, a number of scalable algorithms have been developed. Alternatively, we can only mine the set of closed sequential patterns, where a sequential pattern s is closed if it is a correct subsequence of s' and s' has the same support as s.

For example:

if where a, b, c, d and e are items, then S is a subsequence of S'.

3. Mining Biological Sequence

Biological sequences are made up of nucleotide or amino acid sequences. In bioinformatics and modern biology, biological sequence analysis compares, aligns, indexes, and analyzes biological sequences. Biological sequences analysis plays a crucial role in bioinformatics and modern biology. Such analysis can be partitioned into two tasks- pairwise sequence alignment and multiple sequence alignment.

Biological Sequence Methods:

- Alignment of Biological Sequences:
    - Pairwise Alignment
    - The BLAST Local Alignment Algorithm

    - Multiple Sequence Alignment Methods
- Biological Sequence Analysis Using a Hidden Markov Model:
    - Markov Chain
    - Hidden Markov Model
    - Forward Algorithm
    - Viterbi Algorithm
    - Baum-Welch Algorithm

Application of Data Mining:

- **Financial Information Analysis:**
    - Loan payment prediction/consumer credit policy analysis
    - Design and construction of information warehouse
    - Financial information collected in banks and money establishments area units are typically comparatively complete, reliable, and of top quality.
- **Retail Industry:**
    - Multidimensional analysis(sales, customers, products, time, etc.)
    - Sales campaign analysis
    - Customer retention
    - Product recommendation
    - Using visualization tools for data analysis
- **Science and Engineering:**
    - Data processing and data warehouse
    - Mining complex data types
    - Network-based mining
    - Graph-based mining

Trends of Data Mining:

- Exploration of applications: addressing application-specific issues
- Data mining approaches that are scalable and interactive
- Data mining integration with Web search engines, database systems, data warehouse systems, and cloud computing systems

- Mining social and information networks
- Mining spatiotemporal, moving objects, and cyber-physical systems
- Mining multimedia, text, and web data
- Mining biological and biomedical data
- Visual and audio data mining
- Distributed data mining and real-time data stream mining.

1. Mining Time Series

A specified number of data points are recorded at a specific time or events obtained over repeated measurements of time in a mining time series. The values or data are typically measured in equal time intervals like- hourly, weekly, or daily. Time-series data is also recorded at regular intervals, or characteristic time-series components are the trend, seasonal, cycle, or irregular.

Application of Time Series

- Financial: Stock market analysis
- Industry: Power consumption
- Scientific: Experiment result
- Meteorological: Precipitation

Time Series Analysis Methods

Trend Analysis: Categories of Time Series movements:

- **Long-term or Trend Movements:** General direction in which a time series moves over a long time interval.
- **Cyclic Movements:**Long-term oscillation about a trend line or curve.
- **Seasonal Movements:**A time series appears to follow substantially identical patterns during the corresponding months of subsequent years.
- **Irregular or Random Movements:** Changes that occur randomly due to unplanned events.

Similarity Search:

- Data Reduction
- Indexing Methods
- Similarity Search Methods

- Query Languages

2. Mining Symbolic Sequence

A symbolic sequence comprises an ordered list of elements that can be recorded with or without a sense of time. This sequence can be used in various ways, including consumer shopping sequences, web clickstreams, software execution sequences, biological sequences, etc.

Mining sequential patterns entail identifying the subsequences that frequently appear in one or more sequences. As a result of substantial research in this area, many scalable algorithms have been developed. Alternatively, we can only mine the set of closed sequential patterns, where a sequential pattern s is closed if a correct subsequence of s' and s' has the same support as s.

3. Mining Biological Sequence

Biological sequences are made up of nucleotide or amino acid sequences. Biological sequence analysis compares, aligns, indexes, and analyzes biological sequences in bioinformatics and modern biology. Biological sequences analysis plays a crucial role in bioinformatics and modern biology. Such analysis can be partitioned into pairwise sequence alignment and multiple sequence alignment.

Biological Sequence Methods:

i. Alignment of Biological Sequences:
   - Pairwise Alignment
   - The BLAST Local Alignment Algorithm
   - Multiple Sequence Alignment Methods

ii. Biological Sequence Analysis Using a Hidden Markov Model:
   - Markov Chain
   - Hidden Markov Model
   - Forward Algorithm
   - Viterbi Algorithm
   - Baum-Welch Algorithm

Application of Data Mining:

i. **Financial Information Analysis:**
   - Loan payment prediction/consumer credit policy analysis

- Design and construction of information warehouse
- Financial information collected in bank and money establishments area units is typically comparatively complete, reliable, and top-quality.

ii. **Retail Industry:**

- Multidimensional analysis (sales, customers, products, time, etc.)
- Sales campaign analysis
- Customer retention
- Product recommendation
- Using visualization tools for data analysis

iii. **Science and Engineering:**

- Data processing and data warehouse
- Mining complex data types
- Network-based mining
- Graph-based mining

Trends in Data Mining

Businesses that have been slow in adopting the process of data mining are now catching up with the others. Extracting important information through the process of data mining is widely used to make critical business decisions. We can expect data mining to become as ubiquitous as some of the more prevalent technologies used today in the coming decade. Data mining concepts are still evolving, and here are the following latest trends, such as:

1. Application exploration

Data mining is increasingly used to explore applications in other areas, such as financial analysis, telecommunications, biomedicine, wireless security, and science.

2. Multimedia Data Mining

This is one of the latest methods which is catching up because of the growing ability to capture useful data accurately. It involves data extraction from different kinds of multimedia sources such as audio, text, hypertext, video, images, etc. The data is converted into a numerical representation in different formats. This method can be used in clustering and classifications, performing similarity checks, and identifying associations.

3. Ubiquitous Data Mining

This method involves mining data from mobile devices to get information about individuals. Despite having several challenges in this type, such as complexity, privacy, cost, etc., this method has a lot of opportunities to be enormous in various industries, especially in studying human-computer interactions.

4. Distributed Data Mining

This type of data mining is gaining popularity as it involves mining a huge amount of information stored in different company locations or at different organizations. Highly sophisticated algorithms are used to extract data from different locations and provide proper insights and reports based on them.

5. Embedded Data Mining

Data mining features are increasingly finding their way into many enterprise software use cases, from sales forecasting in CRM SaaS platforms to cyber threat detection in intrusion detection/prevention systems. The embedding of data mining into vertical market software applications enables prediction capabilities for any number of industries and opens up new realms of possibilities for unique value creation.

6. Spatial and Geographic Data Mining

This new trending type of data mining includes extracting information from environmental, astronomical, and geographical data, including images taken from outer space. This type of data mining can reveal various aspects such as distance and topology, which are mainly used in geographic information systems and other navigation applications.

7. Time Series and Sequence Data Mining

The primary application of this type of data mining is the study of cyclical and seasonal trends. This practice is also helpful in analyzing even random events which occur outside the normal series of events. Retail companies mainly use this method to access customers' buying patterns and behaviors.

8. Data Mining Dominance in the Pharmaceutical And Health Care Industries

Both the pharmaceutical and health care industries have long been innovators in the category of data mining. The recent rapid development of coronavirus vaccines is directly attributed to advances in pharmaceutical testing data mining techniques, specifically signal detection during the clinical trial process for new drugs. In health care, specialized data mining techniques are being used to analyze DNA sequences for creating custom therapies, make better-informed diagnoses, and more.

9. Increasing Automation In Data Mining

Today's data mining solutions typically integrate ML and big data stores to provide advanced data management functionality alongside sophisticated data analysis techniques. Earlier incarnations of data mining involved manual coding by specialists with a deep background in statistics and programming. Modern techniques are highly automated, with AI/ML replacing most of these previously manual processes for developing pattern-discovering algorithms.

10. Data Mining Vendor Consolidation

If history is any indication, significant product consolidation in the data mining space is imminent as larger database vendors acquire data mining tooling startups to augment their offerings with new features. The current fragmented market and a broad range of data mining players resemble the adjacent big data vendor landscape that continues to undergo consolidation.

11. Biological data mining

Mining DNA and protein sequences, mining high dimensional microarray data, biological pathway and network analysis, link analysis across heterogeneous biological data, and information integration of biological data by data mining are interesting topics for biological data mining research.

Data Mining:

Data mining refers to filtering, sorting, and classifying data from larger datasets to reveal subtle patterns and relationships, which helps enterprises identify and solve complex business problems through data analysis. Data mining software tools and techniques allow organizations to foresee future market trends and make business-critical decisions at crucial times.

Data mining is an essential component of data science that employs advanced data analytics to derive insightful information from large volumes of data. If we dig deeper, data mining is a crucial ingredient of the knowledge discovery in databases (KDD) process, where data gathering, processing, and analysis takes place at a fundamental level.

Businesses rely heavily on data mining to undertake analytics initiatives in the organizational setup. The analyzed data sourced from data mining is used for varied analytics and business intelligence (BI) applications, which consider real-time data analysis along with some historical pieces of information.

With top-notch data mining practices, enterprises can make several business strategies and manage their operations better. This can entail refining customer-centric functions, including advertising, marketing, sales, customer support, finance, HR, etc.

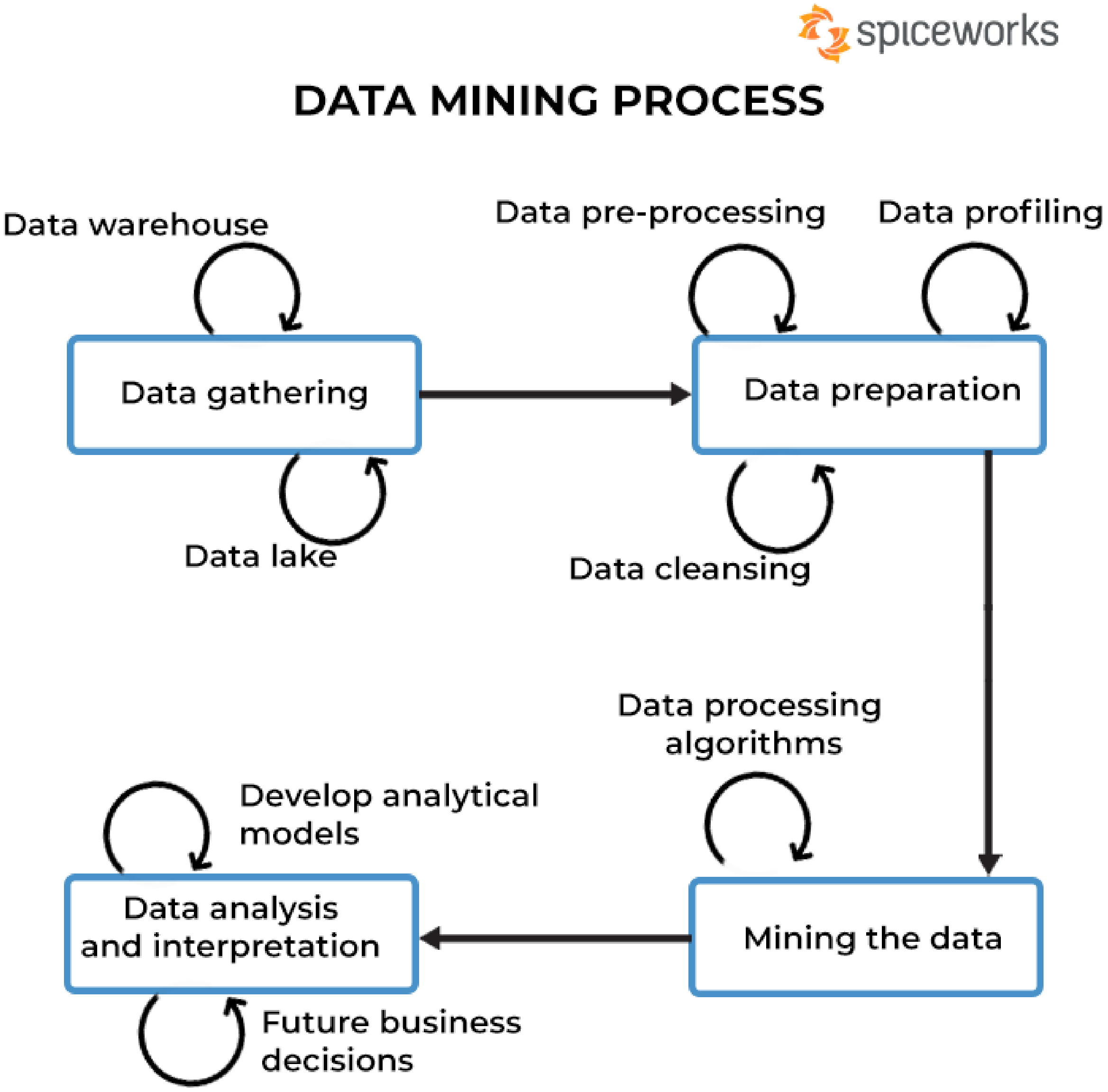
Data mining also plays a vital role in handling business-critical use cases such as cybersecurity planning, fraud detection, risk management, and several others. Data mining finds applications across industry verticals such as healthcare, scientific research, sports, governmental projects, etc.

How does data mining work?

Data mining is predominantly handled by a group of data scientists, skilled BI professionals, analytics groups, business analysts, tech-savvy executives, and personnel having a solid background and inclination toward data analytics.

Fundamentally, machine learning (ML), artificial intelligence (AI), statistical analysis, and data management are crucial elements of data mining that are necessary to scrutinize, sort, and prepare data for analysis. Top ML algorithms and AI tools have enabled the easy mining of massive datasets, including customer data, transactional records, and even log files picked up from sensors, actuators, IoT devices, mobile apps, and servers.

Key stages involved in the data mining process:



Data Mining Process

-

  - **Data gathering:** Data mining begins with the data gathering step, where relevant information is identified, collected, and organized for analysis. Data sources can include data warehouses, data lakes, or any other source that contains raw data in a structured or unstructured format.

-
  - **Data preparation**: In the second step, fine-tuning the gathered data is the prime focus. This involves several processes, such as data pre-processing, data profiling, and data cleansing, to fix any data errors. These stages are essential to maintain data quality before following up with the mining and analysis processes.

-
  - **Mining the data**: In the third step, the data professional selects an appropriate data mining technique once the desired quality of data is prepared. Here, a proper set of data processing algorithms are identified where sample data is trained initially before running it over the entire dataset.

-
  - **Data analysis and interpretation**: In the last step, the results derived in the third step are used to develop analytical models for making future business decisions. Moreover, the data science team communicates the results to the concerned stakeholders via data visualizations and other more straightforward techniques. The information is conveyed in a manner that makes the content digestible for any non-expert working outside the field of data science.

Benefits of data mining

Data mining is beneficial for most businesses primarily because it can run through vast volumes of data and identify hidden patterns, relationships, and trends. The results are helpful for predictive analytics that help in strategic planning while keeping a stock of the current business scenario.

Benefits of data mining for enterprises:

-
  - **Targeted marketing & advertisements:** Data mining allows marketing teams to comprehend customer behavior and preferences better. It will enable them to direct targeted advertisements to respective customers showing a pattern of behavior. Moreover, the sales department benefits from data mining as it helps them target customers with a particular inclination toward specific products. It additionally allows them to sell more services and products to older customers.

- **Identifying customer service issues:** Data mining is an effective tool to keep track of customer service issues when customers interact with contact center agents through calls and online chats. It gives them a chance to provide better customer service, thanks to the in-depth analysis possible through data mining.
- **Improved supply chain management (SCM)**: With data mining, businesses can identify market trends and predict future customer behavior that can impact product demand. This allows enterprises to plan for the future and manage the supply of goods and services to meet market demands. Moreover, SCM managers can plan their logistic operations accordingly, streamline product distribution, and optimize warehousing services.
- **Maintaining production uptime**: Gathering and mining data from sensors, IoT devices, manufacturing machines, and industrial equipment aids in creating predictive maintenance applications that determine potential problems before the actual incident hurts the machinery. Such pre-timed warnings reduce the unscheduled downtime for machines, thereby boosting overall productivity.
- **Better assess risks**: Data mining allows risk managers and concerned business personnel to assess better the risks related to finances, legal matters, or cybersecurity factors that the company may encounter in the future. It gives them the chance to properly prepare for such events and have a plan in place to manage such mishaps better.
- **Drive cost savings**: Data mining can easily identify any operational inefficiencies in a typical business process. This early problem identification helps streamline corporate processes that align with a company's business goals, thereby saving considerably on corporate spending.

Data mining plays a pivotal role in strategizing plans that help companies gain higher business profits and revenues and set them aside from their competitors.

Data Mining Techniques

Every data science application demands a different data mining technique. One of the popular and well-known data mining techniques used includes pattern recognition and anomaly detection. Both these methods employ a combination of techniques to mine data.

Let's look at some of the fundamental data mining techniques commonly used across industry verticals.

1. Association rule

The association rule refers to the if-then statements that establish correlations and relationships between two or more data items. The correlations are evaluated using support and confidence metrics, wherein support determines the frequency of occurrence of data items within the dataset. In contrast, confidence relates to the accuracy of if-then statements.

For example, while tracking a customer's behavior when purchasing online items, an observation is made that the customer generally buys cookies when purchasing a coffee pack. In such a case, the association rule establishes the relation between two items of cookies and coffee packs, thereby forecasting future buys whenever the customer adds the coffee pack to the shopping cart.

2. Classification

The classification data mining technique classifies data items within a dataset into different categories. For example, we can classify vehicles into different categories, such as sedan, hatchback, petrol, diesel, electric vehicle, etc., based on attributes such as the vehicle's shape, wheel type, or even number of seats. When a new vehicle arrives, we can categorize it into various classes depending on the identified vehicle attributes. One can apply the same classification strategy to classify customers based on their age, address, purchase history, and social group.

Some of the examples of classification methods include **decision trees,** Naive Bayes classifiers, logistic regression, and so on.

3. Clustering

Clustering data mining techniques group data elements into clusters that share common characteristics. We can cluster data pieces into categories by simply identifying one or more attributes. Some of the well-known clustering techniques are k-means clustering, hierarchical clustering, and Gaussian mixture models.

4. Regression

Regression is a statistical modeling technique using previous observations to predict new data values. In other words, it is a method of determining relationships between data elements based on the predicted data values for a set of defined variables. This category's classifier is called the 'Continuous Value Classifier'. **Linear regression,** multivariate regression, and decision trees are key examples of this type.

5. Sequence & path analysis

One can also mine sequential data to determine patterns, wherein specific events or data values lead to other events in the future. This technique is applied for long-term data as sequential analysis is key to identifying trends or regular occurrences of certain events. For example, when a customer buys a grocery item, you can use a sequential pattern to suggest or add another item to the basket based on the customer's purchase pattern.

6. Neural networks

Neural networks technically refer to algorithms that mimic the human brain and try to replicate its activity to accomplish a desired goal or task. These are used for several pattern recognition applications that typically involve deep learning techniques. Neural networks are a consequence of advanced machine learning research.

7. Prediction

The prediction data mining technique is typically used for predicting the occurrence of an event, such as the failure of machinery or a fault in an industrial component, a fraudulent event, or company profits crossing a certain threshold. Prediction techniques can help analyze trends, establish correlations, and do pattern matching when combined with other mining methods. Using such a mining technique, data miners can analyze past instances to forecast future events.

Today, data mining is one of the crucial techs businesses need to flourish in this dynamic and volatile consumer-inclined market. It leverages BI and advanced analytics that give organizations a bird’s eye view of evolving market trends, which helps in better strategic planning and optimized decision-making.

According to an April 2021 report by ReportLinker, the global data mining tools market stood at $634.7 million in 2020 and is estimated to reach $1.3 billion by 2027.

Data mining benefits are facilitated through tools essential for anomaly detection in analytics models, trends, and patterns, thereby avoiding the possibility of a system getting compromised in the worst cases.

These are the top ten data mining tools:

1. RapidMiner

RapidMiner is a data mining platform that supports several algorithms essential for **machine learning,** deep learning, text mining, and predictive analytics. The tool provides a drag-and-drop facility on its interface along with pre-built models that help non-experts develop workflows without the need for explicit programming in specific scenarios such as fraud detection.

Subsequently, developers can leverage the benefits of R and Python to build analytic models that enable trend, pattern, and outlier visualization. Moreover, the tool is further supported by active community users that are always available for help.

Pricing: Free and open source data science platform, wherein the free plan analyzes 10k rows of data.

2. Oracle Data Mining

The Oracle Data Mining tool is a part of 'Oracle Advanced Analytics' that creates predictive models and comprises multiple algorithms essential for tasks such as classification, regression, prediction, and so on.

Oracle Data Mining allows businesses to identify and target prospective audiences, forecast potential customers, classify customer profiles, and even detect frauds as and when they occur. Moreover, the programmer community can integrate the analytics model into BI applications using a Java API to see complex trends and patterns.

Pricing: Oracle provides a 30-day free trial to potential buyers.

3. IBM SPSS Modeler

IBM SPSS Modeler is known to fasten the data mining process and visualize processed data better. The tool is suitable for non-programmer communities that can exercise the interface's drag-and-drop functionality to build predictive models.

The tool enables the import of large volumes of data from several disparate sources to reveal hidden data patterns and trends. The basic version of the tool works with spreadsheets and relational databases, while text analytics features are available in the premium version.

4. Weka

Weka is an open-source ML tool written in JavaScript with a built-in framework for various ML algorithms. It has been developed by researchers at the University of Waikato in New Zealand. The tool offers an easy-to-use interface with additional features such as classification, regression, clustering, visualization, and much more. It allows users to build models crucial for testing ideas without writing code. This requires a good knowledge of the algorithms used for such purposes so that the appropriate one is rightly selected.

Weka tools were initially designed to explore the agricultural sector; however, today, it is being extensively used by researchers and scientists to explore the academic sector.

5. KNIME

KNIME is built with machine learning capabilities and an intuitive interface that makes modeling to production much more accessible. The KNIME tool provides pre-built components that non-coders can access to develop analytical models without worrying about a single line of code.

KNIME supports integration features that make it a scalable platform that can process diverse data types and advanced algorithms. This tool is crucial for developing business intelligence and analytics applications. In finance, the tool finds use cases in credit scoring, fraud detection, and credit risk assessment.

6. H2O

The H2O data mining tool brings AI technology into data science and analysis, making it accessible to every user. The tool is suitable for running several ML algorithms with features that support auto ML functions for the build and faster deployment of ML models.

H2O offers integration features through APIs available in standard programming languages and is suitable for managing complex datasets. The tool provides fully-managed options and the facility to deploy it in a hybrid setting.

7. Orange

Orange is a data science tool suitable for programming, testing, and visualizing data mining workflows. It is software that has built-in ML algorithms and text mining features, making it ideal for molecular scientists and biologists.

The tool provides an intuitive interface with add-on graphical features that make data visualization more interactive, such as sieve diagrams or silhouette plots. Moreover, the tool supports visual programming where non-experts in the domain can create models simply by using drag-and-drop interface features. At the same time, skilled professionals can rely on the Python programming language to develop models.

8. Apache Mahout

Apache Mahout is a data mining tool that enables the creation of scalable applications using ML practices. The tool is an open-source platform designed for researchers and professionals who intend to implement their own algorithms.

Apache Mahout is built on a JavaScript foundation on top of the Apache Hadoop framework, known for recommender engines, clustering, and classification applications. The tool can handle large datasets and is preferred by companies such as LinkedIn and Yahoo.

9. SAS Enterprise Mining

SAS Enterprise Miner is a data mining platform that helps professionals better manage data by converting large chunks of data into valuable insights. The tool provides an intuitive interface that aids in faster analytical model building and supports various algorithms that help in data preparation, essential for advanced predictive models.

SAS Enterprise Mining is well-suited for companies intending to implement **fraud detection** applications or applications that enhance customer response rates targeted through marketing campaigns.

10. Teradata

Teradata is a mining tool suitable for enterprises that rely on multi-cloud deployment setups. Such frameworks can easily access databases, data lakes, and even SaaS applications external to the enterprise. Moreover, with no-code deployment features, developing business models and analysing them to make informed decisions becomes more manageable.

Teradata is open to deployment on any public cloud platform such as AWS, Google, and Azure. Data miners can also deploy the tool in on-premise settings or a private cloud.

Data mining has opened up a sea of possibilities for companies by allowing them to improve and work on their bottom lines by identifying patterns and trends in business data. Mining techniques benefit every industry vertical, from retail, finance, manufacturing, insurance, and healthcare, to the entertainment and academic sectors.