ISSN:0975-9646

Data Mining and Knowledge Management Process

G Naveen[#1], K Radha[*2]
#B.TECH IV-CSE,GITAM University
Hyderabad,Sanga Reddy(D),Rudraram,India

**CSE, GITAM University*
Hyderabad, SangaReddy(D), Rudraram,India

Abstract— **Data mining is well known for acquiring knowledge and information from the larger databases has been recognised by many researches as a key topic in data mining and database systems and by many industrial companies as an import area with an opportunity of major revenues. Researchers in many different fields have some great interest in data mining. Several emerging applications in information providing services, such as data warehousing and online services over the internet, also call for various data mining techniques to better understand user behaviour, to improve the service provided and increase business opportunities. In response to such a demand, the article provides a survey, from the database researches point of view, on the data mining techniques developed recently, A classification of the available data mining techniques is provided and comparatively study of such techniques is presented.**

Keywords— **Data mining, Data mining tasks and functionalities of Data mining, Architecture of data mining, KDD, classification of data mining systems, Integration of data mining with database or data Warehouse System, Coupling System, Tools and techniques, Applications of Data Mining.**

I.INTRODUCTION

Data Mining refers to extracting or mining knowledge from large amounts of data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, database systems. The overall goal of the data mining process is to extract information from a data set and transform into an understandable structure for further use.

A. *DATA MINING TASKS:*

The data mining tasks can be classified generally into two types based on specific tasks performed those are: 1. Descriptive Tasks and 2. Predictive Tasks

The descriptive data mining tasks characterize the general information of the general properties of data whereas predictive data mining tasks perform inference on the available data set to predict how a new data set will behave.

B. *FUNCTIONALITIES OF DATA MINING:*

1.Concept /class Description: Characterization and discrimination, Data can be associated with classes or concepts. For example, in the electronics store, classes of the item for sale include computers and printers, and concepts of customers include big Spenders and budget Spenders.

Data characterization: Data characterization is a summarization of the general characteristics or features of target class of data.

Data discrimination: Data discrimination is a comparison of the general of the features of target class data object with the general features of object from one or a set of contrasting classes.

Mining Frequent Patterns, Associations and correlations: Frequent patterns, are patterns that occur frequently in data[6]. There are many kinds of frequent patterns, including itemset, subsequence, and substructures.

Association Analysis: Consider an organization , A Sales Manager, he decides to predict the frequently purchased items among the identical transactions.

Ex:
buy(X,"XeroxMachine")=buys(X,"Scanner")[%support=20%,confidence=80%]

Where, X is a variable representing a customer. Confidence=80% means that if a customer buys a computer, there is a 80% chance that she will buy software as well. Support=20% means that 20% of all of the transactions under analysis showed that computer and software were purchased together.

Classification and Prediction: Classification Main aim is to predict the unknown class labels [2].
"How is the derived model presented?"
Dervied model has various forms such as decision trees,neural networks,mathematical formulae,classification rules.

Decision Tree: A decision tree is a flow-chart like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and leaf nodes of trees,and these leaf nodes represent class distributions.

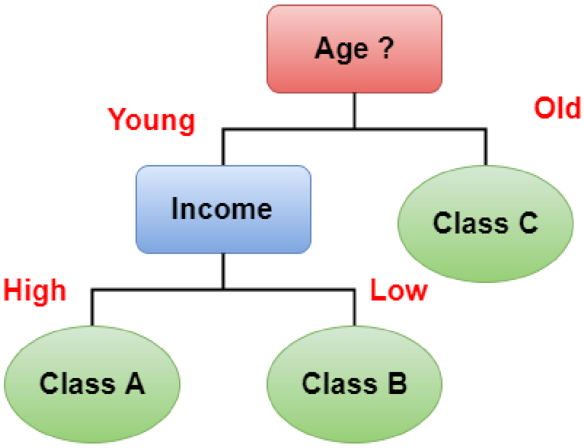

Fig:1. Decision Tree

Neural Network: A Neural network, when used for classification, typically a collection of neuron-like processing units with weighted connections between the units.

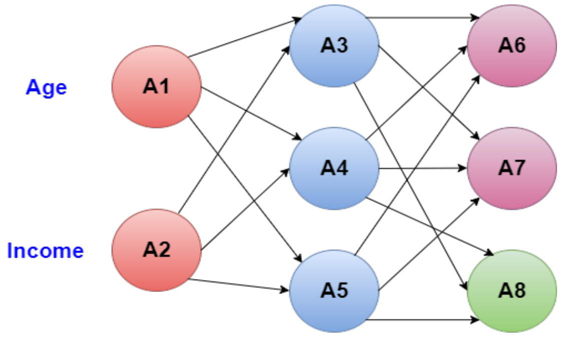

Fig:2. Neural Tree

Cluster Analysis: In classification and prediction analyse class-labelled data objects, where as in the cluster analysis, it analyses the data objects without a known values . The objects are grouped based on the principles of maximizing the intra-class similarity and minimizing the interclass similarity. That is, cluster of objects are formed so that objects within a cluster have a high similarity in comparison to one another but are very dissimilar to objects in other clusters.

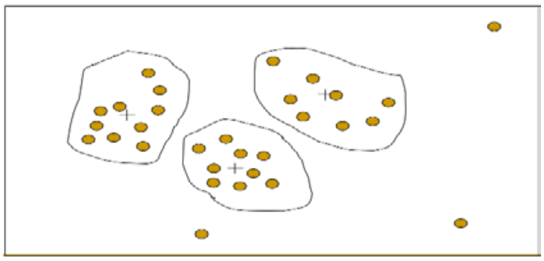

Fig:3. Cluster Analysis

Outlier Analysis: Database consists of data objects that do not follow with the data model. These data objects are outliers. Many of the data mining techniques will ignore the outliers in the datasets. Analysis of the outlier data is called as Outlier Mining.

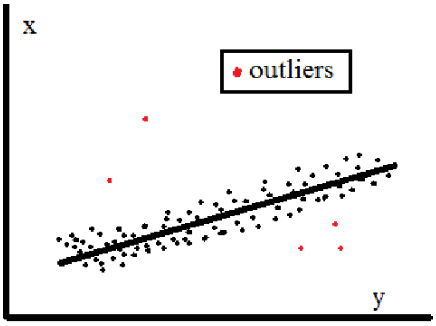

Fig: 4.Outlier Analysis

II.ARCHITECTURE OF DATA MINING:

Data Mining Architecture consists of the below mentioned components. [1]

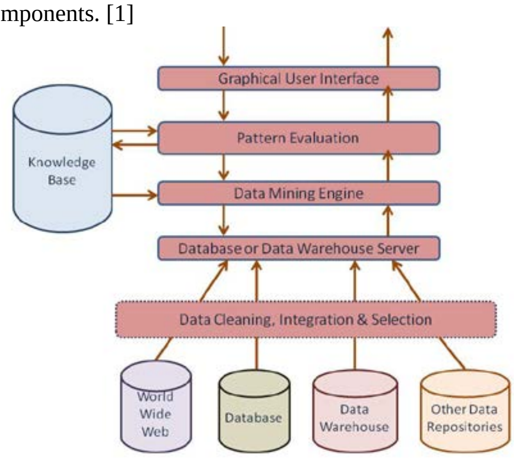

Fig: 5. Architecture of the Data Mining

1. **Knowledge Base:** This domain knowledge that is used to guide the search or evaluate the interestingness of resulting patterns. Such knowledge can include concept hierarchies, used to organise attributes or attribute value into different levels of abstraction. Knowledge such beliefs, which can be used to access a pattern's interestingness based on its unexpectedness, may also be included other examples of domain knowledge are additional interestingness constraints or, and metadata (e.g.: describing data from multiple heterogeneous sources).
2. **Data Mining Engine:** It is mainly used for predicting the various data mining techniques such as association analysis,evolution analysis,characterization and classification analysis.
3. **Pattern Evolution Module:** The pattern evolution module is mainly responsible for measure of interestingness of the pattern by using a threshold value.to focus on interesting patterns it communicates with the data mining engine
4. **4.Graphical User Interface:** The graphical user interface module communicates between the user and the data mining system.GUI helps the user to use efficiently without any difficulty. Whenever the user is posed a query to the data mining system, this data mining system generates results in an easier way to understand.
5. **Data warehouse or Database server:** It contains the actual data that is to be processed. Hence, the server is responsible for retrieving the relevant data based on the data mining request the user.

Summary: Components of Data Mining System has their own roles. To Successfully complete the data mining process and to communicate with each other , these components are used.

3. Classification of Data Mining Systems and Technologies are used: A data mining system can be classified according to the following criteria:

- Database Technology
- Statistics
- Information Science
- Visualization
- Machine Learning
- Other Disciplines

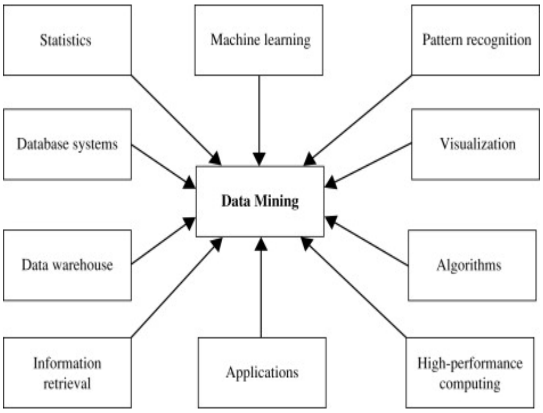

Fig:6. Data Mining Adopts techniques from many domains.

1. **Data warehouse:** To support the decision making process, a data warehouse uses the collection of data in an integrated manner and in subject oriented way.

- **Subject Oriented:** A data warehouse can be used to analyse a particular subject area. For example, sales can be a particular subject.
- **Integrated:** A data warehousing integrates data from multiple data sources. For example, source A source B may have different ways of identifying s product, but in a data warehouse, there will be only a single way of identifying a product.
- **Time-variant:** Historical data is kept in a data warehouse. For example, one can retrieve data from 3 months, 6 months,12 months, or even older data warehouse. This contrast with a transaction system, where often only the most recent data is kept. For example, a transaction system may hold the most recent

address of a customer, where a data warehouse can hold all address associated with a customer.

- **Non-volatile:** Once data is in the data warehouse, it will not change. Hence,from the datawarehouse past historical data cannot be altered.

2.Machine Learning:

By using these Machine Learning techniques, it is investigated that Personnel Computers can learned (or improve their performance) based on data. A main research area is for computer program to automatically learn to recognize complex, a typical machine learning problem is to program a computer so that it can recognize handwritten postal codes on mail automatically after learning from a set of examples. [3]

Machine learning is a fast-growing discipline. Here, we illustrate classical problem in machine learning that are highly related to data mining.

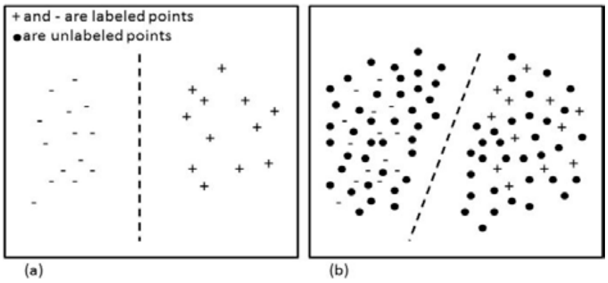

Fig: 7. Semi-supervised learning

Technologies used in machine learning are:

1. Supervised learning It is also termed as classification. In the training dataset , supervised learning is labelled. For example, in the postal code recognition problem, a set of handwritten are used as the training examples, which supervise the learning of classification model.

2.Unsupervised learning It is also called as clustering. The learning process is unsupervised in that the input examples are not class-labelled. Clustering is used to find the classes within the data. In an unsupervised learning technique, Set of images of handwritten digits can be used .this can be used to 10 clusters.These clusters belongs to o to 9 different digits. However, since the training data are not labelled, the learning model cannot tell us the semantics meaning of the cluster found.

3.Semi-Supervised learning It is a class of machine learning technique that makes use of both labelled examples for learning a model. In one approach, labelled examples are used to learn class model. In examples are used to refine the boundaries between classes. For a two-class problem, we can think of set of examples belonging to one class as the positive examples, those belonging to the other class as the negative examples in the above figure like noise or outliers.

4.Active Learning is a machine learning approach that lets users play an active role in the learning process. An example, which may be from a set of unlabelled examples or synthesized by the learning program. The main puropose is to optimize the model quality by acquiring the knowledge from the users based on the connstraints that have been labelled.

We can see that data mining and machine learning shares a lot of similarities. For classification and clustering tasks, machine learning research often focuses on the accuracy of the model.

4. Knowledge Discovery in database (KDD):

Some of them will treat data mining is same as Knowledge discovery process. Here is the list of steps involved in knowledge discovery process: [7]

1. **Data Cleaning:** Data Cleaning removes the noise data and inconsistent data.

- Cleaning in case of missing values.
- Cleaning noisy data, where noise is random or variance error.
- Cleaning with data discrepancy detection and data transformation tools

2. **Data Integration:** Data Integration merges the numerous data sources together.

- Data integration using Data Migration tools.
- Data integration using Data Synchronization tools.
- Data integration using ETL(Extract-Load-Transformation) process.

3. **Data Selection:** Data Selection process analyses the task relevant data and this task relevant data is retrieved from the database.

- Data selection using Neural network.
- Data selection using Decision trees.
- Data selection using Clustering, Regression, etc.

4. **Data Transformation:** In this step data are transferred or consolidated into forms appropriate for mining by performing summary or aggregation operations.

- **Data Mapping:** Assigning elements from source base to destination to capture transformations.
- **Code generation:** Creation of the actual transformation program.

5. **Data Mining:** In this step intelligent methods are applied in order to extract data pattern.

- Transforms task relevant data into patterns.
- Decides purpose of module using Classification or characterization.

6. **Pattern Evolution:** In pattern Evolution ,data is represented in the form of patterns.

- Find interestingness score of each pattern.
- Uses Summarization and Visualization to make understandable by user.

7. **Knowledge Presentation:** In this step, knowledge is represented.
    - Generates reports.
    - Generates tables.
    - Generates discriminant rules, classification rules, characterization rules etc.

****Note:**

- *KDD is an iterative aspect .It evalution can be measured to enhance ,refine,to integrate the data and data transformation to get different and more appropriate results.*
- *Pre-processing of databases consists of Data cleaning and Data integration.*

Above figure 8. depicts the process of knowledge discovery process:

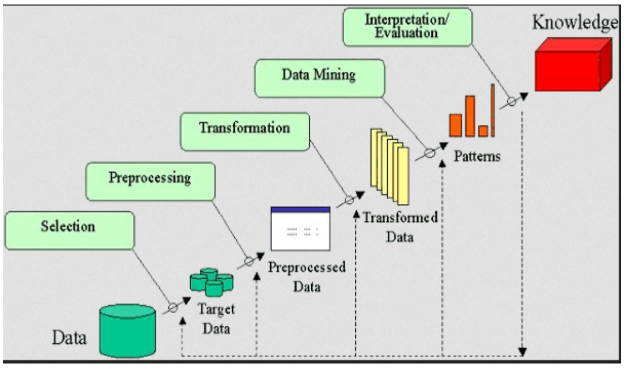

Fig: 8.Knowledge Discovery in database (KDD)

5. Data Mining Tools And Techniques:

5.1. Data mining Techniques are: Extracting important knowledge from a very large amount of data can be crucial to organizations for the process of decision making. Some data mining techniques are: [5]

1. Association
2. Classification
3. Clustering
4. Sequential patterns
5. Decision Tree

1. Association Technique: Association helps to find out the pattern from huge data, based on the relationship between two or more items of the same transaction. The association technique is used to analyse market means it helps us to analyse people's buying habits

Example: you might identify that a customer always buys ice cream whenever he comes to watch movie so it might be possible that when customer again comes to watch movie, he might desire to purchase an ice cream again.

2.Classification Technique: Classification technique is most common data mining technique. In classification method we use mathematical technique such as decision trees, neural network and statistics in order to predict unknown records. This method derives important information about the data.

Let assume you have set of records, each record contains a set of attributes and depending upon these attributes you will be able to predict unseen or unknown records. Consider an organization, predict the list of employees who left the organization and who may leave the organization in the future with classification technique.

3. Clustering Technique: Clustering techniques used in the process of data mining. The main aim of clustering technique is to makes cluster (groups) from piece of data which shares common characteristics. Clustering technique helps to identifies the differences and similarities between the data.

Take an example, A shop in which many items are for sales, now the challenge is how to keep those items in such a way that a customer can easily find his required item. By using the clustering technique, you can keep some items in one corner that have some similarities and other items in another corner that have some different similarities.

4. Sequential Patterns: Sequential patterns are a useful method for identifying trends and similar patterns.

For example, In cluster data you identify that a customer buys particular product on particular time of year, you can use this information to suggest customer these particular product on that time of year.

5. Decision Tree: Decision tree is one of the most common used data mining techniques because its model is easy to understand for users. In decision tree you start with a sample questions which has two or more answers. Each answer leads to a further two or more question which helps us to make a final decision. Question can be formed with the help of the root node of decision tree.

Take example of flood warning system:

First check water level, if water level is >50ft then alert is sent. If water level is<50ft then check water level if water level is >30ft then send warning and if water is <30ft then water is in normal range.

5.2. Data mining tools and techniques are as follows:

Rapid Miner, Orange, Weka, KNIME, Sisense, SSDT, Apache Mahout, Oracle Data mining, Rattle, Data Melt, IBM Cognos, IBM SPSS Modeler, SAS Data Mining, Teradata, Board, Dundas BI, Python, Spark, and H_20. [5]

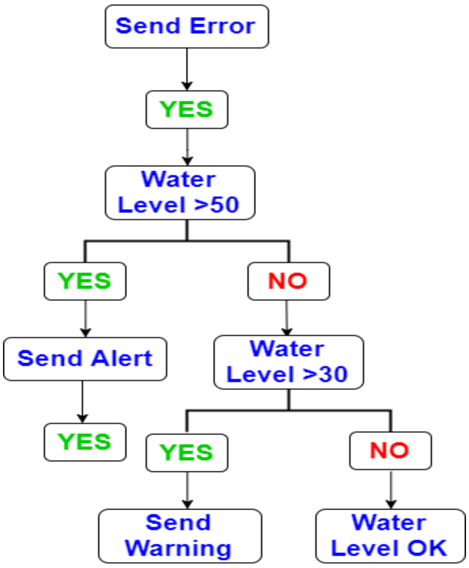

Fig: 9. Decision Tree for Flood Warning System.

Source: https://d2h0cx97tjks2p.cloudfront.net/wp-content/uploads/sites/2/2018/02/Data-Mining-Tools-01-1.jpg

Fig 10: Data Mining Tools

6. Integration of data with Database or Data Warehouse System:

Data Integration is a data pre-processing technique that combines data from multiple sources and provides user a unified view of these data. [7] If the data mining system is not integrated with any database or data warehouse system, then there will be no system to communicate with. This scheme is known as coupling scheme. In this scheme the main focus is on data mining design and for developing efficient and effective algorithms for mining the available data sets.

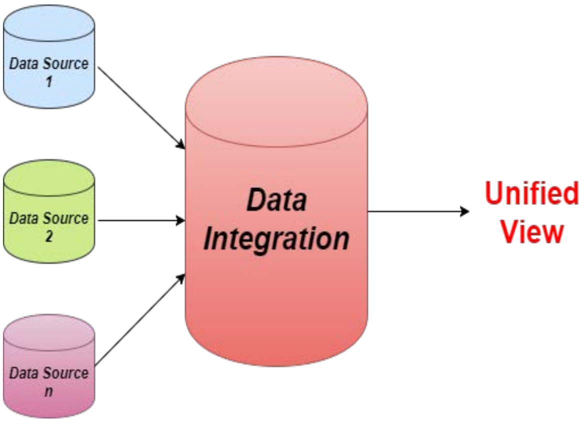

Fig: 10. Data Integration in Data mining.

The list of Integration scheme is as follows:

No Coupling: It will not use any of the data mining ,database or data warehouse functions. It fetches the data from a particular source and process that data using mining algorithms. Results of data mining system will be stored into a file.

Loose coupling: In this scheme, it uses the techniques of data mining system, database and data warehouse system. It fetches the data from the data respiratory managed by these systems and perform data mining on the data and then stores the mining result either in a database or in a data warehouse.

Semi-tight Coupling: In this scheme, the data mining system is linked with a database or a data warehouse system and in addition to that, in the database, efficient implementations of the few data mining primitives can be provided.

Tight Coupling: it is an integration of database with the data mining or data warehouse system. The functional component of a data mining system is data mining sub system.

Applications of Data Mining:

1. **Communications:** Data mining techniques are used in communication sector to predict customer behaviour to offer highly targeted and relevant campaigns [4].
2. **Insurance:** Data mining helps insurance companies to price their products are profitable.They promote a new offers to the existing customers and new customers.
3. **Education:** Data mining benefits educators to access student data, predicts achievements levels and find students or groups of students which need extra attention. Consider that some of the students have less knowledge in their respective subjects such as Information Retrieval Systems.
4. **Manufacturing:** With the help of Data mining Manufacturers can predict wear and tear of production assets. They can anticipate maintenance which helps them reduce them to minimize downtime
5. **Retail:** Data Mining Techniques helps retail malls and grocery stores identify and arrange most sellable items in the most attentive positions. It helps store owner to come up with the offer which encourages customers to increase their spending.
6. **Banking:** Data mining helps finance sector to get a view of market risks and manage regulatory compliance. To issue new loans, credit cards banks need to identify the defaulters, etc.
7. **Service providers:** Service providers like mobile phone utility industries use Data Mining to predict the reasons when a customer leaves their company. They analyse billing details, customer services interactions, complaints made to the company to assign each customer probability score and Offers incentives.
8. **E-commerce:** E-commerce websites use Data mining to offer cross sells to predict their website
9. **Super Markets:** Data mining allows supermarket's develop rules to predict if their shoppers were like to be expecting. They can start target is to start targeting products like: baby, baby shop, diapers and can start
10. **Crime Investigation:** Data mining helps crime investigation agencies to deploy police workforce (When is a most crimes likely to happen and when?) who to search at a border crossing.
11. **Bioinformatics:** Data mining helps to mine biological data from massive datasets gathered in biology and medicine.

Issues of Data Mining:

Data mining is not that easy, the algorithm used are very complex. The data is not available at one place it needs to be integrated from the various heterogeneous data sources. These factors also create some issues. Here we will discuss the major issues regarding: [8]

1. Mining methodology and user interaction
2. Performance issues
3. Diverse data type issues

The following below are some issues:

1. Mining methodology and user interaction issues:

- Mining different kinds of knowledge in database: The need of different users is not the same, and Different users may be in interested in different kind of knowledge. Therefore, it is necessary for data mining to cover broad range of knowledge discovery task.
- **Interactive mining of knowledge at multiple levels of abstraction: Data mining process is interactive and it allows the users to focus on search patterns and refines data mining requests based on the generated results.**
- **Incorporative of background knowledge:** To guide discovery process and to express the discovered patterns, the background knowledge can be used. Background knowledge may be used to discovered patterns, the discovered patterns not only in concise terms but at multiple level of abstraction.
- **Data mining query languages and hoc data mining: Data Mining query language gives the permission to the users to describe the adhoc mining tasks and integrates with a** Data mining query language .It Produces flexible and efficient data mining.
- **Visualization and Presentation of data mining Results: Discovered patterns are expressed in a high level language and it is displayed in** visual representations to easily understandable by the users.
- **Handling noisy or incomplete data: To handle the incomplete objects, noise and mining with regularities data cleaning techniques are used.It maintains the Data Quality.** If the data cleaning methods are not there the n the accuracy of the discovered patterns will be poor.
- **Pattern evolution:** It refers to interestingness of the problem, the pattern discovered should be interesting because either they represent common knowledge.

2. Performance issues:

- **Efficiency and scalability of data mining algorithms: To extract the information from massive databases , data mining algorithms must be applied .These are efficient and scalable.**
- **Parallel, distributed and incremental mining algorithm:** The factors such as huge size of databases, wide distribution of data, complexity of data mining methods motivates the development of parallel and distributed data mining algorithm. These algorithms dived the data into partitions which is further processes parallel. Then the result from the partitions is merged. The incremental algorithms are used to update the databases .These databases may contain the noisy data.to clear the noisy data ,these incremental algorithms are used.

3.Diverse data type issues:

- **Relational and Complex types of data:It contains the multimedia data objects,temporal data,cmplex data objects and spatial data.** It is not possible for one system to mine all these kinds of data.
- **Mining information from heterogeneous database and global information systems:** The data is available at different data sources on LAN or WAN. These data

sources may be structured, semi structured or unstructured. Therefore, mining knowledge from them adds challenges to data mining

Trends of data mining:

Here are some trends in data mining that reflects pursuit of the challenges such as construction of integrated and interactive data mining environments, design of data mining languages:

- Focus on Multimedia
- Applications Exploration
- Artificial intelligence and IoT
- Data mining based on meta data
- Scalable and interactive data mining methods
- Interaction 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.
- Location based data mining
- Mobile data mining

TABLE 1: Data Mining Trends Comparative Statements.

Data Mining trends	*Algorithms/ Techniques used*	*Data Formats*	*Computing Resources*	*Prime areas of application*
Past	Statistical, Machine Learning Techniques	Numerical data and structured data stored in traditional databases	Evolution of 4G PL and various related techniques	Business
Current	Statistical, Machine Learning, Artificial Intelligence, Pattern Reorganization Techniques	Heterogeneous data formats include structured, semi structured and unstructured data	High speed networks, High end storage devices and Parallel, Distributed computing etc…	Business, Web, Medical diagnosis etc.
Future	Soft computing techniques like Fuzzy logic, Neural networks and Genetic Programming	Complex data objects include high dimensional, high speed data streams, sequence, noise in the time series, graph, multi instance objects, multi represented objects and temporal data etc..	Multi-agent technologies and cloud computing	Business, web, Medical diagnosis, Scientific and research analysis fields (bio, remote sensing etc..), social networking etc.

References

[1]. Jaanu Sharma,”A survey on the Data Mining Architecture”, International Journal of Computer Engineering and Applications, Volume XII, I,jan.18,www.ijcea.com. ISSN 2321-3469.

[2]. Mrs.S.P.Deshpande, Dr V.M.Takare, ”Data Mining System and Applications: A Review”, International Journal of Distributed and Parallel Systems (IJDPS) Vol.1.I,No.1,September,2010.

[3]. Shu-Hsien Liao, Pei-Hui-Chu, Pie-Yuan Hsaio,”Data mining technique and applications A decade review from 2000-2011”, Expert System with Applications 39(2012) 11303-11311.

[4]. https://www.guru99.com/data-mining-tutorial.html#11 (Applications of data mining).

[5]. https://data-flair.training/blogs/data-mining-tools-techniques/(Tools of data mining).

[6]. www.lastnightstudy.com/Show?id=37/Data-Mining-Functionalities(Fuctionalities of data mining).

[7]. https://www.tutorialspoint.com/data_mining/dm_systems.htm (Data Integration in data mining).

[8]. http://www.tutorialspoint.com/dm/dm_issues.htm (issue of data mining)