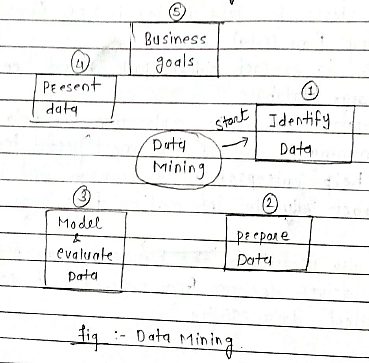
1. **What is data Mining**
2. Data mining is collection of useful data from huge data to identify patterns.
3. Data mining is a process of sorting through large sets to identify patters and relationship that can help. Solve business problems through analysis.
4. Data mining techniques and tools unable enterprises to predict future trends and make more informed business decisions.
5. It is the process of discovering of mining knowledge from a large amount of data.
6. Another term for data mining is KDD (knowledge Discovery from data)
7. Attempts to extract hidden patterns and trends from large database.
8. Also support automatic exploration of data
9. Data mining is one of the most useful techniques that help entrepreneurs researchers and in individuals to extract valuable information from huge set of data.
10. It is also useful for the business to take the data driven decision from huge set of data is called data Mining.
11. finding hidden information in the database
12. Called as axploitarly data analysis data driven
13. Deductive learning

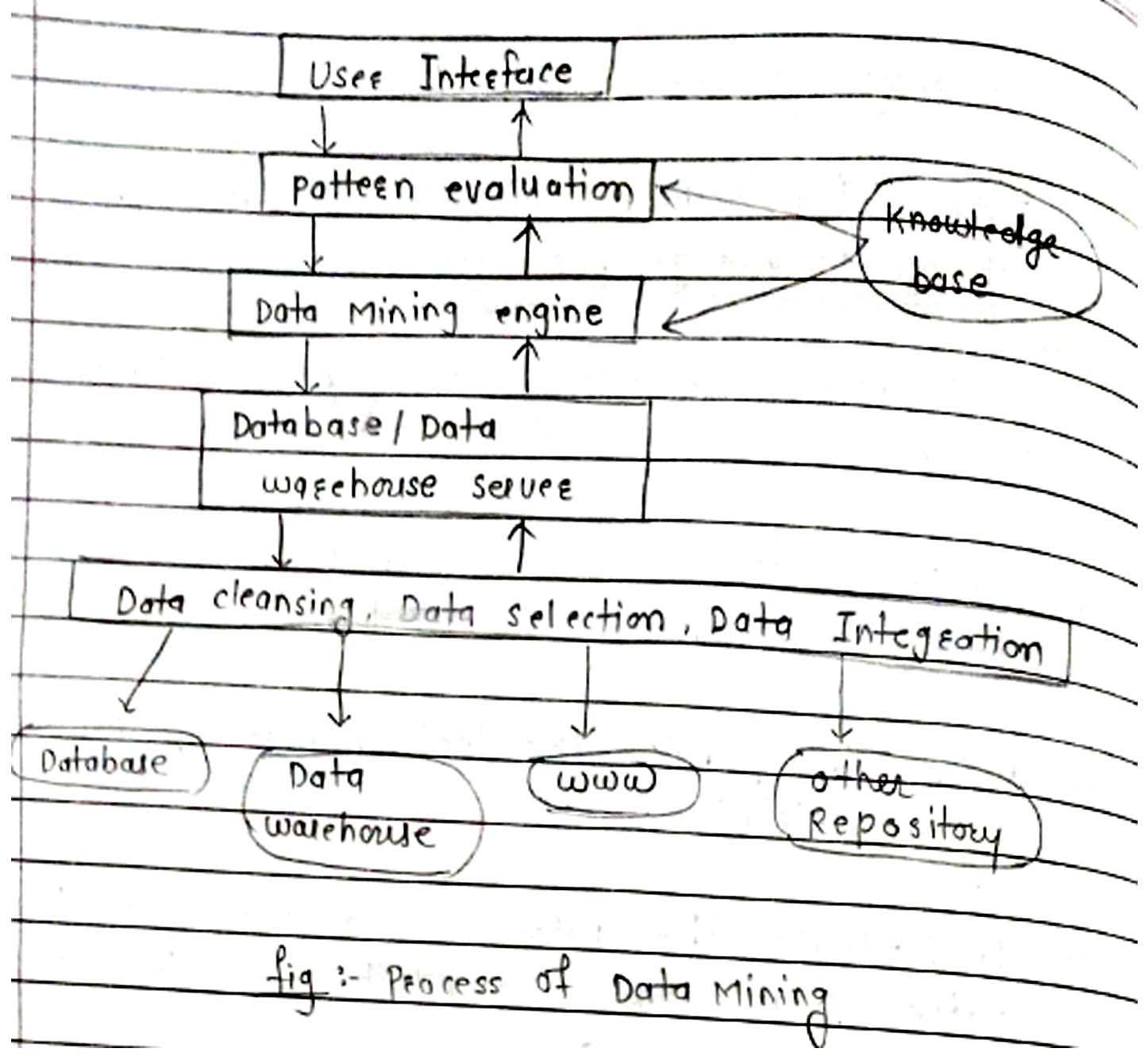
(data driven is data fetching).

1. It is also useful for the business to take the data driven decision from huge set of data is culled data mining
2. finding hidden information in the database
3. called as exploitatery data analysis data driven and reductive learning



* **Advantages of Data Mining :**

1. Data mining techniques enables organization to obtain knowledge basedata.
2. Data mining enables organizations to make modification in operation & production
3. Compare with others statistical data applications data mining is cost efficient
4. Data mining helps the decision making process of an organization
5. If fascinates the automated discovery of hidden patterns as well as predictions of trends and behaviors.
6. It can be induced in the new system as well as existing platform
7. It is quick process that makes it easy for new user to analyze amount of data in a start time.



* Data mining application

1. Healthcare
2. Manufacturing Engineering
3. Education
4. Financial banking
5. Market basket analysis (D.mart)
6. Fraud Detection

* Motivating challenge

There are various challenge of data mining which are as follows

1. Efficiency and scalability of data mining Algorithm:

It can efficient attact large amount of data in database. The knowledge discovery algorithm should be efficient and scalable to huge database, specifically the running time of data mining algorithm should be predictable and expectable in huge database. Algorithm with exponential or even channel order polynomial complexity will not be efficiently use.

1. Usefulness, certainly and expressiveness of data Mining Results

The identified knowledge should exactly property the content of database & be beneficial for specific application the imperfectness must be defined by measures of uncertainty in form of approximate rules or quantitative rules Noise and exceptional data must be managed elegantly in data mining system.

This also simulate a systematic study of measuring quality of the discover knowledge such as reliability of by the development of statistical, analytical and simulative models and tools.

1. Expression of Various kind of data mining Results :

Several kind of knowledge can be discovered from huge amount of data. It can also like to examine discovered knowledge from multiple view and display them in different forms.

This needed us to define data mining request and discovered knowledge in high level languages or GUI, So that data mining task can be defined by non – experts and the discovered knowledge can be understandable and precisely available by users.

This also needed discovery system to select expressive knowledge representation techniques.

1. Interactive mining knowledge at multiple abstraction levels :

Because it is complex to predict what exactly can be discovered from a database, a high level data mining query must be considered as a prob that can disclose some interesting traces for further exploration.

Interactive discovery must be encourage which involves a user to interactive refine a data mining request, dynamically changed data focusing progressively defer a data mining process and flexibly view the information and data mining results at several abstraction level and from multiple angles.

1. Data mining information from different sources of data:
2. The broadly available local and wide area computer Network such as internet can connect various sources of data and from large, huge distributed databases mining knowledge from multiple sources of formatted or unformatted information with divers data semantics process a new requirement of data mining data mining can help disclose high level data regulatories in heterogeneous databases which can barely discoverable by simple query systems.

The huge size of Database, the broad distribution of data and the computational complexity of several data mining method motivated advancement of parallel and distributed data mining algorithems.

* Origins of Data Mining

By the goal of meeting the challenges of the previous section, researches from different disciplines began to focus on developing more efficient and scalable tools that could handle diverse types of data this work which culminated in the field of data mining, Built upon the methodology and algorithms that researches had previously used. Data mining draws upon such as.

1. Sampling

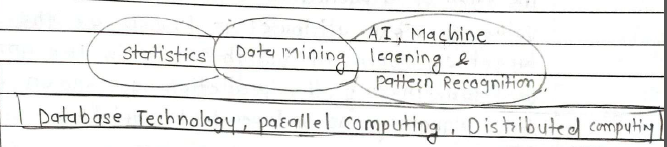
2. Estimation

3. By pothesis testing from statistics

1. Search algorithms, Modeling techniques and learning theories from artificial intelligence and Machine learning.
2. Data mining has also been quik to adopt ideas from other areas, including optimization, evolutionary computing, information theory, single processing, visualization, and information retrieval.

A number of other areas also play key supporting roles. In particulars database systems are need to provide support for efficient storage. Indexing, and query processing. Techniques from high performance (parallel) computing are often important in addressing the massive size from data sets.

Distributed techniques can also help address the issue of size & are essential when the data cannot be gathered in one location.



* Data Mining Tasks

Data Mining tasks are generally divided into two major categories.

1) predictive tasks

2) Descriptive tasks.

1) Predictive tasks

The objective of these tasks is to predict the value of a particular attributes based on the value of other attributes is known as the target or dependent variables while the attributes used for making the prediction are known as the explanatory or independent variables.

2) Descriptive tasks

The objective is to drive patterns (correlations) trends, clusters, trajectories, & anomalies) that summaries the underlying relationship in data. Descriptive data mining tasks are often exploratory in nature and frequency require host processing techniques to validity and explain the results.

1. **Predictive Modeling:-**

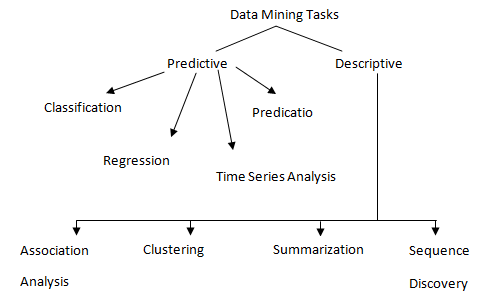
Refers to the task of Building a Model for the target variable as a function are type of predictive Modeling.

1. Classification: - Which is use discrete target variable.
2. Regression: - Which is used for continuous target variable.
3. Time series analysis : -
4. Prediction : -

For example :- Predicting whether a web user will make a purchase at an online bookstore is a classification task because target variable is binary valued on the other hand for casting the further price of a stock is a regression task because price is a continues valued Attributed.

1. **Association analysis : -**

Is used to discover patterns that describe strongly as associated features in the data. The discovered patterns are typically represented in the form of Implication rules or feature subsets. Useful applications of Association analysis include finding group of genes that have related functionality. Identifying web pages that are accessed together, or understanding the relationships between elements of Earth’s climate system.



**Note:-** Predication Method – use some variable to predicate unknown or further value of others variables.

1. **Anomaly detection: -**

Is the task of Identifying observations whose characteristics are significantly different from the rest of the data. Such observations are known as anomalies or outliers the goal of an anomaly detection algorithm is to discover the real anomalies and avoid falsely labeling normal object as anomalies. Applications of anomaly detection include the detection of fraud, Network intrusions, unusual patterns of disease, and ecosystem disturbance.

**Description Method : -**

Find human intrafable pattern that describe the Data.

1. **Cluster Analysis : -**

Cluster analysis seeks to find groups of closely related observation so that observations that belong to the same cluster are more similar to each other than observations that belong to other clusters.

Clustering has been used to group sets of related customers, find areas of the ocean that have a significant impact on the Earth’s climate and compress data.

**Data**: -

1. **The Type of Data : -**

Data sets Differ in a Number of ways.

**For ex : -** The Attributes used to describe data objects can be of different types

1) quantitative

2) qualitative

1. Data sets may have special characteristics e.g. Some data sets contain time series or objects with explicit relationships to one another.
2. Not surprisingly, The type of data determines which tools and techniques can be used to analyze the data.
3. Furthermore, New research in data Mining is often driven by the need to accommodate new application areas and their new types of Data.
4. **The Quality of the Data : -**

The Data is often far from perfect while most data mining techniques can tolerate some level of Imperfection in the data, a focus on understanding and improving data quality typically improves the quality of the resulting analysis..

1. **Preprocessing steps to make the Data more suitable for data Mining : -**

The Raw data must be processed in order to make it suitable for analysis while one object may be to improve data quality other goals focus on modifying the data so that it is better fils a specified data Mining technique or tools.

1. **Analyzing Data in terms of its Relationships : -**

One approach to data analysis is to find Relationships among the data objects and then perform the remaining analysis using these relationship rather than the data objects. Themselves.

Attributes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Tid | Refund | Martial | Income | Cheat |
| 1 | Yes | Single | 12 k | No |
| 2  Objects | No | Married | 15 k | Yes |
| 3 | No | Single | 6 k | No |
| 4 | Yes | Single | 8 k | Yes |
| 5 | No | Married | 9 k | No |

1. **What is Data?**

1) Collection of Data objects and their Attributes.

2) And Attributes it is a property or characteristic of an objects.

3) A collection of Attributes disaibean a object.

**Types of Data**

A data set can often be viewed as a collection of data objects. Other names for a data object are record point, vector, pattern, event, case, sample, observation or entity.

Data object are described by a number of attributes that capture the basic characteristics of an object, such as the mass of physical object or the time at which an event occurred. Other names for an attributes are variable, characteristic, field feature or dimension.

**Attributes &Measurement** : -

1. **What is an Attributes : -**

An attribute is an property or characteristic of an object that may vary, either from one object to another or from one time to another.

For ex : - eye colour varies from person to person, while temperature of an object varies over time Eye colour is an symbolic attributes with a small number of possible values while temperature is a numerical Attribute with a potentially unlimited number of value.

1. **What is Measurement?**

“A Measurement scale is a ruel that associate a numerical or symbolic value with an attribute of an object.

The process of an Measurement is a application of a Measurement scale to associate a value with a particular value of a specific object.

**Types of an Attribute** :-

1. Employee age and ID number
2. Length of line segment

A useful way to specify the type of an attribute is a to identify the properties of numbers that correspond to underlying properties of the attribute.

**For ex :-** An attribute such as length has many of the properties of numbers are typically used to describe attribute.

1) Distinctness = &

2) Order <, ≤, >, &≥

3) Addition + & -

4) Multiplication and/

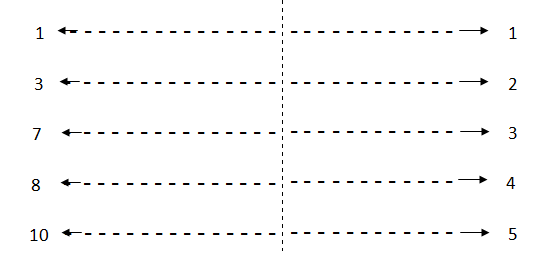
**Given this properties, we can define four types of attribute.**

1) Nominal 3) Interval

2) Ordinal 4) Ratio

Each attribute type prossesses all of the Property and operations of the attributes type above it consequently any property or operation that is valid for nominal, ordinal and interval attributes is also valid for ratio attribute.

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute Type | Description | Example | Operations |
| 1) Nominal | The value of a nominal attribute are just different names i.e. nominal values provide only enough information to distinguish one from another (=, ) | Zip codes, employee ID No. eye color gender | Mode, entropy, contingency, correlation,  X2 test |
| 2) ordinal | The values of an ordinal attribute provide enough information to order objects (<, >) | Hardness of mineral, (good, better, best) grades, street numbers. | Median, Percentiles, rank correlation, run tests, Sign test, |
| 3) Interval | For interval attributes the difference between values are meaningful  l.e, a unit of measurement exists (+, -) | Calender dates, temp. in clesius or fahren beit | Pea ran’s correlation, T & F tests, mean, standard deviatia |
| 4) Ratio | For Ratio variables, both differences and ratio are meaningful (\*, /) | Temp in Kelvin, electrical current, length, age, mass counts | Geometric mean, harmonic mean, percent, variation |



|  |  |
| --- | --- |
| A mapping of length to numbers That  captures only the order properties of length | A mapping of length of numbers That  captures both the order and additivity  properties of length. |

The way to Measure & Attributes is some what may not match the Attribute property.

* **Describing Attributes by the number of values :-**
  1. **Discrete** :- A discrete attribute has a finite or countable infinite set of values such attributes can be categorical such as zip codes or ID number Binary attribute are special case of discrete attribute and assume only two value.

Ex.:- True or false, Yes or No, 0 & 1

* 1. **Continuous : -** A continuous attributes is 1 whose values are real number

Ex :- amples include attributes such as temperature, height or weight continuous attributes are typically represented as floating-point variables. Partially, real values can only be Measured and represented with limited precision.

* **Types of Data sets :-**

There are many types of data sets, and as the field of Data mining develops and matures a greater variety of data sets become available for analysis we have the group type of data sets into three sets.

* 1. Record data
  2. Graph – Base data

1. Ordered data.

* **General characteristics of Data sets :-**

1. **Dimensionality: -** The Dimensionality of Data sets is a number of Attribute that the objects in the data sets possess Data with a small number of dimensions tends to be qualitatively different than moderate or high dimensionally data.
2. **Sparsity :**- Sparsity is a advantage because usually only the non-zero value need to be stored and Manipulated this results in significant saving with respect to computation time and storage some data mining algorithms work well sfor sparse data.
3. **Resolution :-** It is frequency possible to obtain data at different levels of resolution the pattern in the data also depends on the level of resolution is to fine a pattern may not be visible if the resolution is too coarse the pattern may disappear.
4. **Record data :**- Data sets is a collection of record (Data objects) each of which consist of a fix set of data fields (Attribute) for the most Basic form of Record data there is no explicit relationship among records or data fields. Record data visually stored either in flat files or in relational data bases. Relational Database are certainly more than a collection of records but data mining often does not use only of the additional information available in a relational data convenient place to find records.

* **The Data Matrix :-** If the Data object the collection of data all have a the same fix set of numerious

Attributes then the data objects can be thought of as points in a multidimensional space, a set of such data objects can be interpreted as an ‘m’ by ‘n’ matrix, where there are ‘m’ rows one of each object and ‘n’ colums one of each attributes.

This matrix is called Data Matric.

* **The sparse Data Matrix :-** The sparse data matrix is special case of data matrix in which the attributes are of the same types and are symmetric transition data is a example of sparse data matrix that has only ‘0’ or ‘1’ entrys.

**(a) Record Data**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Tid** | **Refund** | **Material Status** | **Taxable Income** | **Defaulted Berrours** |
| **1** | Yes | Single | 12 k | No |
| **2** | No | Married | 16 k | No |
| **3** | Yes | Married | 5 k | Yes |
| **4** | No | Single | 7 k | No |
| **5** | Yes | Single | 25 k | Yes |

**(b) Transaction Data**

|  |  |
| --- | --- |
| **Tid** | **Items** |
| **1** | Bread, Soda, Milk |
| **2** | Beer, Bread |
| **3** | Beer, Soda, Diaper |
| **4** | Beer, Diaper, Milk |
| **5** | Soda, Milk |

**(c) Data Matrix**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Projection of X load** | **Projection of Y load** | **Distance** | **load** | **Thickness** |
| 10 . 23 | 5 . 27 | 15 . 22 | 27 | 1.2 |
| 12 . 23 | 0 . 25 | 16 . 22 | 22 | 1.1 |
| 13 . 64 | 7 . 23 | 17 . 34 | 23 | 1.2 |
| 14 . 27 | 8 . 43 | 18 . 45 | 25 | 0.9 |

**(d) Document term matrix**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Team** | **Coach** | **Play** | **Ball** | **Score** | **Game** | **Win** | **Test** | **Timeout** | **Season** |
| **Document 1** | 3 | 0 | 5 | 0 | 2 | 6 | 0 | 2 | 0 | 2 |
| **Document 2** | 0 | 7 | 0 | 2 | 1 | 0 | 0 | 3 | 0 | 0 |
| **Document 3** | 0 | 1 | 0 | 0 | 1 | 2 | 2 | 0 | 3 | 0 |

1. **Graph Base Data :-**

A Graph Base can sometimes be a convenient and powerful representation for data.

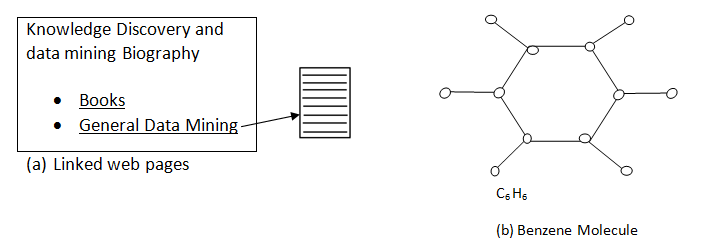
1. The graph captures relationship among data objects.
2. The data objects themselves are represented as graph.
3. **Data with relationships among objects :-**

The relationship among objects frequently convey important information the data objects are mapped to nodes while the relationships among objects are captured by the links between objects and link property such as direction & weight consider web pages or the world wild web (www) which contain Blot text and links to and process.

In order to process search engines collect and process web pages o extract their contents.

1. **Data with objects that are graph : -**

If object have structure that is the object contain subjects that have relationships then such object are frequently represented as graph a graph representation make its possible to determine which sub structure occur frequency in set of compound.



* **Ordered Data :-**

For some type of data the Attributes that Relationship that ordered in time or space.

1. **Sequential Data :**

Sequential data also referred to as temporal data can be through of an extension of record data were each record has a time associated with it.

|  |  |  |
| --- | --- | --- |
| **Time** | **Customer** | **Items Purchased** |
| **t1** | **C1** | **A, B** |
| **t2** | **C3** | **A, C** |
| **t2** | **C1** | **C, D** |
| **t3** | **C2** | **A, D** |
| **t4** | **C2** | **E** |
| **t5** | **C1** | **A, E** |

|  |  |
| --- | --- |
| **Customer** | **Time and items Purchased** |
| **C1** | **(t1 :A, B) (t2 : C, D) (t5 : A, E)** |
| **C2** | **(t3 : A, D) (t4 : E)** |
| **C3** | **(t2 : A, C)** |

* + 1. Sequential transaction Data.

1. **Sequence Data :**

It quite similar to sequential data except that there is no time stamps instead there are positions in an ordered sequence.

1. **Time series Data :**

Time series data is a special type of sequential data in which each record is a time series i.e., a series of Measurements taken over time.

1. **Spatial Data :**

Some objects have spatial attributes, such as positions or areas, as well as other types of attributes an example of spatial data is weather data (Precipitation, temperature, pressure) i.e., collected for a variety of geographical location.

* **Data Quality :-**

Data Mining application are often applied to data that was collected for an other purpose, or for further use.

1. The detection and correction of data quality problems.
2. The use of algorithm that can tolerate poor data quality. The first step detection and correction, is often called data cleaning.
3. **Measurement and Data collection Issues :-**

It is unrealistic to expect that data will be perfect their may be problems due to human error, limitations of Measuring devices or flaws in the data collections process values or even entire data objects may be missing their may be duplicate objects, we begin with a define of measurement and data collection error and then consider a variety of problem that involve measurement error.

1. Noise
2. Artifacts
3. Bias
4. Precision
5. Accuracy

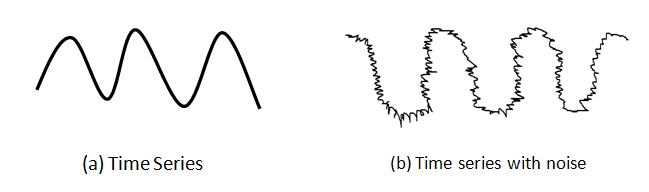
We conclude by discussing data quality issues that may involve both measurement and data collection problems.

1. Outliers
2. Missing & Inconsistent data
3. Duplicate data
4. **Measurement and Data Collection Errors : -**

The term measurement error refers to any problems resulting from the measurement process a common problem is that the value recorded differs from the true value to some extent. The term data collection error refers to error such as omitting data objects or attributes values or inappropriately including a data object.

1. **Noise and artifacts -**

Noise is the Random component of a measurement error it may involve the distraction(Break) of a value if a bit more noise where added to the time series it shape will be lost bit.



1. **Precision –**

The closeness of repeated measurement to one another.

1. **Bias –**

A systematic variation of Measurement from the quantity being measured.

1. **Accuracy –**

The closeness of Measurements to the true value of the quantity being Measure.

1. **Outliers –**

Outliers are either (1) Data objects that in some sense have characteristics that are different from most of the other data objects in the data sets.

(2) Values of an attributes that are unusual with respect to the typical values for that attribute.

1. **Missing Values-**

It is not unusual for a object to be missing one or more attribute value. In some cases the Information was not collected missing value should be taken into account during the data analysis.

1. **Duplicate Data –**

A data set may include data objects that are duplicates or one another.

* **Issues Related to Application : -**

Data quality issues can also be considered from an application view point as expressed by the statement “data is high quality if it is suitable for its intended use”. There are many issues that are specified to particular application and fields data quality have proven quit useful particularly in business and Industry. We consider only a few of the general issue.

1. **Timeliness :**

Some data starts to age as soon as it has been collected. If the data provides the purchasing behavior of customers or web browsing patterns, represte reality for only a limited time. If the data is out of data. Then so are the models and patterns that are based on it.

1. **Relevance :**

The available data must contain the information necessary for the Application considers the task making sure that the objects in a data sets are relevant is also challenging.

1. **Knowledge about the Data :**

Data sets are accompanied by documentation that describes different aspects of the data.

* **Data Preprocessing : -**

Data preprocessing is a broad area and consist of a number of different strategies and techniques that are interrelated in complex ways.

1. Aggregation
2. Sampling
3. Dimensionality reduction
4. Feature subset selection
5. Feature creation
6. Discretization and binarization
7. Variable transformation
8. **Aggregation :-**

Sometimes “Less is more” and this is the case with aggregation the combining of two or more objects into a single objects a qualitative such as atom can either be omitted or summarized as the set of all the atoms that were sold at that location.

Aggregation is the process of elevating Attribute such as the type of atoms or reducing the number of value for a particular attributer.

1. **Sampling :-**

Sampling is a commonly used approach for selecting a subset of the data objects to be analyzed in statistics it has long been used for both the preliminary investigation of the data and the final data analysis sampling can also be very useful in data mining.

**a) Sampling without Replacement -**

As each item is selected it is removed from the set of all objects that together constitute the populations.

**b) Sampling with Replacement –**

Objects are not removed from the population as they are selected for the sample.

1. **Dimensionality Reduction :-**

Data set can have a large number of features. The term dimensionality reduction is often reverse for those tequilas that reduce the dimensionality of the data set by creating new attributes that are a combination of the old attributes. Their a variety of Benefits to dimensionality reduction a key benefits is that many data mining algorithms work better if the dimensionality the number of attributes in the data is lower.

1. **Feature subset selection :-**

Another way to reduce the dimensionality is to use only a subset of the features. Redundant features duplicate much or all the Information’s contained in one or more other attributes.

For Ex :- The purchase price of a product and the amount of sales tax paid contain much of the same information.

**a) Embedded approaches-**

Feature selection occurs naturally as a part of the data mining algorithm during the operation of data mining algorithm the algorithm itself decide which Attributes is to used and which to ignore.

**b) Filter approaches –**

Feature are selected before the data mining Algorithm is run using some Approach that is independent of the data mining task.

**c) Wrapper approaches-**

This methods use the target data mining Algorithms is helps machine learning Algorithms to understand data and determine patterns that can improve the patterns of machine learning Algorithms.

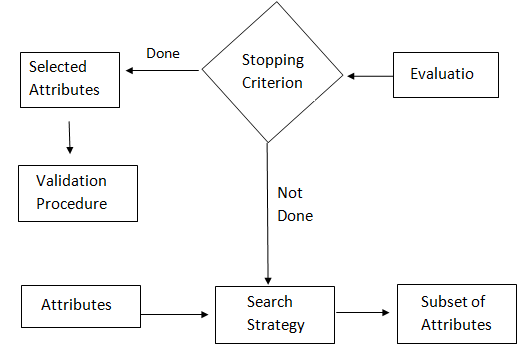


Fig :- Flowchart of a feature subset selection process.

1. **Feature Creation :-**

It is frequently possible to create from the original Attributes a new set of attributes that captures the important information in a data set much more effectively three related methodologies for creating new attributes.

* + - 1. Feature extraction
      2. Mapping the data to a new space
      3. Feature construction.

1. **Feature Extraction :-**

The creation of new set of features from the original raw data is known as feature extraction.

1. **Mapping the data to a new space :-**

A totally different view of the data can reveal important and interesting feactures.

Ex :- Time series data with often contain periodic pattern.

1. **Feature Construction :-**

The features in the original data sets have the necessary information one or more new features constructed out of the original features can be more useful than the original features.

1. **Discretization and Binarization :-**

Data Discretization is a method of converting Attributes value of continues data into a finial set of intervals with minimum data loss.

* **Binarization-**

Data Binarization is used to transform the continous and discrete attribute into Binary Attributes.

* **Variable Transformation :-**

A variable transformation refers to a transformation that is applied to all the values of a variable. In other words for each objects the transformation is applied to the value of the variable for that object two important types of variable transformation.

1. Simple function transformation
2. Normalization.