* **The Iris Data Set :-**

The techniques used in Data Exploration such as visualization can be used to understand and Interpret data mining result Data exploration can Aid in selection the processing and Data Analysis it can even address some of the question typically answer by Data Mining. The Iris data set i.e., available for the university of California at Irvine (UCI) Machine learning Repository. It consists of information on 160 Iris flowers 50 each from one of three Iris species.

1. Setosa
2. Versi colour
3. Vergincia

Each flower is characterized by 5 Attribute

1. Sepal length in centimeters
2. Sepal width in centimeters
3. Petal length in centimeters
4. Petal width in centimeters
5. Class (Setosa, Versiclour, virgincia)

* **Summary Statistics :-**

1. Summary statistics are number that summarized the property of Data.
2. Summary statistics certainty such as the main and standard deviation that computers various characterized sticks of a potentially large set of values with a single number of a small set of numbers.
3. Summarized property included frequency location and spread.

Ex :- Location that – mean

Spread – Standard deviation

Most summary statistics can be calculated in a single pass through the data

* **Frequencies and the mode :-**

1. Frequencies of a attribute value is the percentage of time the value occurs in the data sets.

Ex :- Given the attributer gender and repented the population the gender female occurs about 50% of the time.

1. The mode of an attribute is the most frequent attribute value.
2. The frequency and mode are typically used with categorical data given a categorical attributes ‘X’ which one take a value.

{V1------------- Vi ------------- VK} and a set of m object the frequency of value (Vi) is define as

1. The mode of the categorical attributes is the value that has the highest frequency.

* **Percentiles :-**

For continuous data percentiles is more useful given and ordinal or continuous attribute be and the number of P between O & 100 the pth percentile is the value of Xp such that P present of observed value of X are less than Xp for instance the 50th percentile is the value X 50% such that 50% of all values of X are less than X 50%.

* **Measures of location : Mean and Median : -**

1. The mean is the most common measure of the location of a set of points.
2. The mean is very sensitive to outlier
3. The median or a trimmed mean is also commonly used.

x/r + 1

m is set of object

X is an of attributes

Ex : -(a) 1, 4, 3, 9, 2

1, 2, 3, 4, 5, 9

Median

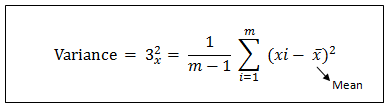
* **Measures of Spread : Range and variance :-**

The simplest measure of spread is the rage which given on attributes ‘x’ with a set of m value (x1 \_ \_ \_ \_ \_ xm), is defined as

Range (x) = max (x) – min (x)

= x (m) – x (1)

1. Range is difference between the max & min
2. The variation an standard deviation is the most common measure of the spread of a set of points.



1. The standard deviation which is the square root of the variance is return as

* **Multivariate Summary Statistics :-**

Measure of location of data that consist of several attributes (Multivariate data) can be obtained by computing the mean or medio separately for each attribute thus given a data set the mean of the data objects , is given by

Where is the mean of the ith attributes

Covariance Matrix = S

ith and jth attributes of the data If and are the ith and jth attribute.

Where and are the value of ith and jth attribute

* **Visualization:-**

Data visualization is a Display of Information in a graphic and tabular format successful visualization requires that the data (information) be converted into a visual format so that the characteristics of the data and the relationship among data atoms or attributes can be analyzed or reported. The goal of visualization is the Interpretation of the visualized information by a person and the formation of a mental model of the information. In ever day life visual techniques such as graphs and tables are often the preferred approach the weather, the economy the results of political elections.

1. Visualization is the conversation of data into a visual or a tabular format so that characteristics of the Data and the relationships among Data atoms or attributes can be analyzes or reported.
2. Visualization of Data is the one of the most powerful appealing techniques for data exploration.
3. Humans have a will develop ability to analyze large amount of Information that is presented visually.
4. Can detect general patterns & trends.
5. Can detect outliers and unusual patterns.
6. **Representation :-**
7. Representation is the mapping of Information to a visual format

i.e., mapping of objects, Attributes and the Relationship in a set of Information to visual objects Attributes and Relationships.

1. The data objects, Their Attributes are the Relationship among Data objects are Translated into Graphical elements such as points, lines shapes, and colours.

For Example :- 1) Objects are Repented as prints

2) There attributes values can be represented as the position

of the point the characteristics of the points colour, size

and shape.

3) If position is used that the relationship of points i.e., whether they form groups or a point is on outlier is easily per give

1. **Arrangement :-**
2. Is the placement of visual element within a display.
3. Can make a large difference in how easy to understand the data.

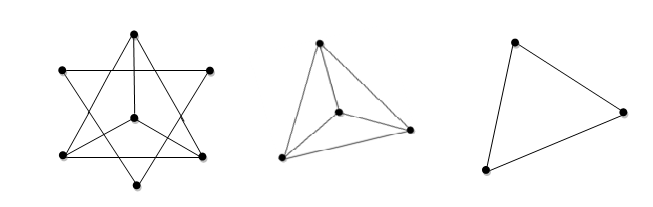
**Example :-**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **.** | **1** | **2** | **3** | **4** | **5** | **6** |
| **1** | **0** | **1** | **0** | **1** | **1** | **0** |
| **2** | **1** | **0** | **1** | **0** | **0** | **1** |
| **3** | **0** | **1** | **0** | **1** | **1** | **0** |
| **4** | **1** | **0** | **1** | **0** | **0** | **1** |
| **5** | **0** | **1** | **0** | **1** | **1** | **0** |
| **6** | **1** | **0** | **1** | **0** | **0** | **1** |
| **7** | **0** | **1** | **0** | **1** | **1** | **0** |
| **8** | **1** | **0** | **1** | **0** | **0** | **1** |
| **9** | **0** | **1** | **0** | **1** | **1** | **0** |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **.** | **6** | **1** | **3** | **2** | **5** | **4** |
| **4** | **1** | **1** | **1** | **0** | **0** | **0** |
| **2** | **1** | **1** | **1** | **0** | **0** | **0** |
| **6** | **1** | **1** | **1** | **0** | **0** | **0** |
| **8** | **1** | **1** | **1** | **0** | **0** | **0** |
| **5** | **0** | **0** | **0** | **1** | **1** | **1** |
| **3** | **0** | **0** | **0** | **1** | **1** | **1** |
| **9** | **0** | **0** | **0** | **1** | **1** | **1** |
| **1** | **0** | **0** | **0** | **1** | **1** | **1** |
| **7** | **0** | **0** | **0** | **1** | **1** | **1** |

(b)

(a)

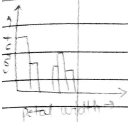
****

(b) Uncoupled view of connected components of the graph.

1. Original view of a graph
2. **Selection :-**
3. Is the elimination or the de-emphasis of certain objects and attributes.
4. Selection may involve the choosing of subset of attribute dimansitionality radiation is often to use the number of dimension to two or three alternatively pair of attribute can be consider.
5. Selection may involve choosing a subset of objects a Region of the screen can only show so many points.

* **Visualization Techniques : -**

1. Visualization techniques are aften specialized to the type of Data being analyzed.
2. Visualization Method can also classified according to the type of attribute involve.



1. **Histogram :-**
2. Usually show the distribution of value of a single variable.
3. Divided the values into bins and showing a bar plot of the number of objects in each bin.
4. The height of the each bar indicates the number of objects.
5. Shape of the histogram depends of upon the number of bin.

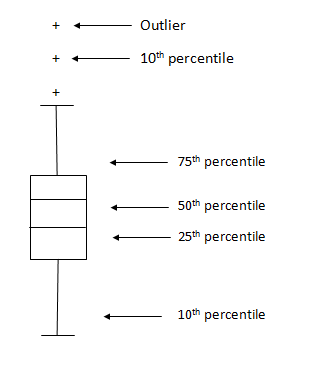
* **Two – Dimensional Histograms :-**

1. Show the joint distribution of the values of two attributes.

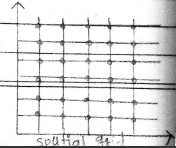
Example : Petal width and petal length.

1. **Box plots :-**

Box plots are another method for showing the distribution of the value of the single numerical attributes the lower and upper end of the Box indicate 25th and 75th percentile while the line inside the Box indicate the value of 50th percentile. The top and bottom lines 19th percentile outliers are shown by “t” marks.



1. **Scatter plots : -**
2. Attributes values determine the position.
3. Two dimensional scatter plots most common but can have 3 – D Scatter plots.
4. Often additional attributes can be displayed by using the shape, size and colour of the market that represent the objects.
5. It is useful to have arrays of scatter plots can compactly summarize the relationships of several pairs of attributes.
6. Each data objects is plotter as a point in the plane using the values of two attributes as X & Y co-ordinates It is assumed that the attributes are either integer or real value.
7. **Contour Plots:-**
8. Useful when a continuous attribute is measured on a spatical grid.
9. They partition the plane into regions of similar values.



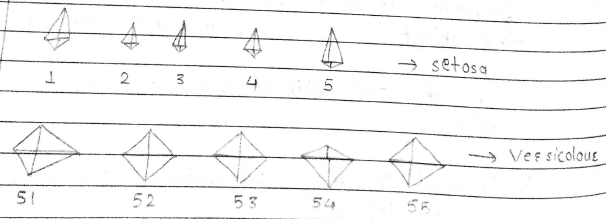
1. The contour line that form the boundaries of these regions connect point with equal values.
2. The most common example is contour maps of elevation can also display temperature, rainfall, air pressure an example for C – surface temperature is provided on the next line.
3. **Matrix plots** **:-**
4. Can plot the data matrix.
5. These can be useful when objects are sorted according to class.
6. Typically the attributes are normalized to prevent one attribute from dominating the plot.
7. Plots of similarity or distance matrices can also be useful for visualizing the relationships between objects.
8. A data matrix is a rectangle array of values.
9. A data matrix can be visualized as an image by associating each entry of the data matrix with a pixel in the image.
10. The brightness or the colour of the pixel is determined by the value of the corresponding entry of the matrix.

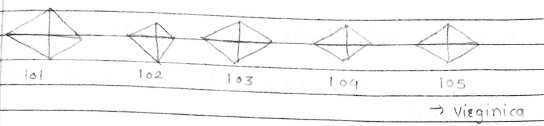
* **Parallel co – ordinates** **:-**

1. Parallel co – ordinates have one co – ordinate arise for each attribute but the different axis are parallel to one another instead of perpendicular.
2. An object is represented as a line instead of as a point.
3. The value of each attribute of on object is mapped to a point on the co – ordinate axis associated with that attribute and these points are thin connected to form the line that represent the object.
4. Use the plot attribute values of high dimensional data.
5. Instead of using perpendicular axis use a set of parallel axis.
6. The attribute values of each object are plotted as a point on each corresponding co – ordinate axis and the point are connected by a line.
7. Thus, each object is represented as a line.
8. Often the lines representing a distinct class of objects group, together at least for some attributes.
9. Ordering of attributes is important in seeing such grouping.

* **Star Coordinates :-**

1. It is similar to parallel co – ordinates but axes radiate flam a central point.
2. The line connecting the values of an object is a polygon.



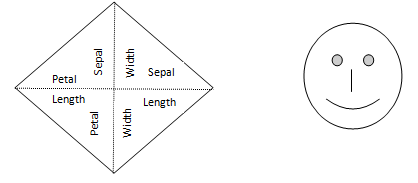


* **Chernoff faces:-**

Approach created by Heeman cheroff. This approach associates each attribute with a character.

The values of each attribute determine the appearance of the corresponding facial characteristics.

Each object becomes a separate phase Rekase on Human’s ability to distinguish faces.



b) Chernoff face of Iris 150

a) Star graph of Iris 150

Fig :- A plot of 15 Iris flowers Using chernoff faces.

* **OLAP [online Analytical processing] :-**

1. OLAP was proposed by E.F code, the fathers of relational database.
2. Relational database put data into tables while OLAP uses a multidimensional array representation such representation of the data previously existed in statistics and other fields.
3. There are a Number of data analysis and data exploration operation that are easier with such a data representation.
4. OLAP systems also have a strong focus on the interactive analysis of data and typically provide extensive capabilities for visualizing the data and generating summary statistics.

* **Creating Multidimensional Array :-**

Two key steps in converting to bluer data into a multidimensional array.

1. Identify which attribute are to be the dimension and which attribute is to be the target attribute whose values appear as entries in the multidimensional array.

* The attribute used as dimensions must have discrete values.
* The target value is typically a count or continuous value.

Ex., The cost of an item.

* Can have no target variable at all except the count of object that have the same set of attribute values.

1. Find the value of each entry in the multidimensional array by summing the values or count of all objects that have the attribute values corresponding to that entry.

Most data sets can be represented as a table where each row is an object and each column is an attribute. It is possible to view the data as multidimensional array.

Example1: (1) we show how the attributes, petal length, petal width and species type can be converted to a multidimensional array.

We discriticize the petal width and length to have categorical values low, medium, high.

The Iris data set that have particular combination of petal width, petal length and species style.

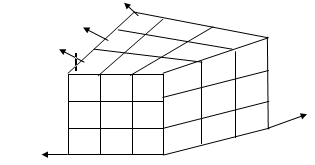
|  |  |  |  |
| --- | --- | --- | --- |
| **Petal length** | **Petal width** | **Species type** | **Count** |
| Low | Low | Setosa | 46 |
| Low | Medium | Setosa | 2 |
| Medium | Low | Setosa | 2 |
| Medium | Medium | Versicolor | 43 |
| Medium | High | Versicolor | 3 |
| Medium | High | Versicolor | 3 |
| High | Medium | Versicolor | 2 |
| High | Medium | Verginica | 3 |
| High | High | Versicolor | 2 |
| High | High | Verginica | 44 |

Note :- Tuple is an ordered sequence of elements of different data type.

Example- (2) Iris data (Continue)

1. Each unique typle of petal width, petal length and species style identifies one element of the array.
2. These element is assigned the corresponding count values.
3. All non – specified typles are zero.

Verginica



46

2

2

0

0

0

0

0

0

Petal width

Versicolou

Setosa

|  |
| --- |
| Low  Medium  High |

High

Medium

Low

Fig :- A multidimensional data representation of the Iris data set.

1. **Data Cube example :-**
2. Consider a data set that record the sales of products at a no of company stores at various dates.
3. These data can be represented as a three dimensional array.
4. There are three dimensional aggregates, three 1-D aggregates and one zero – Dimensional aggregate.

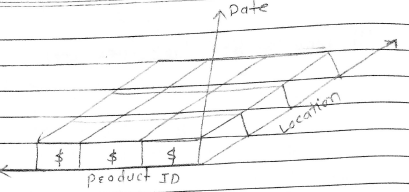


Fig :- Multidimensional data representation for scales data.

A key motivation for taking a multidimensional view point of data is the importance of aggregating data in various ways.

* **OLAP Operations : -**

1) Slicing 3) Roll up

2) Dicing 4) drill down

1. **Slicing –**

Slicing is selecting a group of sales from the entire multidimensional array by specifying a specific value for one or more dimensions.

1. **Dicing-**

Dicing involves selecting a subset of sales by specifying a range of attribute values.

This is equivalent to defining a sub array from the complete array.

Drill down 🡪 split or divide/break in no. of months/days.

Roll up 🡪 Rise of data

* Both operation can also be accompanied by aggregations over some dimensions.

1. **Roll up –** Attribute values often have a hierarchical structure.
2. Each date is associated with a year, month and week .
3. A location is associated with a continent. Country and state.
4. Product can be divided into various categories such as clothing, electronics and furnitures.

Note:- that, these categories often nest and farm a tree.

* A year contains months which contains day.
* A country contains a state which contains a city.
* **Classification :-**

Classification which is a task of objects to one of the several predefined categories is a prevasive problem that encompasses many diverse applications.

1. Give a collection of records each record contains a set of attributes one of the attribute is the class
2. Find a model for class attribute as a function of values of other attributes.
3. Goal – Record should be assigned a class as accurately as possible.

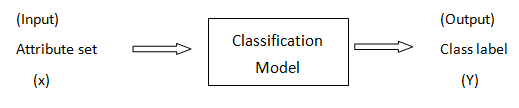


Fig- classification as the tash of mapping an input attribute set x into its class liberty.

* **Preliminaries :-**

The Input data for a classification task is a collection of records.

Each record is also known as an instance or example is categorized by a typle (x,y) where x is attribute set and y is special attribute.

* **Definition (Classification) :-**

Classification is a task of learning a target function of that maps each attribute set x to one of the predefined class label y.

* **General approach to solving a classification problem:-**

1. A Classification technique is a systematic approach to building classification models from an input data set.

Example:- Include decision tree classifier, rule based classifier, Neural networks, support vector. Machines and Bayes classifier.

1. Each technique employer a learning algorithm to identify a model that best feeds the relationship between the attribute set and class label of the input data.

**Training set**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tid** | **Attrib 1** | | **Attrib 2** | **Attrib 3** | **class** |
| **1** | Yes | Large | | 125 k | No |
| **2** | No | Medium | | 100 k | No |
| **3** | No | Small | | 70 k | No  Learning Algorithms |
| **4** | Yes | Medium | | 120 k | No |
| **5** | No | Large | | 95 k | Yes |
| **6** | No | Medium | | 60 k | No |
| **7** | Yes | Large | | 220 k | No  Introduction |
| **8** | No | Small | | 85 k | Yes  Learn  Model |
| **9** | No | Medium | | 75 k | No |
| **10** | No | Small | | 90 k | Yes |

Model

**Test Set**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Tid** | **Attrib 1** | | **Attrib 2** | **Attrib 3** | **class**  Deduction |
| **11** | No | Small | | 55 k | ?  Apply  Model |
| **12** | Yes | Medium | | 80 k | ? |
| **13** | Yes | Large | | 110 k | ? |
| **14** | No | Small | | 95 k | ? |
| **15** | No | Large | | 67 k | ? |

Fig :- General approach for building a classification model.

* **Examples of classification task :-**

1. Predicating tamer cells as malignant.
2. Classifying credit card transactions as legitimate and fraudulent.
3. Classifying secondary structure of protein as & B or random coil
4. Categorizing news stories as finance, weather, entertainment & sports.

* **Classification Techniques :-**

1. Decision tree based methods.
2. Rules based method.
3. Memory based reasoning.
4. Neural Networks.
5. Bayes Network.
6. Support vector Machines.

* **Decision tree Induction :-**
  + 1. How a decision tree works?

🡪 We assign them two categories mammals and non-mammals.

1. A Root nodes that has no incoming edges and zero or more outgoing edges.
2. Internal nodes each of which has exactly one incoming edge and two or more outgoing edges.
3. Leaf or terminal nodes each of has exactly one incoming edge no outgoing edges.

Root node

Internal node

Body temp

erature

Warm

Gold

None - mammals

Gives Birth

Yes

No

Non Mammals

Mammals

Leaf nodes

Fig:- A decision tree for the mammal classification problem.