Ankit Roy 1906534

Q2

Data preprocessing is crucial in any data mining process as they directly impact success rate of the project.

This reduces complexity of the data under analysis as data in real word is unclean

- method involved in data processing

(1) Data cleaning

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

(a) missing data

The situation arises in our

of missing data. It can be handled in various ways:-

Some of them can:

(i) Ignore the tuppies

(ii) fill the missing values.

(b)

Noisy data

Noisy data is an meaningless data that cant be interpreted by machnies. It can be generated due to faulty data collection, data entry error etc.

Q. It can be handled in following ways:

(i) Binning method:

The method works on sorted data in order to smooth it.

The data is devided into segments of equal size and then various methods are performed.

(ii) Regression:

Data can be "made smooth by fitting it to a regression function. The regression used may be linear, or multiple.

(iii) Clustering

"It groups the similar data in the clusters.

"It outliers may be undetected or it will fall outside the clusters.

(2.) Data transformation

"It the step is taken in order to transform the data in appropriate forms suitable for mining process.

It involves the following ways:

(i) Normalization

It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)

(ii) Attribute ~~selb~~ selection:

In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

(iii) Discretization

This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

(iv) concept Hierarchy generation

Here attributes are converted from lower level to higher level in hierarchy. for ex: The attribute "city" can be converted to "country".

(3) Data Reduction

Since Data mining is a technique that is used to ~~mang~~ handle huge amount of data.

The various steps to data reduction are :-

(i) Data cube Aggregation :-

Aggregation operation is applied to data for the construction of the data cube.

(ii) Attribute subset selection

The highly relevant attributes should be used, rest all can be discharged.

(iii) Numerosity Reduction

This enable to store ~~the~~ the model of data instead of whole data, for ex:- Regression models.

(iv) Dimensionality Reduction

Reduction

This reduce the size of data by encoding mechanism. It can be lossy or lossless.