SNS COLLEGE OF TECHNOLOGY

Coimbatore-35.

An Autonomous Institution

COURSE NAME : 19CST203 - DATA ANALYTICS

II YEAR /IV SEMESTER

1.DATA QUALITY

The quality of the models, charts and studies in data analytics depends on the quality of the data being used. The nature of the application domain, human error, the integration of different data sets (say, from different devices), and the methodology used to collect data can generate data sets that are noisy, inconsistent, or contain duplicate records.

Today, even though there is a large number of robust descriptive and predictive algorithms available to deal with noisy, incomplete, inconsistent or redundant data, an increasing number of real applications have their findings harmed by poor-quality data. In data sets collected directly from storage systems (actual data), it is estimated that noise can represent 5% or more of the total data set [13]. When these data are used by algorithms that learn from data – ML algorithms – the analysis problem can look more complex than it really is if there is no data pre-processing. This increases the time required for the induction of assumptions or models and resulting in models that do not capture the true patterns present in the data set.

The elimination or even just the reduction of these problems can lead to an improvement in the quality of knowledge extracted by data analysis processes. Data quality is important and can be affected by internal and external factors.

• Internal factors can be linked to the measurement process and the collection of information through the attributes chosen. 72 4 Data Quality and Preprocessing

• External factors are related to faults in the data collection process, and can involve the absence of values for some attributes and the voluntary or involuntary addition of errors to others.

The main problems affecting data quality are now briefly described. They are associated with missing values, and with inconsistency, redundancy, noise and outliers in a data set.

2.MISSING VALUE

In real-life applications, it is common that some of predictive attribute values for some of the records may be missing in the data set. There are several causes of missing values, among them

• attributes values only recorded some time after the start of data collection, so that early records do not have a value

• the value of an attribute being unknown at time of collection

• distraction, misunderstanding or refusal at time of collection

• attribute not required for particular objects

• non-existence of a value

• fault in the data collection device

• cost or difficulty of assigning a class label to an object in classification problems.

Since many data analysis techniques were not designed to deal with a data set with missing values, the data set must be pre-processed. Several alternatives approaches have been proposed in the literature, including:

• Ignore missing values:– Use for each object only the attributes with values, without paying attention to missing values. This does not require any change in the modeling algorithm used, but the distance function should ignore the values of attributes with at least one missing value; – Modify a learning algorithm to allow it to accept and work with missing values.

• Remove objects: Use only those objects with values for all attributes.

• Make estimates: Fill the missing values with estimates based on values for this attribute in the other objects

3.REDUNDANT DATA

While missing values are a lack of data, redundant data is the excess of it. Redundant objects are those that do not bring any new information to a data set.Thus, they are irrelevant data.They are objects very similar to other objects.

Redundancy occurs mainly in the whole set of attributes. Redundant data could be due to small mistakes or noise in the data collection, such as the same addresses for people whose names differ by just a single letter.

4.INCONSISTANT DATA

A data set can also have inconsistent values.The presence of inconsistent values in a data set usually reduces the quality of the model induced by ML algorithms. Inconsistent values can be found in the predictive and/or target attributes.

An example of an inconsistent value in a predictive attribute is a zip code that does not match the city name. This inconsistency can be due to a mistake or a fraud.

In predictive problems, inconsistent values in the target attribute can lead to ambiguity, since it allows two objects with the same predictive attribute values to share different target values. Inconsistencies in target attributes can be due to labeling errors.

Some inconsistencies are easily detected. For example, some attribute values might have a known relationship to others, say that the value of attribute A is larger than the value of attribute B. Other attributes might only be allowed to have a positive value. Inconsistencies in these cases are easily identified.