Geraldo Braho

Synopsis No. 3

Due Date
02/18/2019

Data Preprocessing

- Data Quality -data quality, including accuracy, completeness, consistency, timeliness, believability, and interpretability.

Data cleaning routines work to “clean” the data by filling in missing values, smooth- ing noisy data, identifying or removing outliers, and resolving inconsistencies. If users believe the data are dirty, they are unlikely to trust the results of any data mining that has been applied.

- Noisy Data

Noise is a random error or variance in a measured variable. Binning: Binning methods smooth a sorted data value by consulting its “neighbor- hood,” that is, the values around it.

Regression: Data smoothing can also be done by regression, a technique that con- forms data values to a function.

Outlier analysis: Outliers may be detected by clustering, for example, where similar values are organized into groups, or “clusters.”

Data mining often requires data integration—the merging of data from multiple data stores. Careful integration can help reduce and avoid redundancies and inconsistencies in the resulting data set. This can help improve the accuracy and speed of the subsequent data mining process.

Redundancy is another important issue in data integration. An attribute (such as annual revenue, for instance) may be redundant if it can be “derived” from another attribute or set of attributes.

Data integration also involves the detection and resolution of data value conflicts. For example, for the same real-world entity, attribute values from different sources may dif- fer.

Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume, yet closely maintains the integrity of the original data.

Dimensionality reduction is the process of reducing the number of random variables or attributes under consideration.

Numerosity reduction techniques replace the original data volume by alternative, smaller forms of data representation.