Data Cleaning

This process involves removing page headers and footers from files, expanding data with more than one row per record, stripping out nonnumeric characters from number fields, and a host of other procedures aimed at standardizing the data to make it suitable for use with the analysis tool selected as most appropriate for the data analysis exercise.

The data-cleaning stage is critical to the data analysis process. It is the only point at which alterations are intentionally made to the extracted data, and it is imperative that amendments made at this stage not affect the accuracy of the information.

Eliminating Duplicate Information

One of the problems forensic technologists encounter is that multiple copies of various materials may be recovered as part of the investigative process. Because of the expense involved in reviewing such duplicative materials, elimination of duplicates (duplication, or deducing) in the recovered data sets is often the first order of business after the data has been acquired and the documentation has been completed.

These issues require the examiner and the entire investigative team to map out their approach in advance and choose a process that is consistent with the particular project's needs.

We have several data mining applications classes used in fraud detection.

Classification

Classification builds up and utilizes a model to predict the categorical labels of unknown objects to distinguish between object of different classes. The categorical labels are predefined, discrete and unordered. Classification can be defined as the process of identifying a set of common features and models that describe and distinguish data concepts. Common classification techniques include neural networks, decision tree and support vector machine. Such classification task are used in the detection of credit card, healthcare and automobile insurance, and corporate fraud.

Clustering

Clustering is used to divide objects into conceptually meaningful group, with the objects in a group being similar to one another but very dissimilar to the objects in other group.