Data Mining

Data Collection Report

2/20/2018

Our current dataset is a dataset representing 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes over 50 features representing patient and hospital outcomes.

1. The dataset is obtained from UCI machine learning repository[1]

2. The dataset is about 18MB, containing about 100,000 hospital encounter records.

3. The original dataset is in .csv format, with about 100,000 entries. Each entry contains 55 attributes representing information about this encounter. Some attributes are numeric, and some are nominal. And there are some entries with missing attributes values. We are going to treat each entry as a feature vector.

4. Since some entry attributes are nominal, we need convert these attribute to numeric or binary. We might also have to do feature scaling and normalization. But basically, this data set is quite well organized and formatted, needing no hard work on converting or cleaning.

5. Our objective is to cluster the patients into several clusters. For each cluster, we are going to simulate similar data with the mean of that cluster and then generate some Gaussian noise to them. We will then use the simulated data to verify our model.

[1] http://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008