Data Mining

Data Collection Report

2/21/2018

Our current dataset represents 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. There are also 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 to convert these attributes to numeric or binary. However, this data set is quite well organized and formatted for the most part, needing little hard work to convert or clean it.

5. To simulate similar data, we could generate a random "patient", deduce what cluster we expect the patient should be in based on the data generated, then run our model with the new patient included to see if the resulting cluster matches what we expected. We could do this for several simulated "patients" to increase the confidence in our model.

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