Project Team 4

Members:
Chang Zhou
Harsha Jakkappanavar
Jenny Johns
Savan Patel

Study of dataset on Diabetes from 130 US hospitals over 10 years

October 12, 2017

Overview

The chosen dataset has been prepared to analyze factors related to readmission as well as other outcomes pertaining to patients with diabetes.

The dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes 100000 instances and over 50 features representing patient and hospital outcomes.

Information was extracted from the database for encounters that satisfied the following criteria:

- If it is an inpatient encounter (a hospital admission).
- If it is a diabetic encounter, that is, one during which any kind of diabetes was entered to the system as a diagnosis.
- If the length of stay was at least 1 day and at most 14 days.
- If the laboratory tests were performed during the encounter.
- If medications were administered during the encounter.

The data contains such attributes as patient number, race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test

result, diagnosis, number of medication, diabetic medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalization, etc.

More information about the attributes can be found under this link: https://www.hindawi.com/journals/bmri/2014/781670/tab1/

A paper written by Strack, DeShazo, Gennings, et al., based on this dataset focuses on how HbA1c levels impact hospital readmission rates. HbA1c is the measurement of the level of hemoglobin in the blood stream that has been attached to glucose. This is referred to as glycated hemoglobin ("What is HbA1c? - Definition, Units, Conversion, Testing & Control," 2017).

Questions we could come up with

Using the dataset we are aiming to find answers to the following questions:

1. **Clustering on types of diabetes:** Are there any clusters (subgroups of patients) in the dataset? E.g. type-1 vs type-2 diabetes.
2. **Prediction model on the dataset:** Can we diagnose the encounter as diabetic based on a model built using the dataset?
3. **Readmission rate:** What percent of the patients diagnosed with diabetes were readmitted within 30 days?
4. **Risk factors for readmission:** What are the potential attributes that can be identified as the factors for readmission.
5. **Reduce the cost of readmission:** By identifying the risk factors can we provide suggestions to reduce the cost of readmissions?

Algorithms we can apply

1. Clustering algorithms like centroid/prototype based(k-means), Hierarchical based
2. Expectation-maximization algorithm

3. FP Growth
4. Researching other algorithms to apply to the dataset
5. Comparison in terms of performance scaling etc.

Milestones

1. The primary milestone would be to preprocess the data by removing the outliers/missing values.
2. The attribute “diag_*” from the database defines a diagnosis coded as first three digits of ICD9. We can cluster on these values to determine different categories of diabetes.
3. Apply FP growth algorithm to identify risk factors for readmission.
4. Also our team feel the need to expose ourselves with more information on algorithms in order to better use them to apply here on this dataset.

Teamwork

1. Coding of algorithms and other features of the project will be split between all group members.
2. All members will work on the YouTube presentation.
3. Jenny will work on the paper.
4. Savan will work on graph generation.
5. Chang will work on analysis of the dataset.
6. Harsha will work on preprocessing of the dataset.
7. Note: This is a preliminary split of the work between the group members, depending on changes of ideas and questions applied to the dataset.

Citations

1. Beata Strack, Jonathan P. DeShazo, Chris Gennings, et al., “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records,” BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014. doi:10.1155/2014/781670
2. What is HbA1c? - Definition, Units, Conversion, Testing & Control. (2017). Retrieved from http://www.diabetes.co.uk/what-is-hba1c.html
3. The name of the dataset, Diabetes 130-US hospitals for years 1999-2008 Data Set http://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008#