Project: Visual Analytics

Overview

The goal of this project is to develop transformations and visualizations capable of developing and testing hypothesis from large amounts of unstructured and structured data. This process is known as "visual analytics." The user should be able to load a JSON, XML, or ASCII (tab or comma-delimited) file and should be able to view a visual representation of the data. The user should be able to apply analysis and transformations on the columns of data to create new columns and filter out some data before visualization. Other more advanced features to be implemented would involve discretizing data or finding relationships in the data that can be visualized in a graph form. The interface should be modular and present simple options for the user to explore various options. In addition to the visualized output data, the configuration settings of the dataset should be able to be exported (and imported).

Background

We do a lot with machine learning. However, visualizing the data used for training machine learning models a priori (often not done) would allow one to gather valuable intuition about the data set. For example, this could help with determining the right data to use (i.e., feature engineering), determining the right machine learning algorithms to use, validating the results of machine learning, and even determining the right problems to solve with machine learning.

Visual analytics is the process of converting data into intuition to *better* extract value from the data. The project will involve working with discrete numbers, continuous numbers, text, and graph-based representations. Any visual analytics framework will have to work with a variety of different input formats. There are quite a few challenges, specifically in data transformation algorithms and displaying large amounts of data (one could use staging of the data).

Objectives

1. Develop an application that can perform data visualization
2. Develop transforms that can be applied to the data before visualizing the data (e.g., mean, standard deviation, min, max, median)
3. Develop a transformation to construct graph-based data representations depicting relationships in the data

Data

We will provide several sample cybersecurity datasets and will outline several visualizations of the data one could perform. We will provide you with support with understanding any transformations. We will provide a larger dataset of binary applications with features extracted for a final performance and visual analytics evaluation.

Considerations

We do not expect it to be perfect, but applying several transformations on the data should be possible. Speed and performance of loading and visualizing the larger datasets is of utmost performance. The user interface should be modular to allow various different visualizations to be performed.

Technology

- Data transformations and filtering should be performed using Python.
- Visualization could be done using one of two different open source visualization platforms, i.e., D3 or TensorFlow Projector.

Related Links

1. Tensorflow projector
   http://projector.tensorflow.org/
   https://www.tensorflow.org/how_tos/embedding_viz
   https://arxiv.org/pdf/1611.05469v1.pdf
   Standalone version to start from.
   https://github.com/tensorflow/embedding-projector-standalone
   Note: I suggest you run the javascript through a pretty printer.