Visual Explorer of Multivariate Data

Tisagh Chase

St. George's University

Introduction

Data visualisation is an integral part of data analysis; visual inspection can help to reveal patterns that would be difficult to reveal computationally. The problem becomes expounded when one deals with large sets of multivariate data. Here, a situation is considered when one has revealed some patterns in one projection using one set of selected variables (left panel) and is interested to find corresponding points on another projection (right panel). One should bear in mind that several data points could be projected to the same point on the left panel, as well as one point on the left panel can correspond to several points on the right panel. The problem of revealing structures and features automatically appeared to be rather complicated algorithmically, so it was decided to make a system for interactive analysis.

Figure 1

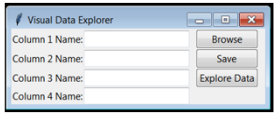


Figure 2

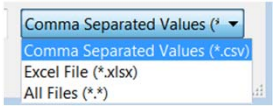


Figure 3

Figure 4

	Entropy	Length	X	Y
0	0.	100	1	1
1	0.6931471805599453	2	2	1
2	0.6931471805599453	2	3	1
3	0.6931471805599453	2	4	1
4	0.6931471805599453	2	5	1

System Design

Typically when the term "big data" is used any subsequent visualisation of the data tends to ignore individual data points and focus on groups and/or diffusion of points. However, the requirements of this system were such that large quantities of data needed to be visually represented while maintaining individual data-point interactivity.

The system was initially developed using Microsoft Visual Studio Community 2017 IDE, Python 3.6 and some of its supporting libraries designed for data analysis.

Tkinter was used to develop the GUI (*Figure 1*), allowing the user to retrieve data files in either CSV or Excel formats (*Figure 2*). The user could then label the respective axes as required before exploring the data (*Figure 3*).

Pandas was used to reconstruct the data on the backend into a data frame for easy translation to plot (*Figure 4*).

Numpy was imported to run any statistical operations required by the user, none were utilised in the test case.

Holoviews (with a Bokeh backend) was used to create a toolbar for various levels of interactivity (*Figure 5*). The main tool functionalities include: zoom, pan, select, save (as .png), reset, and help.

On the backend, data was extracted as required from the Pandas data frame, Holoviews utilised the power of Bokeh to construct plots (*Figure 6*) and assign respective tools for the users use on the front end. Bokeh was also used to translate the code from Python Script to HTML and JavaScript for use on the front end through any browser with JS enabled.

Some JavaScript was used directly in the code for additional functionality and user interactivity such as hover information (*Figure 7*) and the ability to generate a (.txt) file containing a list of data points selected as well as displaying the said list just below the plot with the number of points selected (*Figure 8*).

Interactive data visualisation was achieved by using web browsers through JSON blobs (*Figure 9-12*).

Observation

The system was successfully capable of providing an interactive set of plots allowing the user to select and map chosen columns to respective axes on a projection of choice. It was noted that a high definition display was essential for interactivity above 25,000 datasets. Test data contained 1.7 million datasets.

This outlined the limitations of the chosen libraries and plans for the next version will utilise yet another library specifically designed for massive datasets with less interactivity.

Figure 5

Figure 6

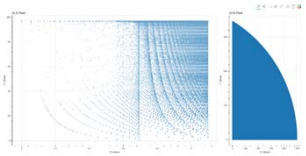


Figure 7

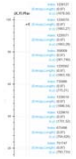


Figure 8

Selected: 226
Indices:
1295,33,384,749,2061,2461,1471,234,262,3

Figure 9

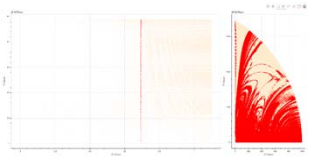


Figure 10

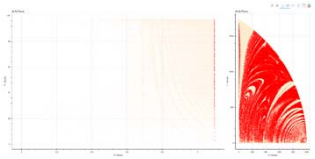


Figure 11

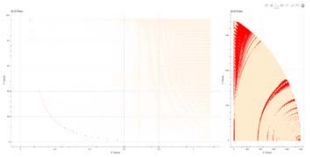
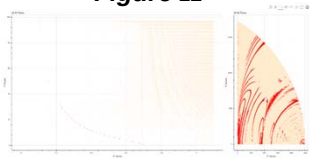
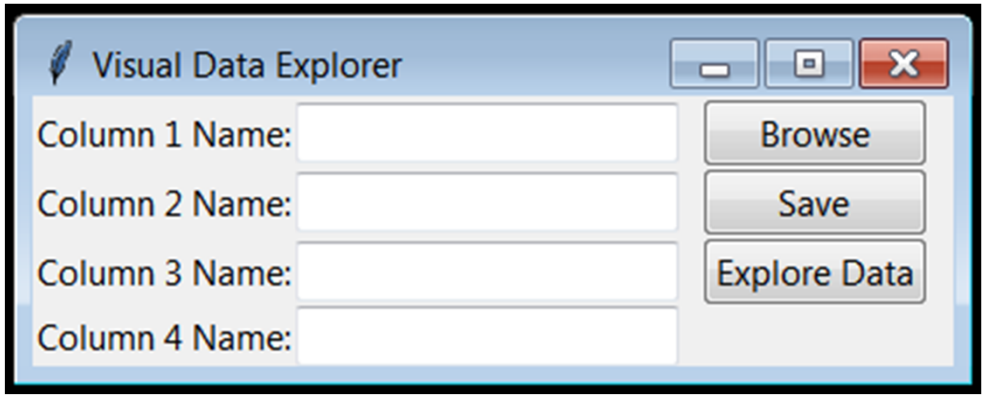


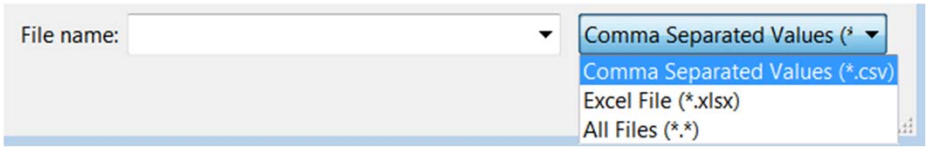
Figure 12

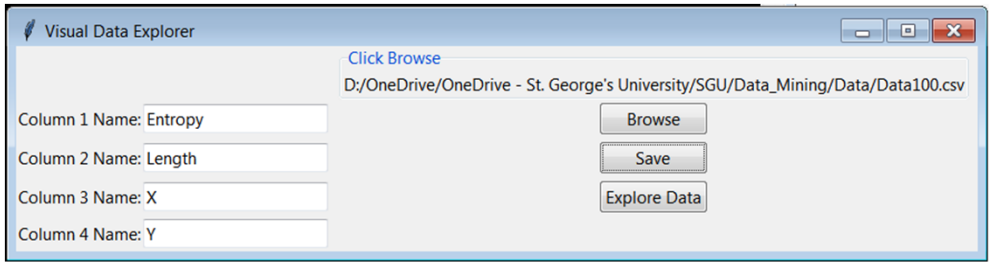


References

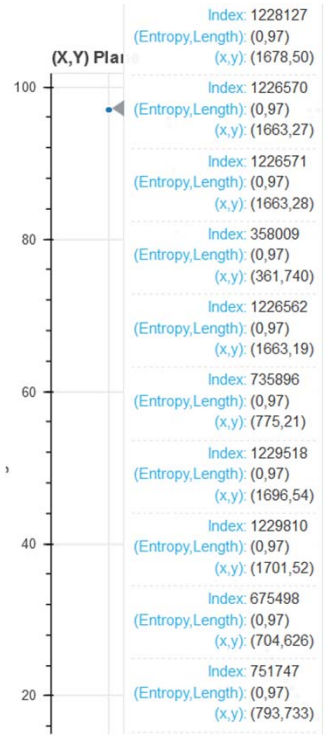
1. Chase, T., Mylläri, A., & Mylläri, T. (2019). Visual Explorer of Multivariate Data. Procedia Computer Science, 150, 416-424.
2. Kirk, A., Timms, S., Rininsland, Æ., Teller, S. (2016). Data visualization: Representing information on modern web. Birmingham: Packt Publishing Ltd.
3. Ward, MO., Grinstein, G., Keim, D. (2015). Interactive data visualization: Foundations, techniques and applications. 2nd ed. Boca Raton: Taylor & Francis Group, LLC.

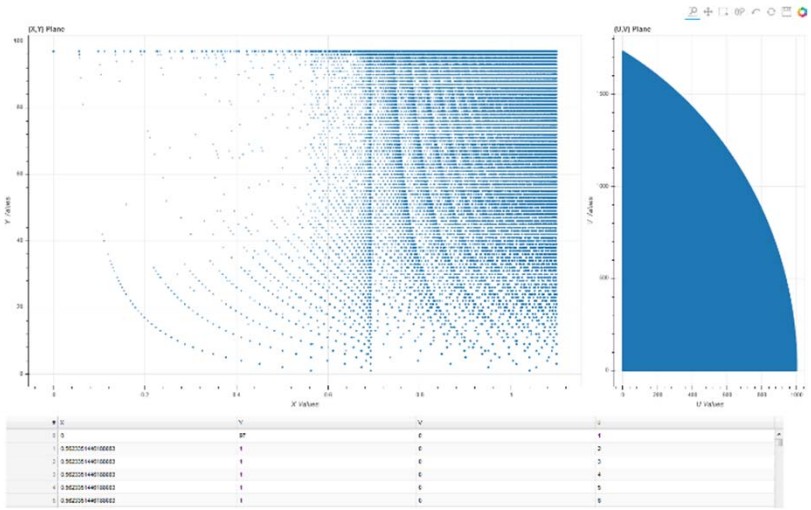






	Entropy	Length	X	Y
0	0.	100	1	1
1	0.6931471805599453	2	2	1
2	0.6931471805599453	2	3	1
3	0.6931471805599453	2	4	1
4	0.6931471805599453	2	5	1





Selected: 226
Indices:
1295,33,384,749,2061,2461,1471,234,262,3