Dr. Eick

COSC 4335*“Data Mining”* Assignment4 Spring 2016

*Design and Implementation of an Outlier Detection Technique for Spatial Data*

Final Draft

Individual Project

Due date: Friday, May 6, 11**a** (students who submit by May 3, 11p get a 5% bonus)

Last updated: April 26, 2016 at 11a



The goal of the project is to design and implement a bivariate, spatial outlier detection technique of your own preference and to apply it to a 2-dimensional dataset called *Complex9\_gn8*[[1]](#footnote-1)(<http://www2.cs.uh.edu/~ceick/UDM/complex9_gn8.txt>) which is the variation of the Complex9 dataset, we used before. Your outlier detection technique should take the dataset and create a copy of the dataset that contains an additional column/attribute called *ols (“outlier score”)* which contains numbers that indicates how much your outlier detection method believes that the particular object is an outlier—the smaller the value of the ols attribute the less likely the object is believed to be an outlier. For the project, you can use any R-library or other software to accomplish the project tasks; just acknowledge what external software you used in your project report.

The complex9\_gn8 dataset has attributes x, y, class; for example, after applying your outlier detection technique to the last 5 examples of the dataset, the result created by your outlier detection technique could look as follows:

728.899,535.627,8,0.24

504.528,-46.2297,8,0.41

373.256,409.026,8,0.12

850.838,242.711,8,0.33

641.676,347.544,8,0.11

This depicted result indicates that the second example is the most likely outlier, the fourth example is the second most likely outlier,…, and that the last example is the least likely outlier.



Complex9\_gn8 Dataset.

**Assignment4 Tasks:**

Task0: Visualize the Complex9\_gn8 dataset; visualize the third attribute using different colors, similar to supervised scatterplots we used in Assignments 1.

Task1: Develop a 2D spatial outlier detection technique of your own preference that identifies abnormal data in datasets which contain pairs of numbers.

Task2: Implement the chosen outlier detection technique for the complex9\_gn8 dataset.

As explained earlier your implementation of your outlier detection technique should add a column/attribute *ols* to the dataset and fill this column with numbers, as explained above.

Task 3: Evaluation

a. Apply our outlier detection to the Complex9\_GN8 dataset obtaining a new file X; your outlier detection method is only applied to attributes x and y of the dataset, and ignores the attribute named class.

b. Sort X in descending order based on the values of attribute ols (the example with the highest ols value should be the first entry in X)!

c. Visualize the first 2% of the observations in X, just displaying their x and y value and the class using a different color for each class, in a display and the remaining 98% of the observations in a second display. In general, the first display visualizes the outliers and the second display visualizes the normal observations in the dataset.

d Visualize the first 5% of the observations in X, just displaying the x and y value and the class using a different color for each class and the remaining 95% of the observations in a second display.

e. Visualize the first 10% of the observations in X, just displaying the x and y value and class using a different color for each class, in a display and the remaining 90% of the observations in a second display.

f. Interpret the 6 displays you generated in steps c-e; particularly, assess how well does your outlier detection method work—intuitively observations that are quite far a way of the 9 natural clusters of the original Complex 9 dataset should be outliers. Also try to characterize which points are picked as outliers first (the top 2%), second (2% to 5% percentile), and third (5-10% percentile).

g. Create a histogram for the ols values of the top 10% entries in file X. Interpret the results. Moreover, look for gaps in the ols values in the file X; if observe any gaps, try you best to interpret why they occur!

Task 4 (optional): Enhance your outlier detection technique based on the feedback of Task3 and redo Task 3

Task 5: Write a 2-5 paragraphs, explaining your outlier detection technique works and how it was implemented. If you enhanced your approach based on feedback to get better results also describe how you enhanced your technique. If your outlier detection technique needs the selection of parameter values before it can be run, describe how you selected those parameter values for conducting Task3. Moreover, mention in a paragraph what (if any) external software packages your used in the project!

Task 6: Submit the code of the implementation of your outlier detection technique!

Finally, paste all project results for the different tasks into a single file and submit this file at the due date! Also be prepared to demo your software system.

.

1. The outlier detection method is only applied to the first 2 numerical attributes and the third attribute is just used for visualization purposes; for that reason, it is called a spatial outlier detection methods; e.g. it can detect outliers for object locations described by (longitude, latitude)-pairs. [↑](#footnote-ref-1)