Dr. Eick

COSC 4335*“Data Mining”* Assignment4 Spring 2018

*Design, Implementation and Comparison*

*of Two Outlier Detection Techniques for a Spatial Dataset*

Group Project (Groups of 2)

Due date: Monday, April 30, 2018 at 11p (4% bonus); however, we will still accept submissions until Wednesday, May 2, 11a; submissions received after this deadline will not be graded!

Last updated: April 18, 9p!



The goal of the project is to design and implement two bivariate, spatial outlier detection techniques of your own preference—each student develops a different one. You will apply the technique of your choice to the 2-dimensional dataset called *Complex9\_RN32*[[1]](#footnote-1) we already used in Assignment2, which, as you recall, is the variation of the Complex9 dataset with 32% random noise added to the original Complex9 dataset. Finally, you will compare the results with those of your team mate and write a summary report. Your outlier detection techniques 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 any other software library to accomplish the project tasks; just acknowledge what external software you used in your project report.

The Complex9\_RN32 dataset has attributes x, y, class; for example, after applying your outlier detection technique to 5 examples of the dataset the result produced by the method you are supposed to implement 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.

**Assignment4 Tasks:**

Task0: Visualize the Complex9\_gn16 dataset; visualize the third attribute using 9 different colors, similar to supervised scatterplots we used in Assignments 1/2.

Task1: Develop a 2D spatial outlier detection technique of your own preference that identifies abnormal data in datasets which contain pairs of numbers; each student develops a different 1

Task2: Design and implement 2 outlier detection techniques—one per student—for the Complex9\_RN32 dataset.

As explained earlier your implementation should add a column/attribute *ols* to the dataset and fill this column with numbers.

Task 3: Evaluation (one for each technique)

a. Apply our outlier detection to the Complex9\_RN32 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/the example that is the most likely outlier should be the first entry in X)!

c. Visualize the first 9% 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 91% 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 18% 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 82% of the observations in a second display.

e. Visualize the first 36% 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 64% 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 Complex9 dataset should be outliers. Also try to characterize which points are picked as outliers first 9%, 18%, and 36%, respectively

g. Create a histogram for the ols values of the top 36% entries in file X. Briefly, interpret the obtained histogram!

Task 4: Write a 2-5 paragraphs (for each technique), explaining each 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. Moreover, mention in an additional paragraph what (if any) external software packages your used in the project!

Task 5: Compare the two outlier detection techniques you developed in 2-4 paragraphs. Try to assess which technique worked better. Finally, write 1-2 paragraphs, describing what you leant from the project and if you enjoyed it or not. A significant amount of points will be allocated to this comparison; therefore, try to do a good job in comparing and evaluating your two techniques.

Submit the code of the implementation of your outlier detection techniques in a separate file! Finally, paste all project results for the different tasks into a single file and also submit this file at the due date! We only want one project report per group! Please upload your submission on Blackboard and name the submission file as HW4\_Group\_# (# denotes the group number).

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)