Hannah Andrews and Angela Hughes

Group 1

Homework 7 Outlier Detection

INFO 523

Spring 2020

Chapter 12 questions 12.1,12.4, 12.8, 12.9

## 12.1 Give an application example where global outliers, contextual outliers, and collective outliers are all interesting. What are the attributes, and what are the contextual and behavioral attributes? How is the relationship among objects modeled in collective outlier detection?

An application example where global, contextual, and collective outliers might be interesting could be a running app. Attributes would include run time, run mileage, date, location, and weather. Global outlier that would be interesting could be a run that was 24 hours long but only 2 miles long. The user likely left the app running and the developers might consider adding notifications for runs with this attributes to remind the user to stop the app or correct the logged run. An example of a contextual outlier might be someone that travels and runs in a different city than usual. This is contextual because whether the location of a specific run is an outlier depends on where the user usually runs. If developers identify when a user is running in a new location, they may consider sending push notifications for different routes to try in that area that the user was not aware of. An example of an interesting collective outlier might be ultra marathon runners, individuals running much higher mileage than other users. If there is a high density of high mileage users, developers might consider developing products/programs specific to that group of customers.

Contextual attributes include date, location, and weather. Behavioral attributes include running time and mileage.

To identify the ultra marathoners, we would use a clustering-based approach that may identify a dense cluster that is distant from the rest of the data.

## 12.4 Using an equal-depth histogram, design a way to assign an object an outlier score.

An equal-depth histogram has the same number of objects in each bin. The width of the bins will be different. In an equal-depth histogram, bins with outliers will have a low density. Density can be calculated by multiplying 1/N (the number of objects) by (the number of objects in the bin)/(bin width). Bins with a low density relative to other bins should be investigated for outliers.

## 12.8 In outlier detection by semi-supervised learning, what is the advantage of using objects without labels in the training data set?

Objects without labels in the training data set allow the model to detect normal objects based on their similarity to other labeled, normal objects. This allows the model to discriminate outliers based on dissimilarity to its detected pattern, whereas a fully-labeled data set would have restricted the freedom of the model to discriminate based on its own learning.

Semi-supervised learning works best when most of the unlabeled data objects are normal. However, problems arise when too many of the outliers are unlabeled, as it becomes harder to discern a pattern that comprises normal objects.

## 12.9 To understand why angle-based outlier detection is a heuristic method, give an example where it does not work well. Can you come up with a method to overcome this issue?

Angle-based outlier detection is only an approximation of what it would be like to compute the Angle-Based Outlier Factor for every point. Computing angles for each point would be too costly in high-dimensional data. Angle-based outlier detection uses the variance of angles between 3 points to determine whether the point o in the angle (a,o,b) is an outlier. outliers will have less angle variability than points that are in the center or at the border of clusters.

In high-dimensional space, it is difficult to use distances between points to distinguish outliers from noise. High-dimensional sparse data is especially difficult. Angle-based outlier detection might not work well in sparse situations. It would be easier to distinguish outliers if there were more data points filling in the sparse gaps, to which the outlier could be compared.

One method to overcome this issue would be dimensionality reduction, such as Principal Component Analysis (PCA) and Discrete Fourier Transform.