Data Cleaning - Project 1

AUTHOR
Robert Johnson

PUBLISHED
October 14, 2023

Table of contents

1 What is an outlier?

The book offers a broad definition for an outlier "an outlier is an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism" (Hawkins 1980). For example, if the average score on an exam was an 80 a grade of 0 would be considered an outlier. Different outlier detection techniques have differing ways in how they determine normal behavior and consequently what is and is not an outlier. Data visualization tools can be helpful for identifying potential outliers and selecting the right outlier detection techniques.

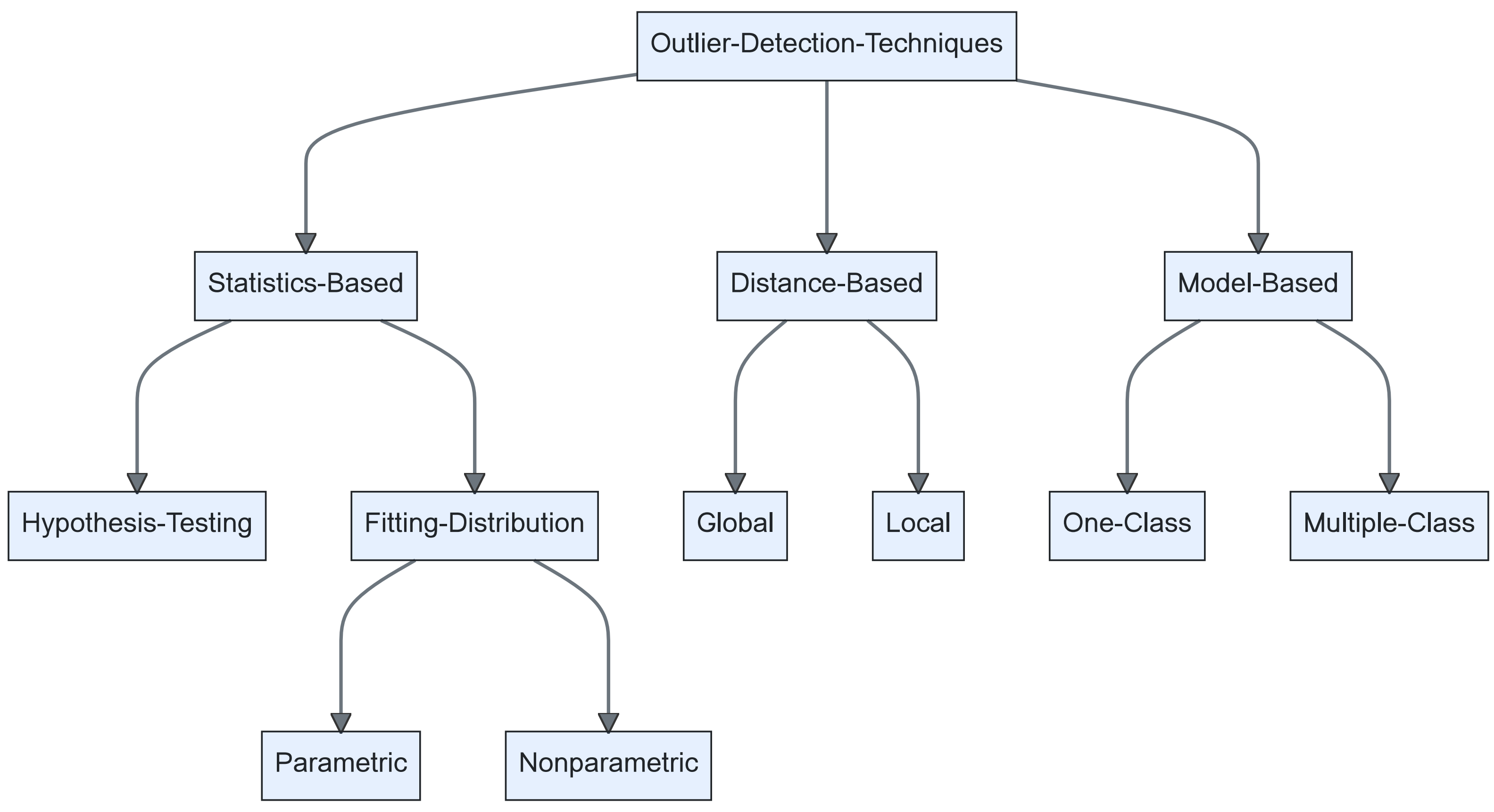
▶ Code

```
# A tibble: 9 × 4
  Name                  Age Income   Tax
  <chr>               <dbl>  <dbl> <dbl>
1 Vivian Baskette         1     70     7
2 Jamison Marney         25    110    11
3 Marie Mulero           27     80     8
4 Trudi Kimmell          30    130    13
5 Stephanie Lindemann    32    120     7
6 Dia Werley             35     80     8
7 Abbie Lama             40     90     9
8 Misti Luce             41    100    10
9 Wilda Byerly         1000    120    12
```

We can see in the above example the Age feature has two potential outliers that immediately stand out. The other features are not so obvious where outliers may be.

2 Outlier Detection Techniques

A basic taxonomy of detection models can be seen as follows (Ilyas et al., 2019)



2.1 Statistics-Based

These techniques define a normal value as one that appears in the high probability region of a stochastic model, and an outlier vice-versa.

2.1.1 Hypothesis Testing

This type of testing can be seen in the Grubbs Test, and Tietjen Moore Test. Typically these methods will define and calculate a test statistic and declare a significance level/critical area. The null hypothesis states that no outliers exist, and the alternative states the opposite. When the test statistic is computed for the observed data if it falls within the critical region the null hypothesis is rejected.

2.1.2 Model Fitting

Model fitting attempts to fit a known distribution or *pdf* to the data-set. Data-points that have a low probability according to the fitted distribution are declared as outliers.

2.2 Distance-Based

These techniques used the distance between the data-points to define normal behavior. Data-points that are far away from others are assumed to be outliers.

2.2.1 Global Approach

This approach considers all other data-points when determining distance for each individual point.

2.2.2 Local Approach

This approach uses a 'neighborhood' that can be defined as needed to determine the distance for each data-point.

2.3 Model-Based

These techniques take advantage of labeled data-sets to train a classifier-model then apply this model to a data-point to determine whether it is normal or not.

2.3.1 Multi-Class

This approach makes the assumption that the training data contains data that has multiple normal classes

2.3.2 One-Class

This approach makes the assumption that all of the training data points to one normal class

3 PCA

Principal Component Analysis is a popular technique used to analyze large data-sets that have high dimensionality. It reduces the dimensionality of the data-set without losing much information. It works by computing two principal components. The first component is computed so that it explains the most variance in the original features, and the second is computed to explain the most variance left after the first component. This is extremely useful in large datasets as it reduces it down to just a few key features.