Visualizing Big Data Outliers through Distributed Aggregation

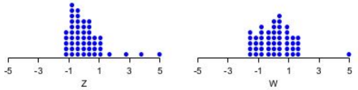
Leland Wilkinson. Proc VAST 2017, TVCG to appear.

Theodore Smith
CPSC 547
Nov 7, 2017

Outliers

- General definition
  - Observations which appear to be inconsistent with the remainder of a set of data (Barrett and Lewis)
- Principles of detection
  - Each observation represents a point in vector space of a random variable
  - Likelihood that a point outlies the distribution of a sample is proportional to the probability that the point is a member of the distribution

Example



The Gaps Rule

- Looks for gaps in data that do not match assumed generating distribution
- Can detect aberrations in the middle of a distribution, not just at its extremes

$$Q = \frac{x_n - x_{n-1}}{x_n - x_1}$$

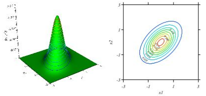
Dixon

$$f(x_i;\theta_i,\phi) = \exp\left[\frac{x\theta - a(\theta)}{b(\phi)} + c(x,\phi)\right]$$

Burridge and Taylor

Higher-Dimensional Outlier Detection

- **Mahalanobis Distance**
  - Detects outliers based on Euclidean distance of multidimensional point from centroid of multivariate Normal distribution
    - Only valid if assumption of normality is satisfied
  - Squared Mahalabobis distance = chi-square variate with *p* degrees of freedom



Higher-Dimensional Outlier Detection

- **Clustering**
  - Process:
    - Pre-cluster data
    - Target points with large distance from nearest cluster
  - Effective for samples of moderate size with limited singleton frequency
  - Does not typically scale well for larger data sets
    - Outlier aggregation
    - Convergence in Euclidean space
    - Efficiency
  - Generally not based on probability model
    - Susceptible to error

hdoutliers

- **Purpose**
  - Statistical method for identifying subsets of data which do not match underlying distribution of sample
  - Generate highlighted points representing outliers in visualization of data
- **Design Criteria**
  - Identify outliers in mixed data sets containing both ordinal and categorical variables
  - Exploit random projection for a large number of dimensions
  - Handle large sets through single-pass aggregation
  - Overcome masking effects resulting from interaction of outlying points
  - Function for both univariate and multivariate data

hdoutliers

- **Algorithm**
  1. Convert all categorical variables to continuous variables
     a. Correspondence Analysis
  2. If > 10,000 columns, reduce via random projections using error bound to squared distances
  3. Normalize resultant columns
  4. Initialize *exemplars*
     a. Initializes with row 1 as sole member of set
     b. Rows added to *exemplar* set if row distance from existing exemplars exceeds threshold
  5. Initialize *members*
     a. List of lists with initial entry defined by rows in *exemplars*.
     b. Each *exemplar* has list of affiliated members

hdoutliers

- **Algorithm**
  6. Single pass

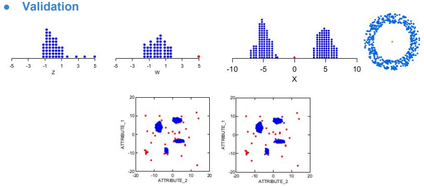
```
forall the row(i), i = 1, ..., n do
    d = distance to closest exemplar, found in exemplars
    if d < δ then
        add i to members list associated with closest exemplar
    else
        add row(i) to exemplars
        add new list to members, initialized with [i]
    end
end
```

$$\delta = .1/(\log n)^{1/p}.$$

  7. Compute nearest distances between all pairs of exemplars
  8. Fit exponential distribution to upper tails nearest-neighbor distances
  9. Flag members associated with exemplars exceeding distance cut-off (1-0.05 from CDF of previous step) from other exemplars as outliers

hdoutliers

- **Validation**



hdoutliers

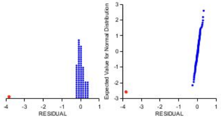
- **Visualization**

Core principles:

1. **Probability-grounded algorithm necessary for reliable outlier detection**
   a. Risk of outlier classification unknown without statistical foundation
2. **Visual analysis necessary to derive meaning from algorithmic detection**
   a. Highlighting cases based on probabilistic detection guides discovery

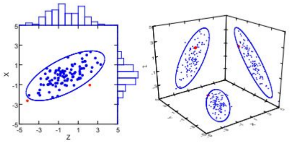
hdoutliers

- **Visualization**
  - **Univariate data**
    - Dot plots and probability plots



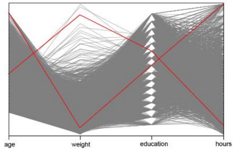
hdoutliers

- **Visualization**
  - **Low-Dimensional Visualizations of High-Dimensional Data**



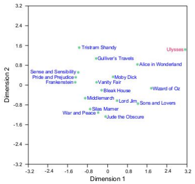
hdoutliers

- **Visualization**
  - **Parallel Coordinates**



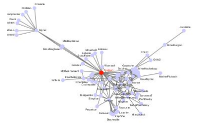
hdoutliers

- **Visualization**
  - **Text Data**



hdoutliers

- **Visualization**
  - **Graph Outliers**
    - Featurize nodes based on some metric (betweenness centrality, prominence, average degree of neighbors, etc.)
    - Feed features into hdoutliers
    - Highlight outlying nodes



Conclusions

- Identification of outliers is only valuable if the assumptions that differentiate them from a sample are valid
- Methods that include outliers in estimation of parameters for a given distribution are circular and unreliable
- The risk of excluding outliers is unknown if the probability of accurate detection is not calculated
- VIsualization of outliers in context, particularly for high-dimensional data, is essential for extracting information regarding the features which set them apart