Toward An Outlier Uncertainty Model: A Comparative Analysis

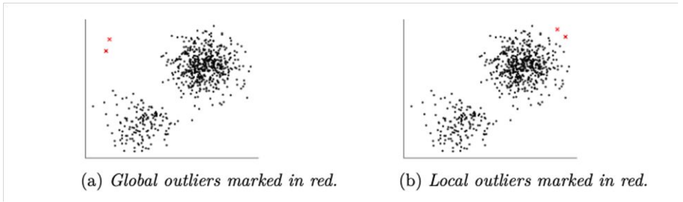
Sophia Lin
slin02@email.unc.edu

Ghulam Quadri
quadri@ou.edu

Danielle Szafir
dnszafir@cs.unc.edu

Motivation:

Outliers can make data visualizations difficult to interpret, and are usually removed. However, there exists ambiguity in outlier identification due to the messy data and unclear cluster boundaries of real world datasets. I investigated the practical applicability of eighteen different methods for a perception-based model that estimates outlier uncertainty in two-dimensional scatterplots through a comparative analysis.



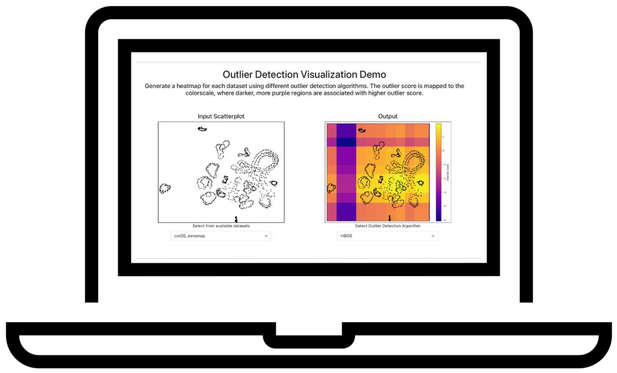
(a) *Global outliers marked in red.* (b) *Local outliers marked in red.*

Feature Analysis of Algorithms

Type	Algorithm	Size of cluster	Distance to nearest cluster	Tail probabilities	Probability	Distance to kth nearest neighbor	# of nearest neighbors	Pairwise distance between points	Dimensionality	# of input points
Proximity-based	Cluster-based Local Outlier Factor (CBLOF)	☑	☑	☐	☐	☐	☐	☐	☐	☐
	Histogram-based Outlier Detection (HBOS)	☐	☐	☐	☐	☐	☐	☑	☐	☐
	K-Nearest Neighbors (KNN)	☐	☐	☐	☐	☑	☑	☑	☐	☑
	Local Outlier Factor (LOF)	☐	☐	☐	☐	☑	☑	☑	☐	☐
	Density-based Spatial Clustering of Application with Noise (DBSCAN)	☐	☐	☐	☐	☑	☑	☑	☐	☐
Probabilistic	Angle-based Outlier Detection (ABOD)	☐	☐	☐	☐	☐	☐	☑	☑	☐
	Probabilistic Mixture Modeling for outlier analysis (GMM)	☐	☑	☑	☐	☐	☐	☐	☑	☑
	Outlier Detection with Kernel Density Functions (KDE)	☐	☐	☐	☐	☑	☑	☑	☑	☑
	Unsupervised Outlier Detection Using Empirical Cumulative Distribution Functions (ECOD)	☐	☐	☑	☐	☐	☐	☐	☑	☐
	Copula-based Outlier Detection (COPOD)	☑	☐	☑	☐	☐	☐	☐	☑	☐
Ensembles	Feature Bagging (FB)	☐	☑	☐	☐	☑	☑	☑	☐	☑
	Isolation Forest (IF)	☐	☐	☐	☐	☐	☐	☐	☐	☐
	Locally Selective Combination of Parallel Outlier Ensembles (LSCP)	☐	☐	☐	☐	☑	☑	☐	☐	☑
	Isolation-based Anomaly Setection using Nearest Neighbor Ensembles (INNE)	☑	☑	☐	☐	☑	☐	☑	☐	☐
Linear Models	Minimum Covariance Determinant (MCD)	☑	☐	☐	☐	☐	☑	☑	☑	☐
	One-class SVM (OCSVM)	☐	☐	☐	☑	☐	☐	☐	☑	☐
	Principal Component Analysis (PCA)	☐	☐	☐	☑	☐	☐	☐	☑	☐
	Deviation-based Outlier Detection (LMDD)	☐	☑	☐	☑	☐	☐	☐	☑	☐

Outlier Detection Visualization Demo:

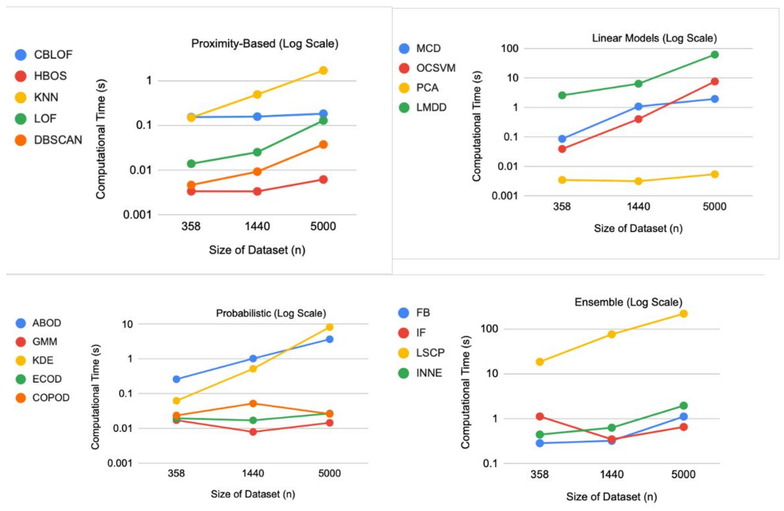
I created a simple demo using React that can be used to view the heatmap visualizations of outlier scores for different algorithms and datasets.



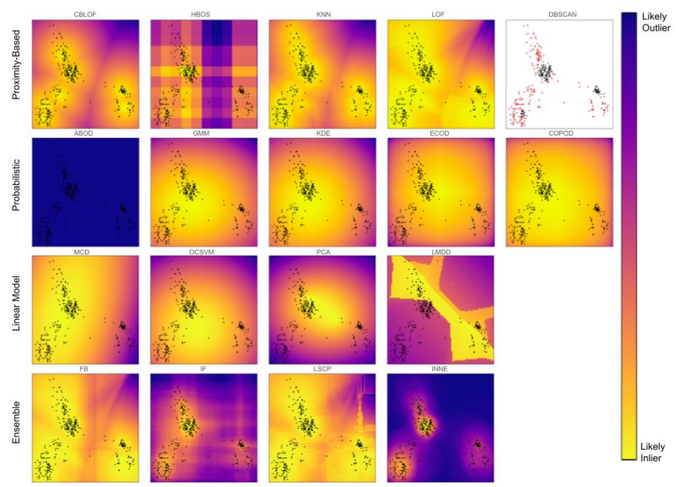
Available at https://slin02.github.io/outlier_vis/

Computational Complexity

Both theoretical complexity and real runtimes were examined. All algorithms were run on two-dimensional datasets of different sizes (n=358, n=1440, n=5000) for three trials, taking the average time.
Proximity-based and probabilistic methods tended to be the fastest, such as PCA and HBOS. Linear models and methods that rely on nearest-neighbor calculations like KNN and LSCP tended to scale poorly.



Heatmap Visualization:



Conclusion and Future Work:

This comparative analysis provides insight into some strengths and limitations of different outlier detection methods.
Key points:

- Many methods employ the classical techniques of KNN and LOF
- Outlier ensembles tend to be the most robust, but can also be more computationally complex
- While overall time complexity is important, scalability should also be a significant consideration
- Proximity and ensemble models tend to make for better heatmap visualizations, as the color gradients tend to better reflect the shape of the clusters

This study is a preliminary step towards the development of a model for estimating outlier uncertainty in 2D scatterplots. Future research efforts should aim to ensure that the model is robust and versatile enough to accommodate a broad range of clustering patterns while remaining scalable and efficient.

Next steps: A qualitative study to examine the visual factors involved in perceptual ambiguity in order to design a perception-based model for estimating outlier uncertainty.

Additional thanks to members of VisuaLab

UNC COLLEGE OF ARTS AND SCIENCES
Computer Science