Charu C. Aggarwal
IBM T J Watson Research Center
Yorktown Heights, NY

Ensemble-Centric Evaluation of Outlier Detection Algorithms

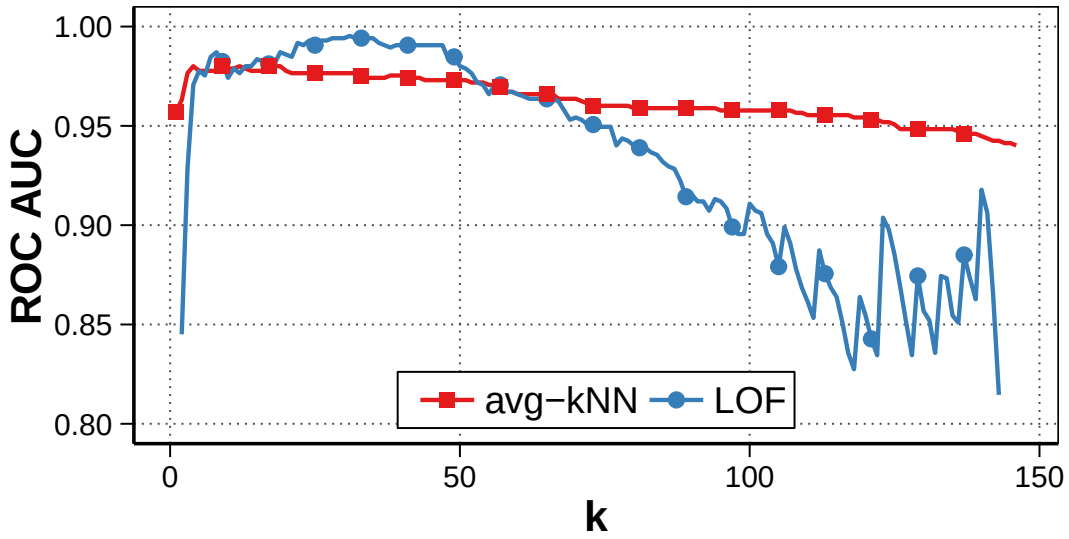
Introduction

- Evaluation of outlier detection algorithms is notoriously difficult
  - Outlier detection is unsupervised $\Rightarrow$ No labels at the time of algorithm execution
  - Relative performance of algorithms depends on setting and data distribution $\Rightarrow$ No guidance from supervision
  - Need a testing approach that is robust to setting and data distribution $\Rightarrow$ Ensembles

Comparability of Different Algorithms

- Cannot compare algorithms with different hypotheses of what a "normal" class distribution is:
  - The nearest neighbor class behaves very differently from kernel Mahalanobis method
  - Relative performance of different models may vary drastically with data distributions
  - Want to use ensembles to obtain design principles that are robust across data distributions

Parameter Setting Dilemma



- *Example from Lymphography Data Set:* LOF has better best-case performance but average k-nearest neighbor is stable.
  - Unsupervised nature of the problem makes it challenging to make decisions
  - Analyst has no idea of quality (until after the fact)

Other Parameter Setting Dilemmas

- An algorithm with many parameters can be easily optimized with cross-validation in supervised settings
  - Notoriously hard to optimize in unsupervised settings
  - One class-SVMs: Can perform excellently for some parameter choices and very poorly for others
- **Inadvertent benchmarking problem:** Contamination from labels via repeated testing favors algorithms with many parameters

The Effect of Ensembles

- Performance of algorithm is sensitive to whether it is wrapped in an ensemble
  - Ordering of base detectors different from ensemble-centric variations
  - Cannot compare an ensemble detector (e.g., isolation forest) easily with a single implementation of k-NN

Goals of Testing

- Not intended to provide a panacea or single solution
  - Intend to show how complicated the picture really is
  - Some methods like Isolation Forest do really well in our tests
    - Seems to do well in common rare class settings
    - Can be disastrously bad for pathological cases
- Main takeaway is to emphasize the importance of ensembles

Integrate Testing with Design Principles

- Find which algorithms are robust within a particular family of algorithms (e.g., NN-algorithm or linear model).
  - Not meaningful to compare different distribution assumptions of what the normal class is.
- Identify how one can optimize a particular type of outlier model in parameter independent way (e.g., ensembles) before evaluation
- Find which algorithms are most correlated in terms of performance
  - Useful for creating a heterogeneous detector combining multiple algorithms

Creating an Ensemble-Centric Model from a Single Family

- A useful ensemble method is variable subsampling.
- Create a subsample of size varying between two ranges.
- Apply outlier detection algorithm to each subsample.
- Average outlier scores from different models.
- Comparing unoptimized k-NN against an isolation forest is not fair!
  - Ensembles of 1-NN algorithms perform similar to an isolation forest and even have semantic connections with one another

Advantages of Variable Subsampling

- Variable subsampling can often perform implicit parameter space exploration of some detectors by fixing the parameter values and varying the size of the sampled data set.
  - A k-nearest neighbor detector would automatically use varying percentile values of k over different subsample sizes.
  - Plays the dual role of enabling both diversity and parameter-space exploration.

Models Tested

- **Distance- or Density-based:** Average k-NN method, Exact k-NN method, LOF, Isolation Forest
- **Soft PCA (Mahaanobis):** Distance from centroid after scaling each principal component to unit norm
- **Kernel Mahalanobis:** Do the above in kernel space
- **Regression-based:** (GASP, ALSO)– Predict each feature from remaining features and average RMSE across features

Hidden Gems

- Many top detectors such as isolation forests are well known.
- Some lesser known detectors seem to work well in ensemble context (Hidden Gems)
  - Linear regression-based detectors (GASP, ALSO)
  - Kernel Mahalanobis method
- Described in subsequent slides.

ALSO: Base Detector

- Decompose an outlier detection problem on d-dimensional data into d supervised problems.

- These d supervised problems are constructed one by one by selecting each of the d dimensions as the target attribute in turn and using the other $(d-1)$ attributes to predict it.
  - Off-the-shelf classification/regression modeling problem
  - One can use any model such as random forests, least-squares regression, and kernel regression.

- Combine scores into a single RMSE score

ALSO-E: Ensemble-Centric Variant

- For each base detector, a value of s is chosen uniformly at random from $(\min\{n, 50\}, \min\{n, 1000\})$.
- A sample of size s is drawn from the data and a training model is constructed on using an off-the-shelf learner.
- An attribute is randomly sampled from the data as the target attribute and the remaining attributes are used as the predictors.
- The RMSE error is computed for this target attribute on the remaining $(n - s)$ out-of-sample points.
- Averaged over multiple samples.

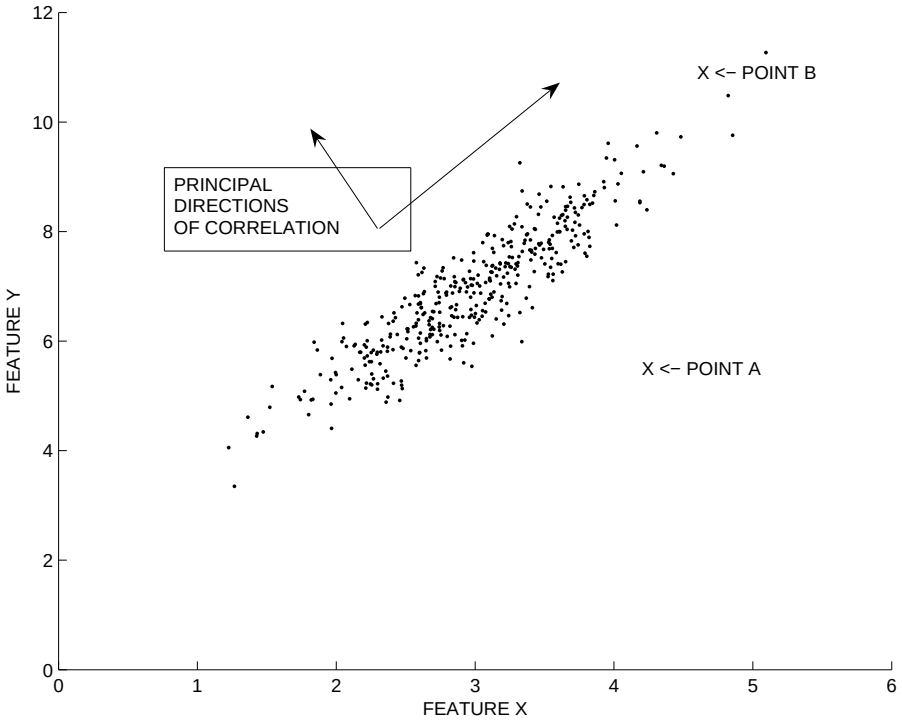
GASP: Group-wise Approach

- Partitions the d attributes into a set of r target attributes and $(d-r)$ predictor attributes.
- In each iteration, we randomly sample r target attributes and the remaining $(d-r)$ attributes are treated as predictor attributes.
- Use RMSE error to create outlier scores.
- Similar ensemble-centric variants as ALSO

Recap of Mahalanobis Method

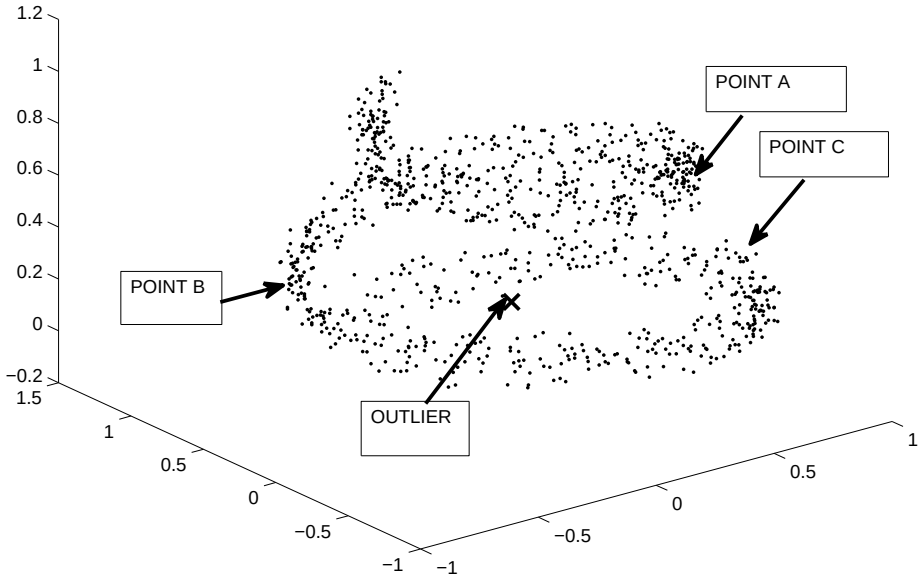
- Outlier score is Mahalanobia distance to centroid of data set
- Equivalent algorithm: Transform the data using PCA and normalize along principal component directions
- Distance to centroid is the Mahalanobis distance
- Principal components for transformation $\Rightarrow$ eigenvectors of covariance matrix
  - Useful alternative: Eigenvectors of similarity matrix directly provides embedding

Outliers that Mahalanobis method finds



- Global extreme values

Outliers that Mahalanobis method can't find

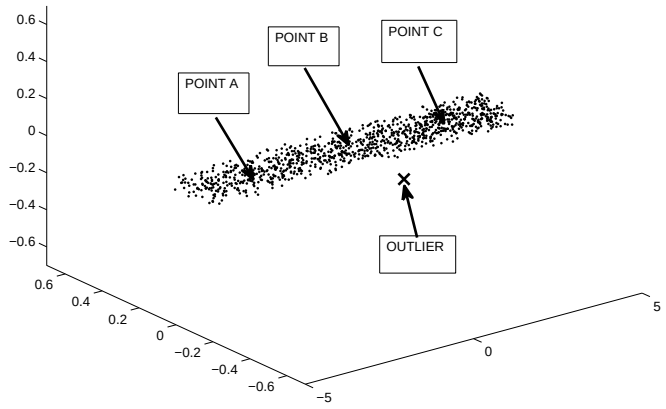


- Not extreme values

Kernel Mahalanobis Method

- Compute the similarity matrix of the data set using a kernel
- Find *all* its normalized eigenvectors: Embedding
  - Similarity matrix is efficient to compute in subsampling settings
  - Can be generalized to subsample-based approach.

Outliers that Kernel Mahalanobis method finds



- Merit of space transformation

Kernel Mahalanobis Method: Generating Full Representation from Subsample

- Find eigenvector matrix P of $s \times s$ similarity matrix (subsampled) $S_{in} = P\Sigma^2 P^T$ and eigenvalues in Σ^2

- Find $n \times s$ similarity matrix S between all points and in-sample points.

- Rows of $n \times s$ matrix $F = SP\Sigma^{-1}$ provide unnormalized embedding of all points

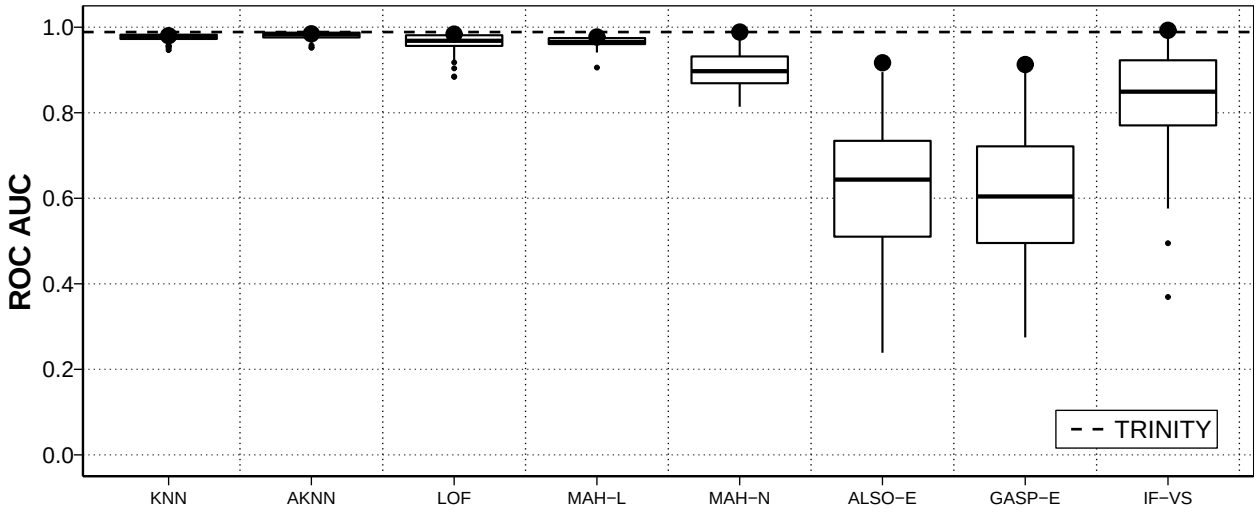
Generating Outlier Scores from Subsample

- Standardize F to have zero mean and unit variance.
- The L_2-norm of each row provides the subsampling-specific outlier score.
- Repeat over many subsamples and average the outlier score.

Experimental Results

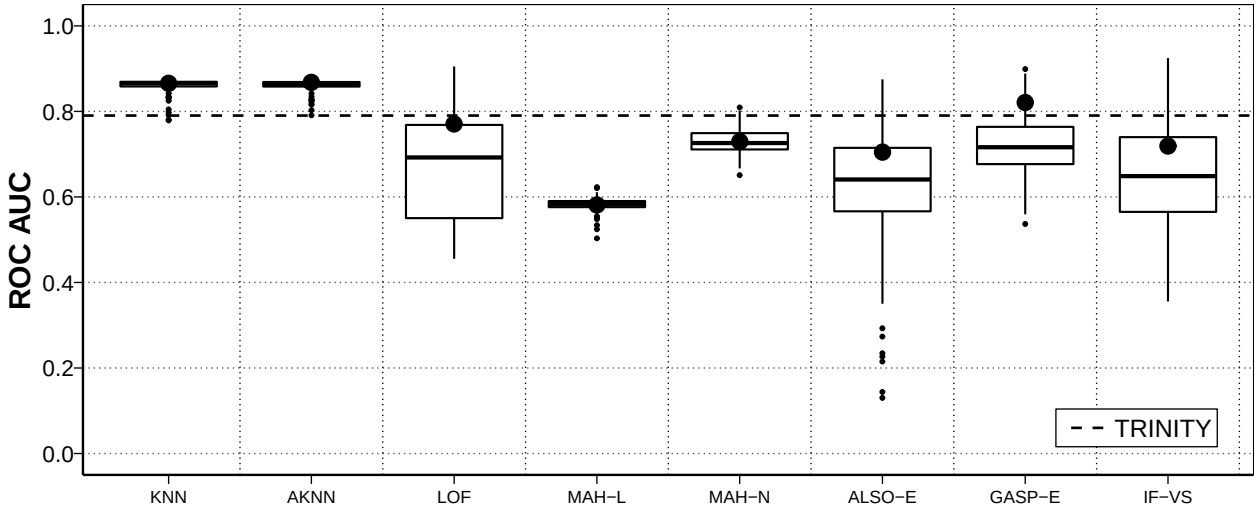
- Implemented each of the detectors with variable subsampling
- Show both performance of base detectors and ensembles with box plots.

Lymphography



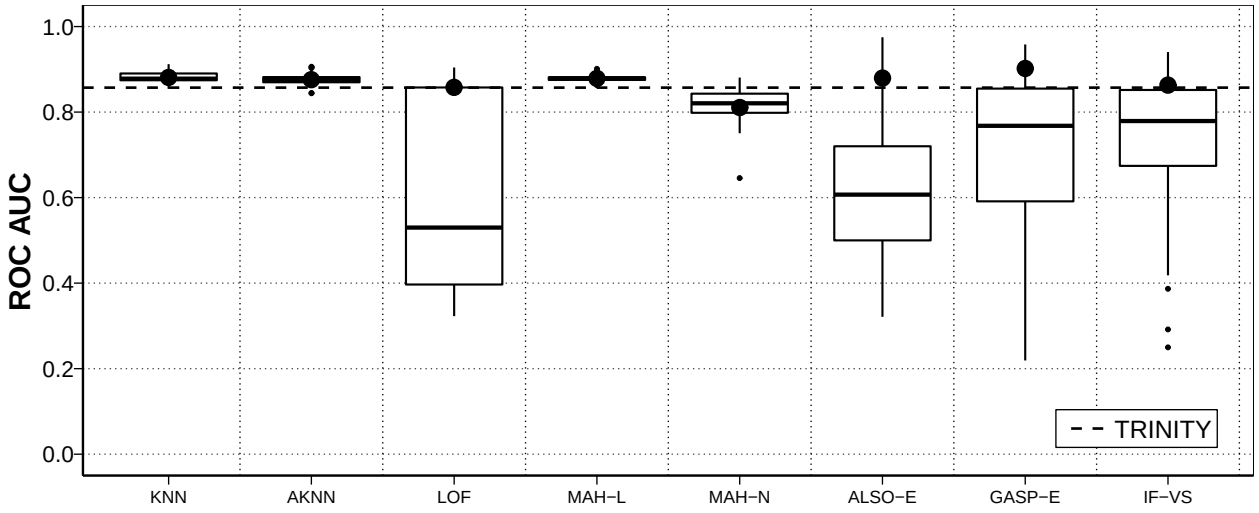
- Large variation in base detectors but similar performance of ensemble

Glass



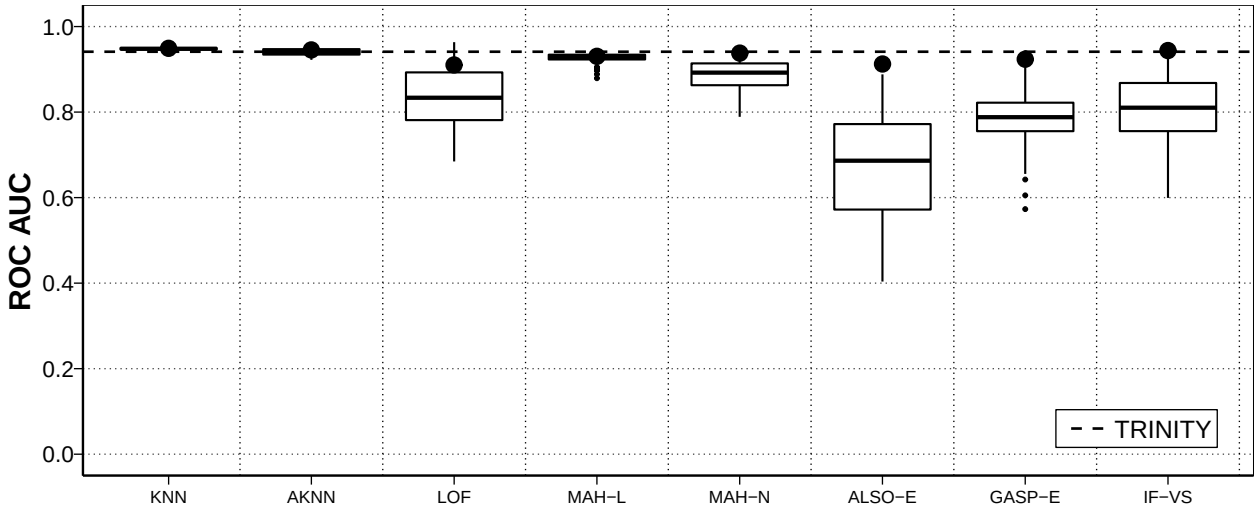
- Glass is among the few data sets were Isolation Forest does poorly

Ecoli



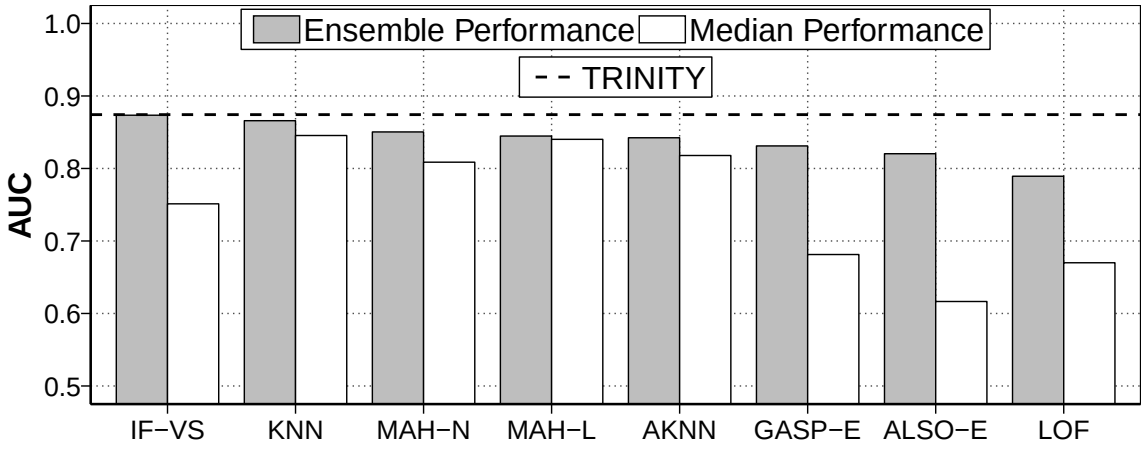
- Huge variation in base detectors but similar performance of ensemble

Wisconsin Breast Cancer



- Huge variation in base detectors but similar performance of ensemble

Median Base Performance vs Ensemble Performance



- The great equalizing power of ensembles

Some Observations

- Extreme value detectors like the Mahalanobis method seem to do surprisingly well
  - Data sets derived from high-dimensional classification problems
  - Linear separability of high dimensional data tends to favor detectors biased towards extreme value analysis
  - *Detectors that are not specifically focused on finding extreme values but are biased towards it will tend to do very well.*

Observations

- The differential performance between the Mahalanobis method and kernel Mahalanobis method can be used to infer when outliers are not (multivariate) extreme values

- Kernel Mahalanobis method is able to find both types of outliers

Observations

- The kernel Mahalanobis method was highly uncorrelated to other detectors and performed very well.

- Isolation Forests and k-NN were somewhat correlated to one another and performed impressively.

- The Isolation Forest provided the best overall results.
  - Occasionally performed poorly on some data sets

TRINITY Ensemble

- Combine three different subsampled ensembles:
  - Kernel Mahalanobis ensemble
  - Exact k-NN ensemble
  - Isolation Forest with Variable Subsampling
- TRINITY is an ensemble of ensembles

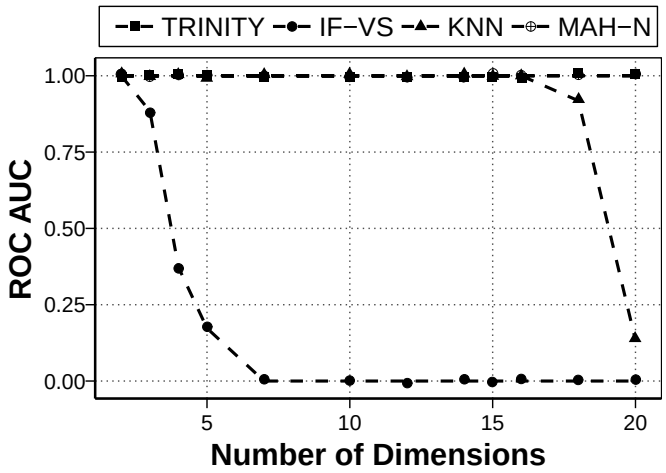
The Case of Isolation Forests

- Excellent performance across a range of data sets
- Occasionally performs very poorly (e.g., Glass data set).
- The cut-based approach tends to discriminate against outliers in the interior of the data.
- Could this excellent detector have hidden weaknesses?

A Pathological Data Set for Isolation Forests

- Generate normal data on the surface of a spherical ball
- Place a single outlier at the center of the ball
- Referred to as *ball-and-speck data set*.
- An isolation forest tends to do surprisingly poorly with increasing dimensionality on the ball-and-speck data set

A Pathological Case for Isolation Forests



- Effect of increasing dimensionality

What Can be Achieved with Ensembles

- Reducing effects of variance for a single detector type

- Protecting against disastrous performance of a particular detector because of quirks of data set

What Cannot be Achieved

- Consistently performing better than all detectors in a heterogeneous combination

- Tailoring detector selection and importance to specific types of data sets

Overview and Future Directions

- Real data sets have inbuilt disadvantages

- What about synthetic data?
  - Danger that algorithm can be tailored to data
  - Variational models can be used to encode outliers modeled on real data