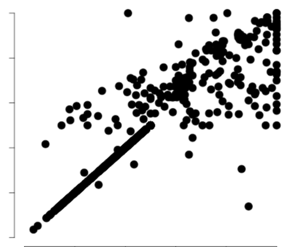
Classifier-Adjusted Density Estimation for Anomaly Detection and One-Class Classification

Lisa Friedland, Amanda Gentzel, David Jensen

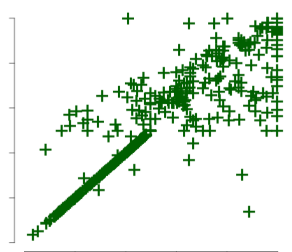
School of Computer Science
University of Massachusetts Amherst

Method Overview

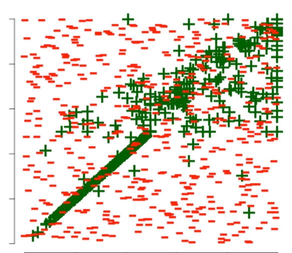
- Classifier-adjusted density estimation (CADE) detects anomalies by identifying low-probability instances in large, multivariate data sets.
- CADE estimates the joint probability density function of its training data by using a classifier to “correct” a naive density estimate.



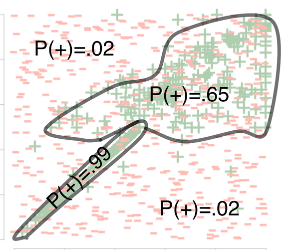
1. Start with unlabeled data.



2. Label original data positive (non-anomalous). Construct a naive density estimate of the positives → P(X | A).



3. Generate pseudo-negatives (pseudo-anomalies) from P(X | A).



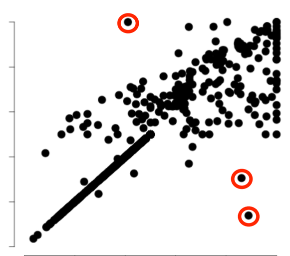
4. Train a classifier to distinguish the positives from the pseudo-negatives.

5. Combine classifier’s prediction with initial density estimate to compute a final density estimate → P(X | T)

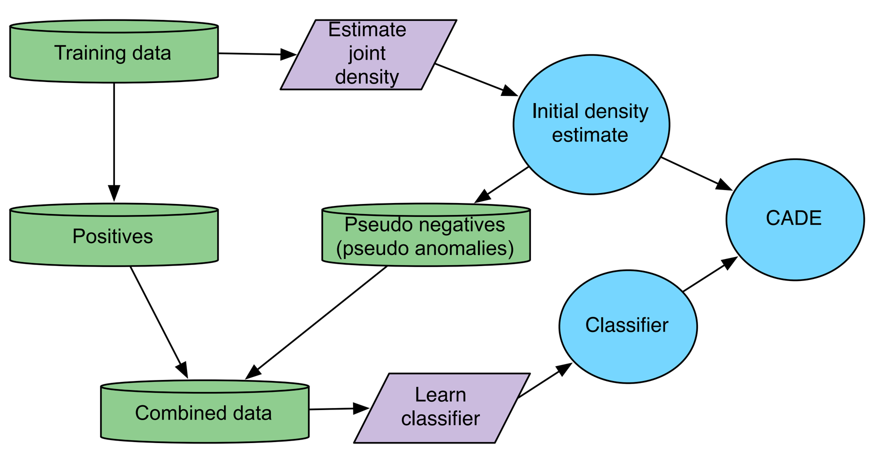
$$\underbrace{P(X \mid T)}_{\text{Final density estimate}} = \frac{\overbrace{P(X \mid A)}^{\text{Initial density estimate}} P(A)\ \overbrace{P(T \mid X)}^{\text{Classifier prediction}}}{\underbrace{P(T)}_{\text{Training set class proportions}}\,(1 - P(T \mid X))}$$

[Hempstalk, Frank, Witten. PKDD 2008]

6. Apply final density estimator P(X | T) to unlabeled data to identify anomalies.



Workflow



Experiments for semi-supervised anomaly detection:

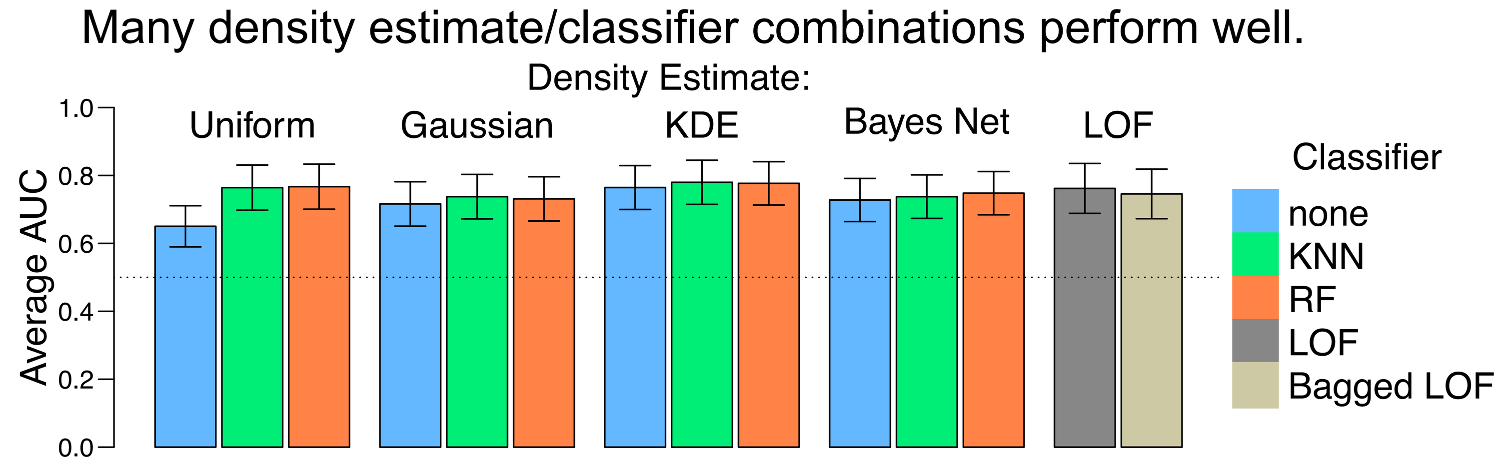
- 13 UCI datasets → 76 class divisions (positive vs. anomalous)
- 5 classifiers (Weka)
- 4 initial density estimates
- 10-fold cross validation

Summary

- High-quality anomaly detection is possible in multivariate data with a relatively simple method that estimates a joint probability function.
- Experimental evidence across a range of data sets shows CADE to be competitive and scalable.
- Within CADE, simple components often work well:
  - Marginally independent initial density estimates
  - Adjusted by random forest or *k*-nearest neighbor classifier
- Probability density estimators are more robust than local outlier factor methods to the challenge of irrelevant attributes.

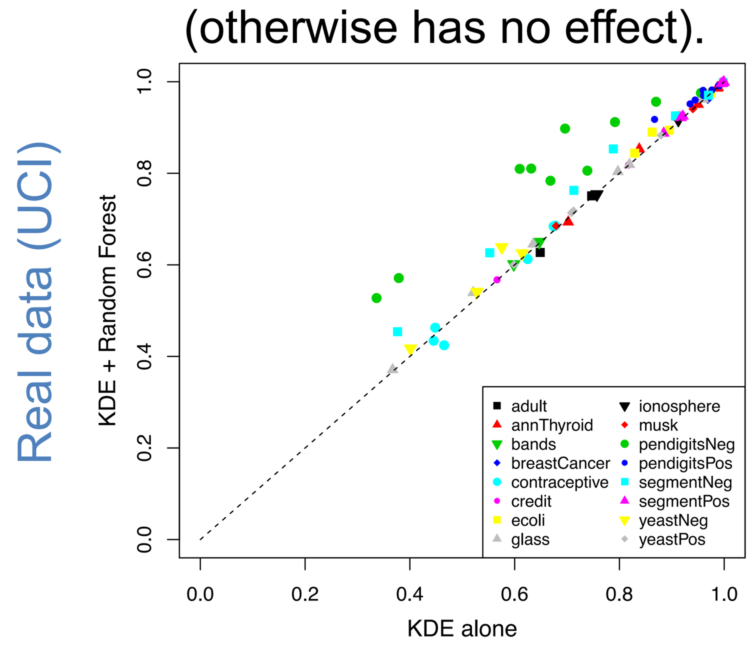
Algorithm Components and Performance

Many density estimate/classifier combinations perform well.

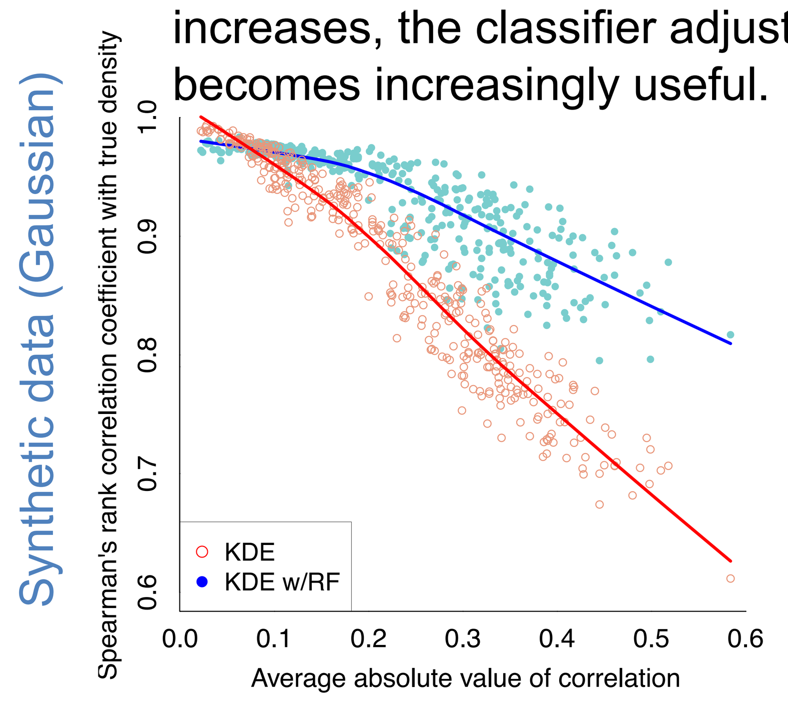


Average area under ROC curve over 76 class divisions of UCI data

Classifier adjustment often improves upon initial density (otherwise has no effect).



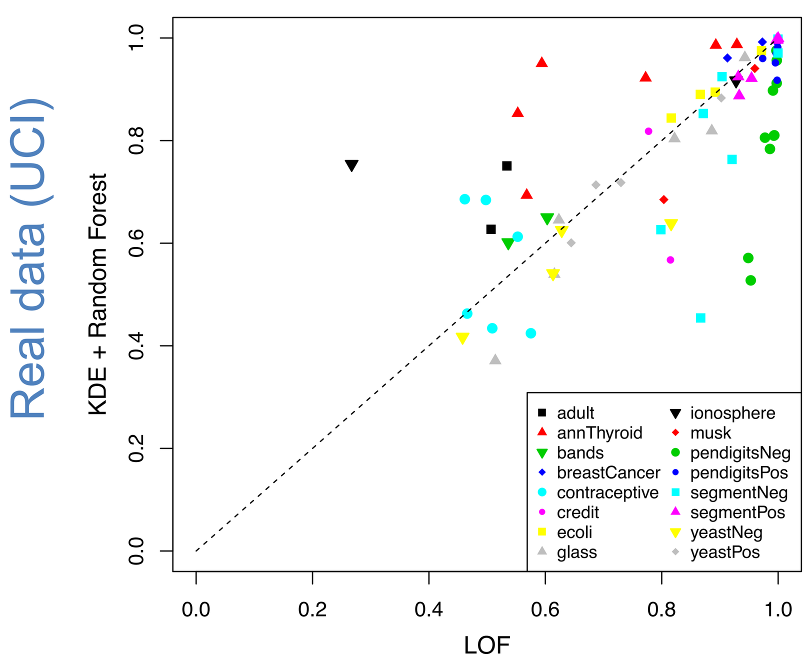
As correlation among features increases, the classifier adjustment becomes increasingly useful.



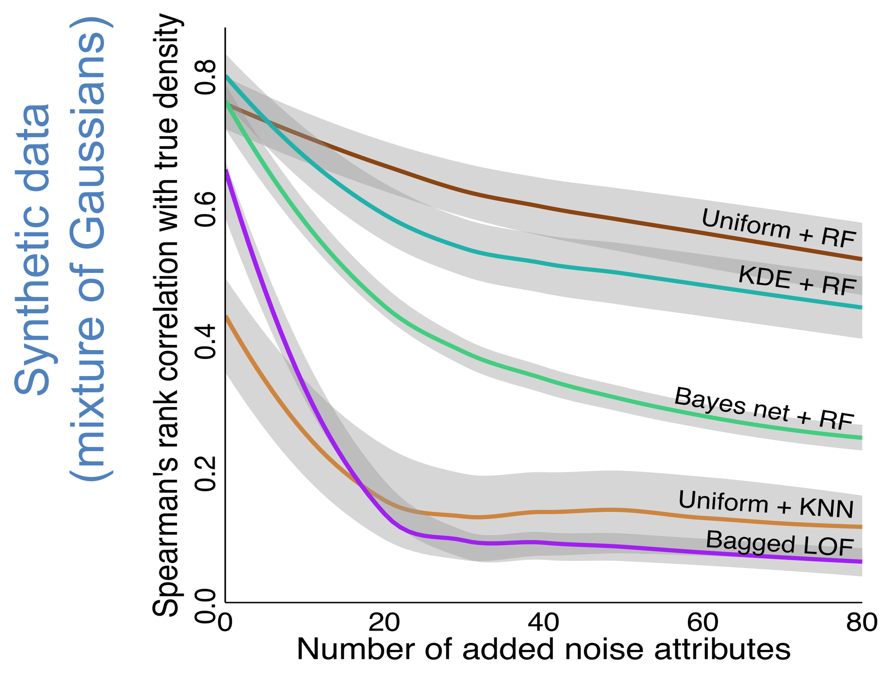
Comparison with Local Outlier Factor

[Breunig, Kriegel, Ng, Sander. SIGMOD 2000]

CADE performs competitively with LOF (varies by data set).



Robustness to irrelevant attributes: when uniform noise attributes are added, LOF degrades quickly. CADE is much more resistant.

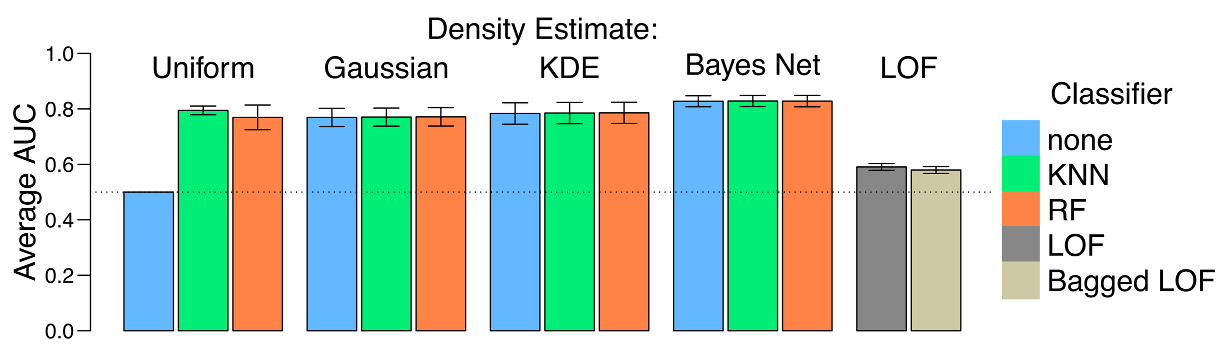


Unsupervised Runs on Large Data Sets

Data Set	Employee
Source	Collected for DARPA ADAMS project on insider threat detection. Describes computer activities of ~5500 employees of a real company.
# features	88
# instances	108,215 to 133,770 (6 separate months)
# anomalies	8 to 98
Avg. runtime	368.1 sec

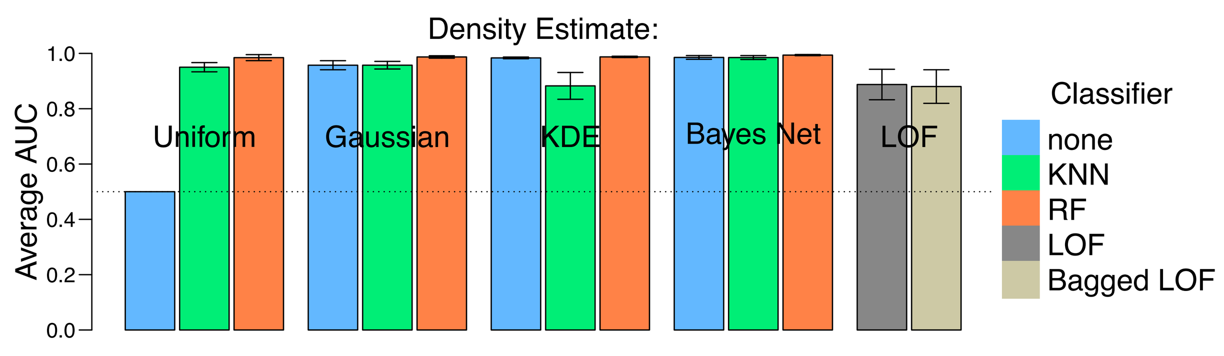
Data Set	Shuttle
Source	UCI
# features	9
# instances	45,596 to 54,489
# anomalies	10 to 8903
Avg. runtime	104.3 sec

Employee



Average area under ROC curve over 6 months x 10 runs each

Shuttle



Average area under ROC curve over 5 class divisions x 10 runs each

Funding was provided by the U.S. Army Research Office and Defense Advanced Research Projects Agency under Contract Number W911NF-11-C-0088. The content of the information in this document does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.