# Anomaly Detection by Combining Decision Trees and Parametric Densities

**Matthias Reif**[1], **Markus Goldstein**[1], **Armin Stahl**[1], **Thomas M. Breuel**[2]

[1]German Research Center for Artificial Intelligence (DFKI), Kaiserslautern
[2]Technical University of Kaiserslautern

## Introduction

- Anomalies are very different from normal data points and occur only rarely
- One class missing at training
- Classify unusual instances as anomaly
- Different approaches: statistical, distance or model based, profiling methods
- Generating counter-example to represent missing class at training

## Our Approach

- Extension for standard decision tree algorithms
- Able to deal with symbolic and continuous features
- **Instead of artificial counter-examples, use a parametric distribution for the anomaly**
  → Avoids trade-off between precision of sampling and the priors of the classes
  → More accurate split points
  → Faster training due to fewer samples

## Decision Tree

- Node divides feature space from its parent into two or more disjoint ranges
- Algorithm selects split according to an impurity measure of node t, e.g.

$$i(t) = -\sum_{c \in C} \frac{N_c(t)}{N(t)} \log_2 \left( \frac{N_c(t)}{N(t)} \right) \quad (1)$$

- Best split $s$ has highest decrease of impurity:

$$\Delta i(s,t) = i(t) - \frac{N(t_L)}{N(t)} i(t_L) - \frac{N(t_R)}{N(t)} i(t_R) \quad (2)$$

- $N_c(t)$ is number of instances of a class $c$ at node $t$
- No samples of anomaly class $c_A$ that can be count
  → Use density distribution to estimate $N_{c_A}$
- We use uniform distribution with a defined prior probability $P(c_A)$
  → anomaly distribution comparatively small in areas with many given samples but dominates regions without regular instances

## Uniform Distribution of Anomaly Class

**Symbolic features** → discrete uniform distribution

- Defined over a finite set $S$ of possible values, all equally probable: $\frac{1}{|S|}$
- probability that a feature has a value out of a set $M \subset S$: $P(X \in M) = \frac{|M|}{|S|}$

**Continuous features** → continuous uniform distribution

- constant probability density over a finite interval $[r^{min}, r^{max}]$:

$$f(x) = \begin{cases} \frac{1}{r^{max} - r^{min}} & x \in [r^{min}, r^{max}] \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

- probability that a data point is located inside a specific interval $[a,b] \subset [r^{min}, r^{max}]$:

$$P(X \in [a,b]) = \int_a^b f(x)dx = \frac{b-a}{r^{max} - r^{min}} \quad (4)$$

- $r^{min}$ and $r^{max}$ have to be defined before the training with $P(X \in [r^{min}, r^{max}]) = 1$

### Joint Distribution

- testing a split requires the number of instances which fall into the resulting subspaces
- a subspace $Q$ is defined by:
  - intervals $[a_i, b_i] \subset [r_i^{min}, r_i^{max}]$ for all $k_c$ continuous features
  - subsets $M_j \subset S_j$ of all $k_s$ symbolic features
- probability that a instance falls into $Q$ is the joint probability:

$$P(X \in Q) = \prod_{i=1}^{k_c} P(X_i \in [a_i, b_i]) \prod_{j=k_c+1}^{k_c+k_s} P(X_j \in M_j) \quad (5)$$

- expected number of instances within subspace $Q_t$ of node $t$:

$$N_{c_A}(t) = N_{c_A} P(X \in Q_t) = \frac{P(c_A)}{1 - P(c_A)} N_n P(X \in Q_t) \quad (6)$$

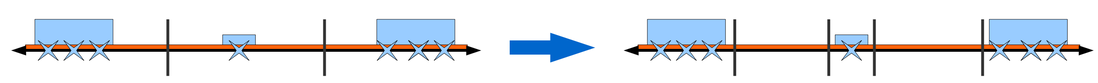
- $P(c_A)$ is the prior of the anomalous class and controls the trade-off between detection rate and false alarm rate

## Methodology

Use Equation 6 when number of instances of the anomaly class is needed
→ no major changes in procedure of finding the best split

### Suitable Split Points

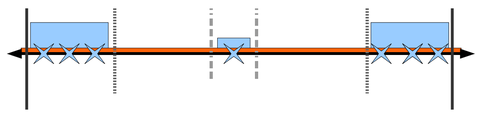
- No changes for symbolic features required
- But mean of two successive values for continuous features does not work since it would lead to splits between known classes only
  → cannot delimit regular classes from areas without training samples
- grid search over feature space too time intensive
  → test split points close to given samples



probable split points before and after redefinition of possible split points

### Stopping Criterion

- Typical way of creating the tree recursively until training error is zero not possible since classifying a region as "known" causes always an error
  → we have to define an error limit $\epsilon > 0$ w.r.t. $P(c_A)$ in relation to the dimensionality of the data
- $P(c_A)$ small, dimensionality rather high:
  → use smaller $\epsilon$ to force cuts closer to the samples and increase detection rate



Use smaller $\epsilon$ to cut off smaller regions

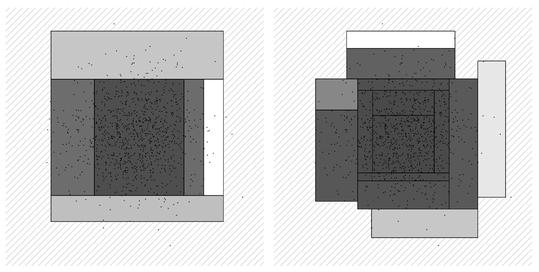
### Pruning

- most pruning techniques still applicable
- lower effect on methods dividing the training set because dividing the anomaly class makes no difference

## Experiments and Evaluation

### Synthetic Example

- two-dimensional space
- 1000 normally distributed data points
- illustrates trade-off between detection rate and false alarm rate

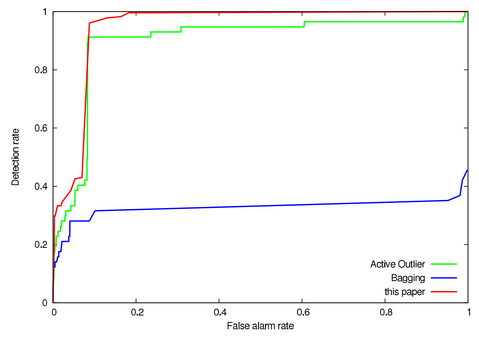


(a) lower prior $P(c_A)$ (b) higher prior $P(c_A)$

Different prior probabilities of anomaly class (filled: regular; hatched = anomaly; brightness indicates confidence)

### Real Life Data Sets

- three different Datasets from the UCI Machine Learning Repository
- training only on most common class
- at testing also one of the rare classes
- compared to other approaches [1][2][3]




ROC curves of KDD-Cup 99 data for Active Outlier, Bagging, and this paper

| Dataset | Regular Class | Anomaly Class | Active Outlier | Bagging | Feature Bagging | Boosting | LOF | This Paper |
|---|---|---|---|---|---|---|---|---|
| Ann-thyroid | 3 | 1 | 0.97 | 0.98 | 0.869 | 0.64 | 0.869 | **0.993** |
| Ann-thyroid | 3 | 2 | 0.89 | 0.96 | 0.769 | 0.54 | 0.761 | **0.977** |
| Shuttle (avg.) | 1 | 2,3,5,6,7 | **0.999** | 0.985 | 0.839 | 0.784 | 0.825 | 0.994 |
| KDD-Cup 99 | normal | U2R | 0.935 | 0.611 | 0.74 | 0.510 | 0.61 | **0.946** |

## References

[1] Naoki Abe, Bianca Zadrozny, and John Langford. Outlier detection by active learning. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, NY, USA, 2006. ACM Press.

[2] Edwin M. Knorr and Raymond T. Ng. Algorithms for mining distance-based outliers in large datasets. In *Proc. 24th Int. Conf. Very Large Data Bases, VLDB*, 1998.

[3] Aleksandar Lazarevic and Vipin Kumar. Feature bagging for outlier detection. In *KDD*. ACM, 2005.