A Transductive Forest for Anomaly Detection with Few Labels

https://github.com/jzha968/transForest

Jingrui Zhang, Ninh Pham, Gillian Dobbie

University of Auckland

ECML-PKDD, Sep 19, 2023

Anomaly detection

- Definition:
  - Identify rare items, events, or observations which deviate significantly from the majority of the data
- Unsupervised methods:
  - Common assumption: "*Anomalies are on sparse regions while normal points are on dense regions.*"
- Challenges in high dimensions:
  - The common unsupervised prior is not true due to irrelevant features
  - There are exponential number of subspaces to investigate

This work

- Research question: How could we find relevant subspaces to identify anomalies given few labels?

- Solution: Semi-supervised TransForest
  - Push the classification boundaries towards sensitive subspaces containing both normal and abnormal points
  - Provide feature importance ranking
  - Competitive with recent semi-supervised SOTA with **2%** labeled data

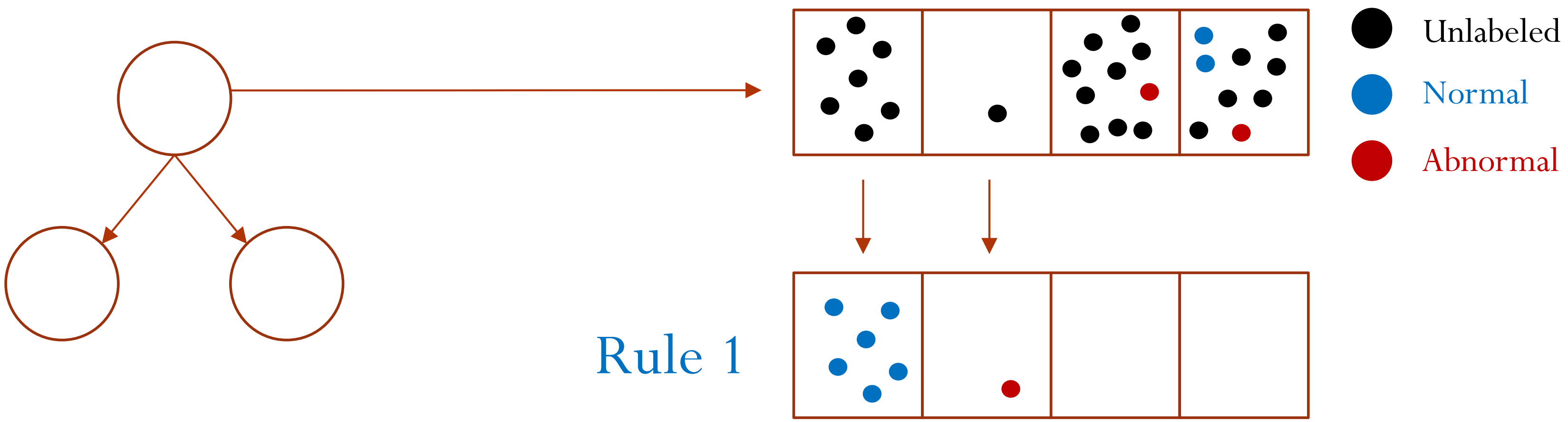
How TransForest works?

- Setting:
  - Collection of trees (similar to iForest, Random Forest, Extra Trees)
  - Tree's input: all available labeled points + a subset of unlabeled points

- Training: Pseudo-labeling + Extra Trees learning
  - For a randomly selected feature $\mathbf{f_i}$, build a histogram, and use label information to pseudo-label unlabeled points in each bin
  - For a random feature $\mathbf{f_i}$ and a random cut $\mathbf{v_i}$, compute the InfoGain based on the labeled and pseudo-labeled points; then, select $\mathbf{(f_i, v_i)}$ that maximizes InfoGain

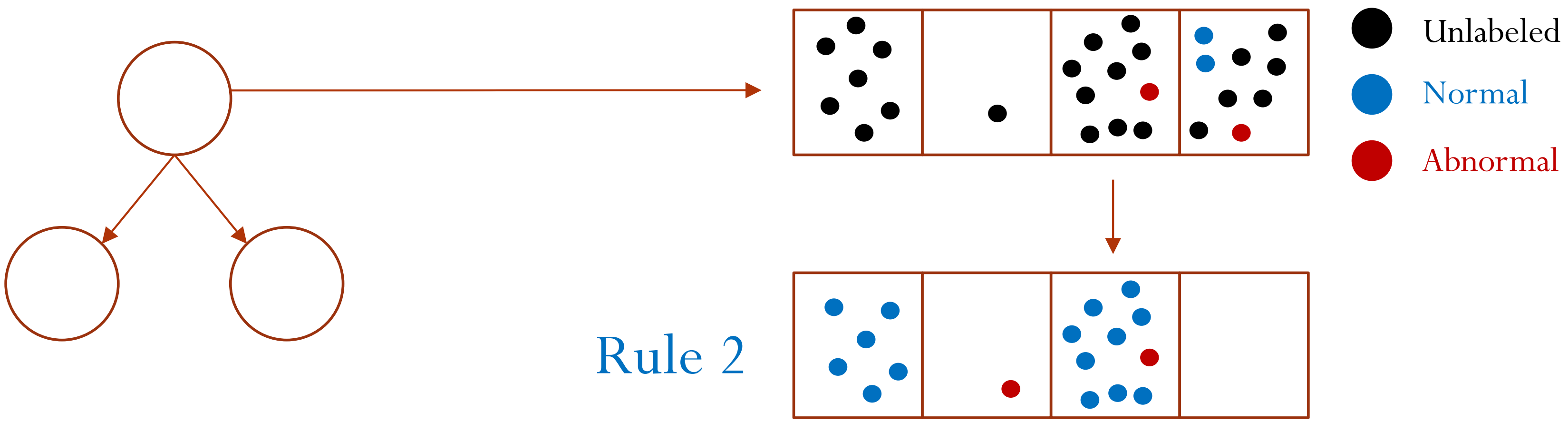
Pseudo-label points in a bin B

- Rule 1 (unsupervised):
  - If **B** has no labeled points, resort unsupervised prior given a threshold Δ
    - If **B** is dense ($|B| \geq \Delta$), all points in **B** are normal
    - If **B** is sparse ($|B| < \Delta$), all points in **B** are abnormal



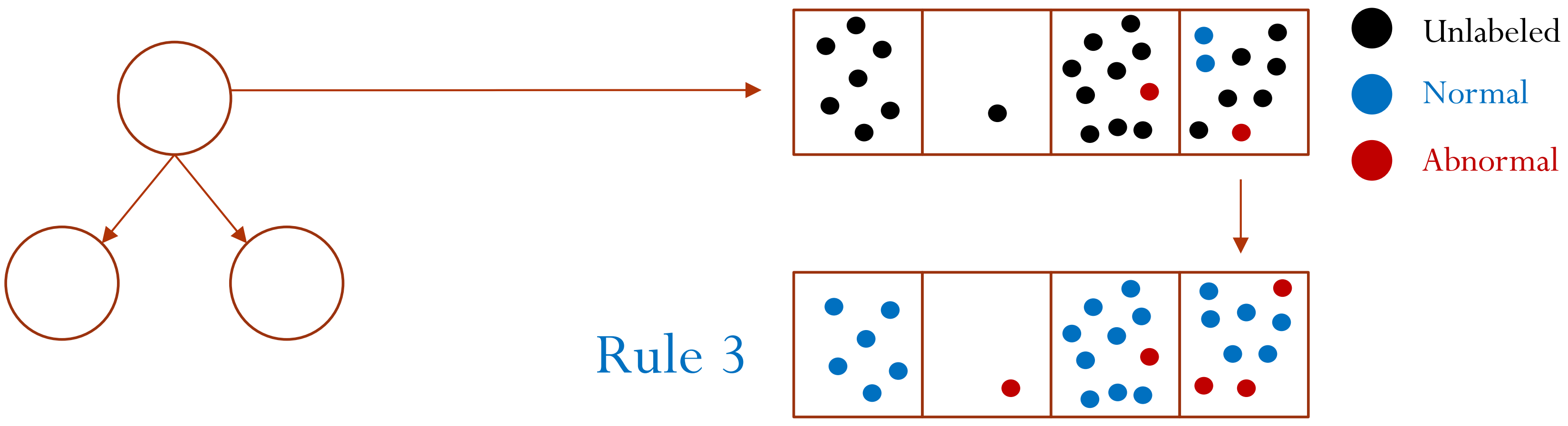
Pseudo-label points in a bin B

- Rule 2 (anomalies are rare):
  - If **B** is dense and contains only labeled anomalies
    - **B** will have **0.9|B|** normal and **0.1|B|** abnormal points



Pseudo-label points in a bin B

- Rule 3 (semi-supervised):
  - If **B** has $\mathbf{m_0}$ normal points and $\mathbf{m_1}$ anomalies
    - **B** will have $\frac{m_0}{m_0+m_1}|\mathbf{B}|$ normal and $\frac{m_1}{m_0+m_1}|\mathbf{B}|$ abnormal points



Pseudo-label points in a bin B

- Rule 1 (unsupervised):
  - If **B** has no labeled points, resort unsupervised prior given a threshold Δ
    - If **B** is dense $(|B| \geq \Delta)$, all points in **B** are normal
    - If **B** is sparse $(|B| < \Delta)$, all points in **B** are abnormal

hyperparameters

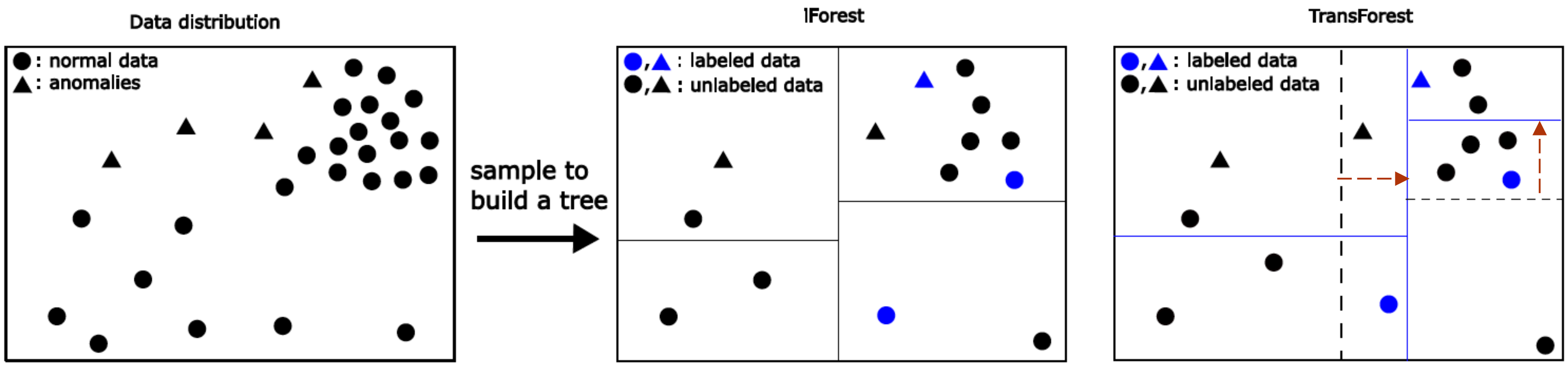
- Rule 2 (anomalies are rare):
  - If **B** is dense and contains only labeled anomalies
    - **B** will have $\mathbf{0.9|B|}$ normal and $\mathbf{0.1|B|}$ abnormal points

hyperparameters

- Rule 3 (semi-supervised):
  - If **B** has $\mathbf{m_0}$ normal points and $\mathbf{m_1}$ anomalies
    - **B** will have $\frac{m_0}{m_0+m_1}|\mathbf{B}|$ normal and $\frac{m_1}{m_0+m_1}|\mathbf{B}|$ abnormal points

TransForest vs iForest

- TransForest vs iForest:
  - TransForest pushes boundaries towards sensitive areas



- Testing:
  - Using pathLength as anomaly score (similar to iForest) with adjustment
    - o Labeled anomalies on isolated node: **pathLength = 1**
    - o Labeled normal points on isolated node: **pathLength = log(s)**

Hyperparameter setting

- Hyperparameters:
  - Density threshold of node **S**: $\Delta = \mathbf{0.1\,|S|}$
  - Number of histogram bins of node **S**: $\mathbf{\log(|S|) + 1}$
  - Number of trees: $\mathbf{t = 100}$
  - Number of trials to find the best split: $\mathbf{k = 10}$
  - Tree size: $\mathbf{s = max(256, 2\, n_l)}$ where $\mathbf{n_l}$ = # labels

- Complexity:
  - Training in $\mathbf{O(t\ s\ k\ \log(s))}$ and testing in $\mathbf{O(n\ t\ \log(s))}$ as similar as iForest

- Feature importance ranking:
  - Derived from the estimated InfoGain

Experiment

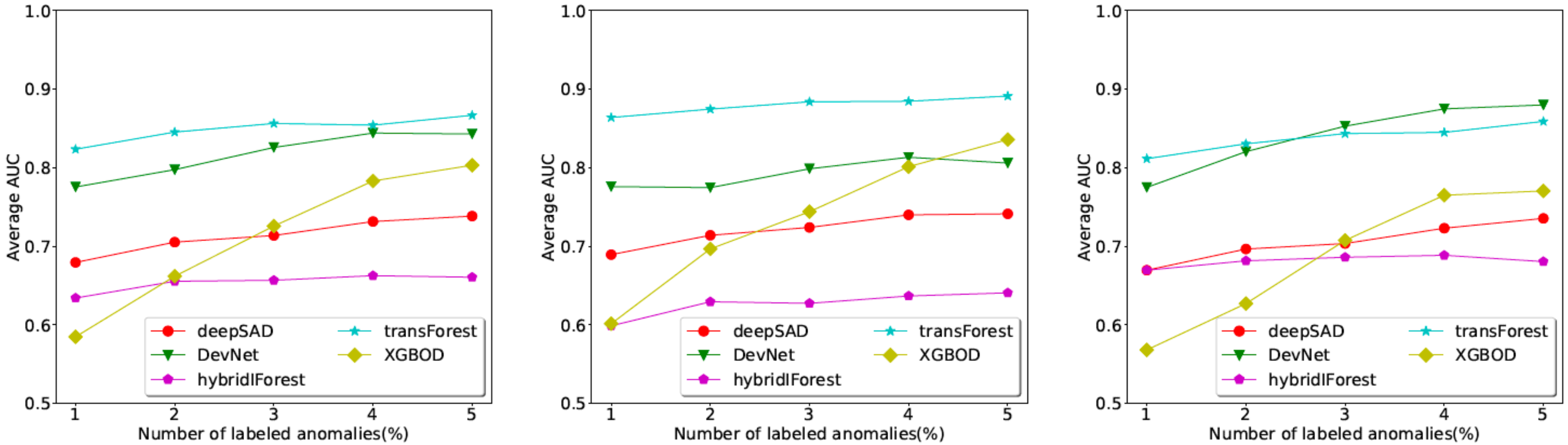
- Datasets:
  - **15** tabular and **20** continuous real-world data sets

- Experiment setting:
  - Randomly take **2% - 10%** labeled data for training (# labeled normal points = # labeled anomalies) and use the rest for testing
  - Measure average AUC for 10 runs

- Semi-supervised competitors:
  - Tree and boosting: Hybrid iForest, XGBOD
  - Deep learning: DeepSAD, DevNet

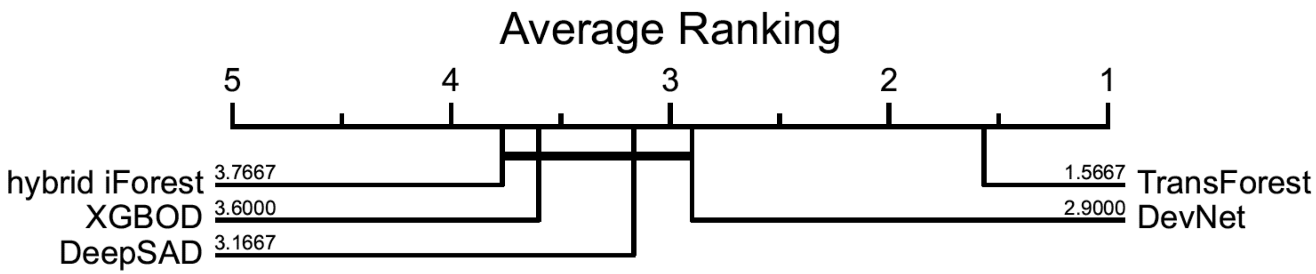
Experiment

- Avg AUC using **2% - 10%** labels on various data sets:



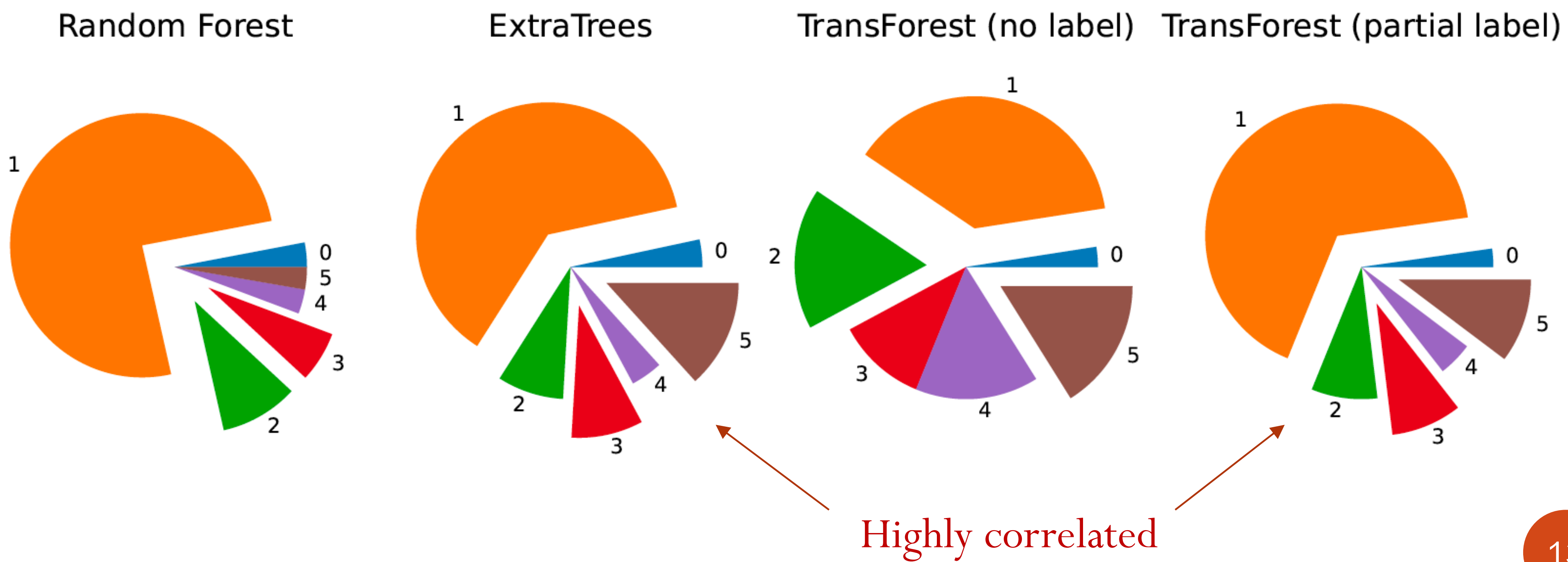
(a) All datasets. (b) Tabular datasets. (c) CV & NLP datasets.

- Critical diagram using **2%** labels on all data sets:



Feature importance ranking

- Annthyroid (**n = 7200, d = 6, # anomalies = 534**):
  - We use ~**2%** of labeled points (**10** anomalies and **10** normal points)
  - Number is dimension index
  - Wedge size reflects feature importance



Conclusion

- Few labels are useful for identifying relevant features for anomaly detection
- TransForest:
  - Simple but competitive with other semi-supervised models
  - Consistent importance feature ranking with supervised models on low-dimensional data sets
  - Robust against irrelevant features
  - https://github.com/jzha968/transForest