Unsupervised Outlier Detection on Thyroid Data

DSAA2011 Course Project

Member A: **Yuqi Ren**
Member B: **Anyi Wang**
Member C: **Feng Liang**

Hong Kong University of Science and Technology (Guangzhou)

Project Presentation

Project Overview

- Goal: detect thyroid-related anomalies using unsupervised methods.
- Dataset: **Annthyroid_unsupervised_anomaly_detection.csv**.
- **Team Roles:**
- **Member A (Yuqi Ren)**
    - Data acquisition & preprocessing
    - Baseline detector: Isolation Forest
    - Feature engineering & PCA visualization
- **Member B (Anyi Wang)**
    - Second detector: LOF
    - Evaluation: Precision@K, PR-AUC, ROC-AUC
    - Confusion matrix & second visualization
- **Member C (Feng Liang)**
    - Stability analysis on IF
    - Contextual Ensemble
    - Comparative experiment

Dataset Overview & Preprocessing

- **Dataset:** Annthyroid Unsupervised Outlier Detection
- **Size:** 6916 samples, 21 numerical features
- **Outlier ratio:** 3.6%
- Removed two empty columns; no missing values remain
- Features include:
  - Hormone measures (TSH, T3, TT4, T4U, FTI)
  - Clinical indicators (on_thyroxine, surgery, etc.)
- **Preprocessing:**
  - Log(1+x) transform to reduce skewness
  - StandardScaler for feature normalization

Baseline Isolation Forest

Initial model performance (before optimization):

- ROC-AUC: **0.5975**
- PR-AUC: **0.0532**
- Precision@50: **0.0400**

Observation: Baseline Isolation Forest performs poorly on this dataset due to high skewness and strong class imbalance.

Feature-Weighted Isolation Forest (FWIF)

Goal: Improve anomaly ranking by emphasizing important features.

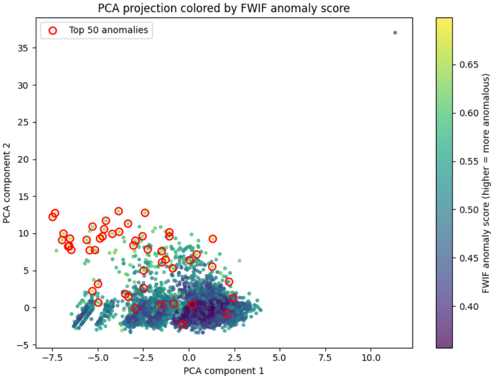
- Compute split-based feature importance from Isolation Forest.
- Select **top 12** influential features.
- Apply feature weighting:
  - High-importance features repeated ×3.
  - Medium-importance features repeated ×2.
  - Low-importance features kept as is.
- Retrain IF on the weighted + log-transformed + standardized matrix.

Final FWIF performance:

- ROC-AUC: **0.7330**
- PR-AUC: **0.1835**
- Precision@50: **0.4600**

Improvement: Precision@50 improved from 4% (baseline) to 46% ($\approx$ 11.5× increase).

PCA Visualization of FWIF Anomaly Scores



Key Insights:

- Normal points form a dense central cluster.
- High anomaly scores appear in low-density peripheral regions.
- Top-50 anomalies (red circles) lie on cluster boundaries.
- A few isolated points indicate extreme abnormalities.

Stability Analysis of FWIF(Bootstrap + Seeds)

1. **Bootstrap Resampling (10 runs)**
   - Mean ROC-AUC: 0.7206 std: 0.0052
   - Mean PR-AUC: 0.1645 std: 0.0113
   - Precision@50: 0.3920 std: 0.0646
2. **Random Seed Test (10 seeds)**
   - ROC-AUC std: **0.0060**
   - PR-AUC std: **0.0097**
   - Precision@50 std: **0.0367**

Conclusion:

- ROC-AUC and PR-AUC remain highly stable.
- Precision@50 shows acceptable variation due to top-50 sensitivity.
- FWIF is robust across sampling variations and random seeds.

Second Detector: Local Outlier Factor (LOF)

Motivation

- Try a neighborhood-density detector as a contrast to Isolation Forest.

Settings

- Features: same **top-12** as FWIF; **log(1+x)** + **RobustScaler**.
- Grid: $k \in \{5, 10, 15, 20, 30, 50\}$, contamination $\in \{0.01, 0.02, 0.03\}$.
- Scoring: negative_outlier_factor inverted; higher = more anomalous.

Best config (by PR-AUC)

- contamination = $\mathbf{0.01}$, $\quad k = \mathbf{5}$.

LOF Performance (Best Config)

Offline evaluation (labels only for assessment)

- ROC-AUC: **0.5286**
- PR-AUC: **0.0703**
- Precision@50: **0.1800**

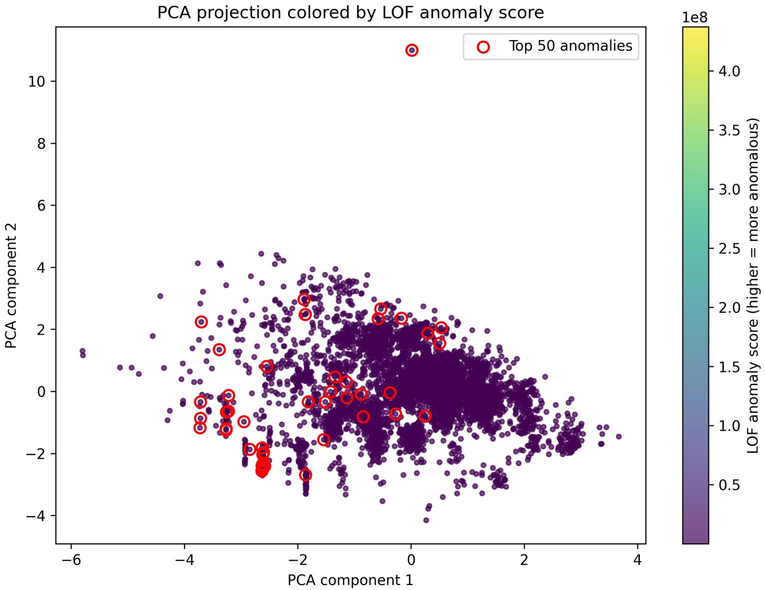
Threshold & confusion matrix

- Threshold: 99% quantile (aligned to 1% contamination)
- Confusion: TN=6614, FP=52, FN=232, TP=18

Observations

- LOF underperforms IF on PR-AUC and Precision@50.
- Low-density areas yield unstable scores; sensitive to duplicates/local scale.

PCA Visualization of LOF Anomaly Scores



IF vs LOF: Comparison & Takeaways (Before Ensemble)

Metrics

- IF (FWIF): ROC-AUC 0.7330, PR-AUC 0.1835, Precision@50 0.4600
- LOF (best): ROC-AUC 0.5286, PR-AUC 0.0703, Precision@50 0.1800

Visualization

- IF: Top-K along cluster boundary; smooth score gradient.
- LOF: Top-K dispersed; extreme scores mixed into the main cluster.

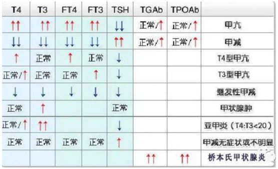
Takeaways

- Keep LOF as a reported attempt/baseline; choose IF as the primary detector.
- Reason: LOF is sensitive to duplicates/local scale and generalizes poorly here.

Contextual Ensemble: Motivation and Design

Motivation

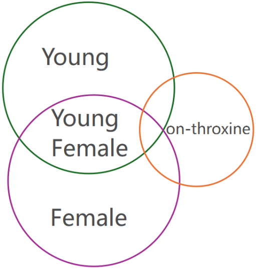
- Hormone measures (TSH, T3, TT4, T4U, FTI) vary across demographic & clinical subgroups.
- Global anomaly detection may miss context-specific deviations.



T4	T3	FT4	FT3	TSH	TGAb	TPOAb	
↑↑	↑↑	↑↑	↑↑	↓↓	正常/↑	正常/↑	甲亢
↓↓	↓↓	↓↓	↓↓	↑↑	正常/↑	正常/↑	甲减
↑	正常	↑	正常	↓			T4型甲亢
正常/↑	正常	正常	↑	↓			T3型甲亢
↓	↓	↓	↓	↓			继发性甲减
正常	↑			正常			甲状腺肿
正常/↑	↑↑			↓			亚甲炎（T4:T3<20）
正常	正常	正常	正常	↑			甲减无症状或不明显
					↑↑	↑↑	桥本氏甲状腺炎

Context Construction

- Medical-informed subgroups:
  - Age × Gender
  - Treatment status (no/thyroxine/antithyroid/post-surgery)
  - Special status (pregnant, lithium medication)



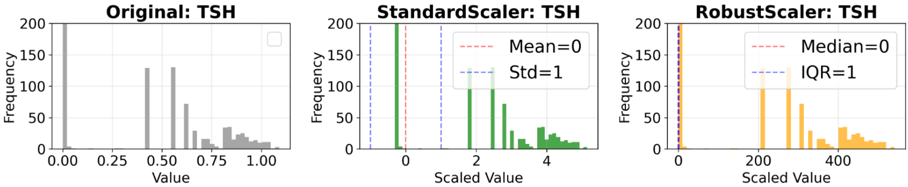
Pipeline

- Train separate LOF/IF within each context.
- Aggregate scores across contexts via averaging.

Contextual Ensemble: Main Findings

Configuration	**ROC-AUC**	**PR-AUC**	**Precision@50**
Baseline LOF	0.5286	0.0703	0.1800
LOF + SS + c. + discrete avg	0.6719	0.1387	0.1200
LOF + RS + c. + continuous avg	**0.8812**	0.2015	0.3200
LOF + RS + c. + discrete avg	**0.7604**	**0.2662**	**0.7600**

Table: SS: StandardScaler, RS: RobustScaler, c.: with context ensemble

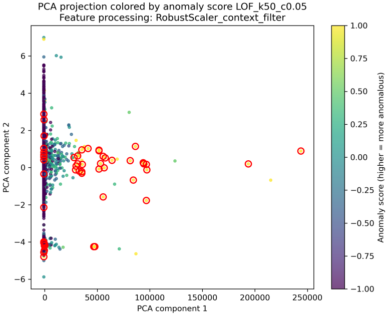


Key Insights

- **RobustScaler** handles outlier hormone data better
- **Discrete averaging** yields 4× higher Precision@50 (0.76 vs 0.18)

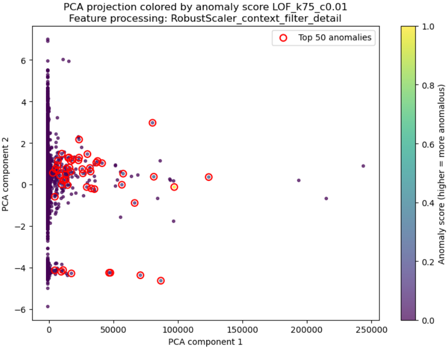
Discrete vs Continuous Aggregation: A Visual Comparison

Discrete Voting Aggregation



- **Method**: Context output ±1, sum votes
- **Result**: Amplify outliers
- **Precision@50**: **0.76**

Continuous Score Aggregation



- **Method**: Average normalized scores
- **Result**: cautious predictions
- **Precision@50**: 0.32

Summary & Future Directions

Model	ROC	PR	P@50
Baselines			
IF (initial)	0.60	0.05	0.04
LOF (initial)	-	-	-
Optimized Single			
FWIF	0.73	0.18	0.46
LOF (best)	0.53	0.07	0.18
Ctx. Ensemble			
FWIF + ctx	0.74	0.21	0.46
LOF + RS + disc	0.76	**0.27**	**0.76**

Future Work

- **Hybrid ensemble**: Combine FWIF + contextual-LOF scores
- **Random subspaces**: Ensemble on feature subsets
- **Deep learning**: Autoencoders for reconstruction-based detection

Q&A

Thank You!