

RESEARCH PRESENTATION

Cluster-Guided Minority Enhancement for Imbalanced Classification

A simple way to make SMOTE more reliable

Xiaorui Huang · Pyeoungkee Kim

Why is imbalanced classification difficult?

Too few positive examples

- Models tend to favor the majority class.
- Rare but important cases are missed.
- Accuracy can hide this problem.

Example: stroke data

42,617

negative cases

783

positive cases

Why is standard SMOTE not enough?

SMOTE creates new minority samples between neighboring minority points.

Problem 1 | Outliers

An isolated or noisy point can generate misleading synthetic data.

Problem 2 | Different regions

Interpolation may connect unrelated minority groups.

Main claim

Better minority samples matter more than simply adding more samples.

Our hypothesis

Filter atypical minority groups, then generate samples only within reliable local groups.

OA-KM-SMOTE: a two-stage clustering approach

- 1 Split the data** Keep the test set untouched.
- 2 First K-means** Find sparse minority clusters.
- 3 Filter outliers** Exclude these groups from oversampling.
- 4 Second K-means** Group the remaining minority points.
- 5 Local SMOTE** Generate samples within each group.

Why two K-means stages?

Stage 1 · Filter

Cluster minority samples.
Identify unusually small clusters.
Exclude them from sample generation.

Stage 2 · Generate

Recluster retained minority samples.
Find neighbors inside each group.
Apply SMOTE locally.

First remove atypical seeds; then avoid mixing different minority regions.

How local SMOTE improves sample quality

Standard SMOTE

Uses nearby minority points without checking cluster structure.

Outliers or separate groups may lead to poor synthetic samples.

OA-KM-SMOTE

Removes sparse outlier groups first.

Interpolates only within retained local clusters.

Better preserves local structure.

Key idea: control WHERE new samples are generated, not just HOW MANY.

How did we evaluate the method?

3 datasets

Stroke

11 features · 15 detected outliers

Credit-card default

8 features · 60 detected outliers

Car-insurance claims

43 features · 80 detected outliers

Same split, same classifiers

75% train / 25% test · Recall and F1 score

4 classifiers

- XGBoost
- Random forest
- K-nearest neighbors
- Decision tree

Does it improve performance?

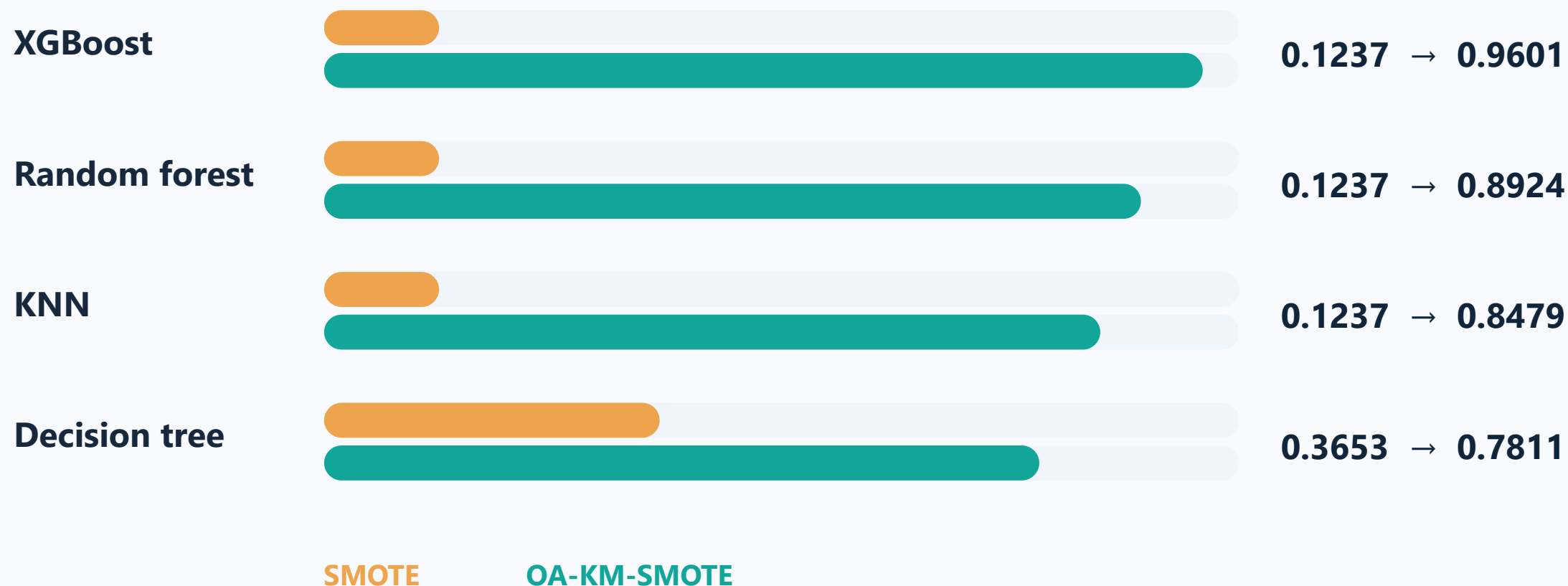
Average F1 improvement across 4 classifiers



Recall and F1 improved in all 12 dataset–classifier comparisons.

The largest gain: car-insurance claims

Reported F1 score: standard SMOTE vs. OA-KM-SMOTE



What can we conclude?

Filtering first + local sampling = more reliable minority examples

- The same resampling pipeline works with different classifiers.
- The reported gains are strongest for difficult, high-dimensional data.
- A risk: real but rare minority groups may be removed.

Take-home message: Where we generate minority samples matters.

RESEARCH PRESENTATION

THANK YOU
