THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Position-Aware Safe Boundary Interpolation Oversampling

Yongxu Liu, Yan Liu

The Hong Kong Polytechnic University, Hong Kong SAR, China

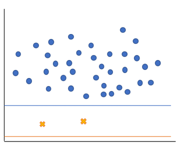
18041824r@connect.polyu.hk, csyliu@comp.polyu.edu.hk

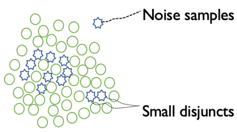
Definition

- Imbalanced data: unequal distribution of different class samples [1, 2].
- Interpolation-based oversampling: the synthetic samples are interpolated along the line segment between the reference and candidate points.

Problem

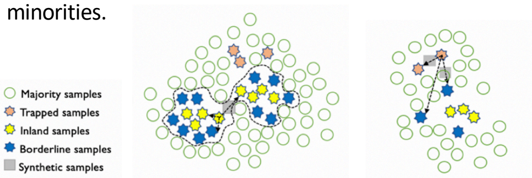
- Imbalance data usually compromise the performance of standard classifiers.
- Not only imbalance ratio hinder classifier, but also noise and small disjuncts hinder classifier.





Generating synthetic samples to create balanced datasets has two challenges:

- **Over-constraint**: generate overlapped synthetic samples for the inland because of improper clustering.
- **Over-generation** of erroneous samples [3]: generate synthetic samples for the trapped based on the nearest minorities.



(a) Over-constraint (b) Over-generation

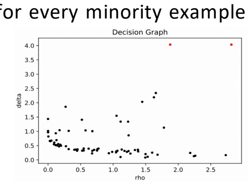
Approach

Clustering

- Adopt CFSFDP clustering algorithm to separate two dense clusters [5].
  - Centers are with high local density (**high ρ**) and large relative distance (**high δ**) to points with higher density

Given a distance matrix $[d_{ij}]_{n*n}$, for every minority example x_i compute:

- $\rho_i = \sum_{j:j\neq i} e^{-(\frac{d_{ij}}{d_c})^2}$
- $\delta_i = min_{j:\rho_j>\rho_i}(d_{ij})$



Generation

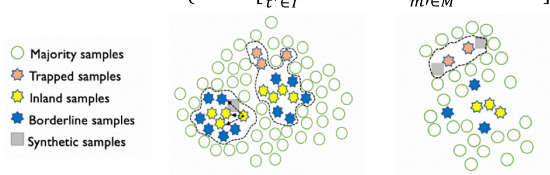
Use interpolation-based method for generating synthetic inland and borderline example.

- **Inland**: the candidate set is the same cluster $L_c \backslash x_i$, where $x_i \in L_c$.
- **Borderline**: the candidate set is k_{maj} nearest majority neighbors $N_{maj}(x_i)$.

Propose a novel approach of generating safe boundary for generating synthetic trapped example.

- **Trapped**: for any trapped example t $\in T$, the candidate set is T and nearest majority neighbors set M.

$$l = \max\left\{0, 2 - \left[\max_{t'\in T} \phi(t', s) - \max_{m'\in M} \phi(m', s)\right]\right\}$$



(a) CFSFDP clustering (b) Proposed safe boundary

Experimental Evaluation

- **Datasets**: Five classical Imbalanced data sets from UCI repository

Data	# Min class	# Maj class	# Min Instances	# Maj Instances	# F_num	IR
Pima	1	1	268	500	8	1.866
Ecoli	5	3	64	272	7	4.25
Vowel	1	1	90	900	8	10
Yeast	2	8	81	1403	8	17.32
AB1	2	26	99	4078	7	41.19

- **Baselines**: Nine other oversampling algorithms
- **Classifiers**: Linear-SVM and C4.5 decision tree
- **Metrics**: F1-score, G-mean, AUC [4]

Conclusion

F1-SCORE, G-MEAN, AND AUC OF ALL THE OVERSAMPLING METHODS ON EACH NUMERICAL IMBALANCED DATASET USING LINEAR-SVM.

Data	Metrics	None	ROS	SMOTE	Safe-SMOTE	MWMO	SMOM	INOS	MDO	RACOG	PAIO	PABIO
Pima	F1-score	0.6253	0.6188	0.6638	0.6609	0.6547	0.6639	0.6527	0.641	0.5339	0.6596	**0.6667**
	G-mean	0.6996	0.7002	**0.7384**	0.7359	0.731	0.7383	0.7282	0.7193	0.6248	0.7348	0.7070
	AUC	**0.8294**	0.7676	0.8274	0.8265	0.8202	0.8275	0.8239	0.8264	0.7304	0.8241	0.7500
Ecoli	F1-score	0.6935	0.7388	0.751	0.7478	0.7262	0.7526	0.7427	0.758	0.5998	0.7434	**0.7710**
	G-mean	0.7724	0.8849	0.8866	0.8778	0.8698	0.8857	0.8811	0.8807	0.7463	0.8855	**0.9000**
	AUC	0.9392	0.9372	0.9387	0.938	0.9314	0.9392	0.9389	0.9391	0.7929	**0.9405**	0.9000
Vowel	F1-score	0.3106	0.5071	0.5071	0.5066	0.5031	0.5058	0.509	0.4934	0.4937	0.5061	**0.5385**
	G-mean	0.1805	0.8765	0.8701	0.8646	0.8614	0.8677	0.8677	0.8529	0.8151	0.873	**0.9090**
	AUC	0.8934	0.9151	0.9127	0.913	0.9116	0.913	0.9124	0.9092	0.8942	0.915	**0.9409**
Yeast	F1-score	0.2532	0.3282	0.3258	0.3801	0.3183	0.3219	0.3439	0.4334	0.3328	0.3337	**0.4715**
	G-mean	0.3202	0.8077	0.8029	0.7986	0.8016	0.8002	0.7991	0.75	0.7721	0.8132	**0.8320**
	AUC	0.7364	0.856	0.8583	0.8597	0.856	0.86	0.8525	0.8588	0.8216	0.8569	**0.8920**
Abalone	F1-score	NaN	0.1608	0.1614	0.2021	0.1599	0.1643	0.1977	0.1778	0.0855	0.1667	**0.2286**
	G-mean	0	0.7689	0.766	0.7194	0.7546	0.7607	0.7494	0.712	0.6625	0.7719	**0.8035**
	AUC	0.6635	0.8764	0.8782	0.8592	0.8645	0.8758	0.8796	0.8701	0.6974	0.8803	**0.9006**

- **F1-score**: our PABIO achieves the best results of all five data sets.
- **G-mean:** our PABIO outperforms most of the five data sets.
- **Robustness**:
  - Vowel dataset (no trapped example), PABIO discovers more dense minority groups, and generates synthetic inland samples safely.
  - Abalone dataset (only has trapped examples), PABIO learns safe boundary and expands minority efficiently.

Hyperparameters

- Adopt the recommended values of the common parameters in PAIO
- If the hyperparameter d_c of clustering falls in appropriate value range, it would affect the performance of our proposed PABIO.

Reference

1. Guan, Hongjiao, et al. "WENN for individualized cleaning in imbalanced data." *ICPR*, 2016.
2. Ghanem, Amal S., Svetha Venkatesh, and Geoff West. "Multi-class pattern classification in imbalanced data." *ICPR*, 2010.
3. Sandhan, Tushar, and Jin Young Choi. "Handling imbalanced datasets by partially guided hybrid sampling for pattern recognition." *ICPR*, 2014.
4. García, Vicente, Ramon A. Mollineda, and J. Salvador Sanchez. "Theoretical analysis of a performance measure for imbalanced data." *ICPR*, 2010.
5. Rodriguez, Alex, and Alessandro Laio. "Clustering by fast search and find of density peaks." *Science, 2014.*