Position-Aware Safe Boundary Interpolation Oversampling

Yongxu Liu, Yan Liu

Department of Computing

The Hong Kong Polytechnic University, Hong Kong SAR, China

Outline

- Background
- Literature
- Proposed Method
- Future Work

Outline

- **Background**
- **Literature**
- **Proposed Method**
- **Future Work**

Imbalance Data Classification

- Classification is a supervised learning task.
  - **Learn** from training data, then **predict** categorical classes on test data.
- **The imbalanced data sets**
  - The number of data in some classes are **extremely smaller** than other classes.
  - Widely existing in many applications, such as:
    - fraud detection, disease diagnosis, oil spill detection from satellite images, etc.
- Imbalanced data **significantly compromise** the performance of most standard learning algorithms.

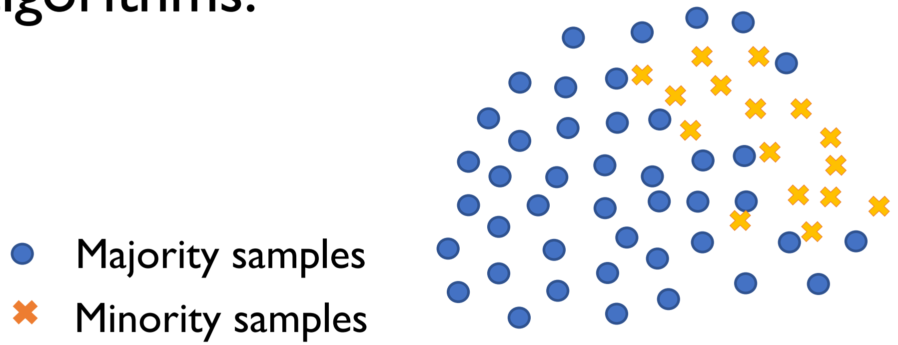


Fig. Imbalanced data

Effects of Imbalance on SVM

- SVM works well on balanced datasets. But it towards the majority class and has low performance on the minority on imbalanced datasets.
- SVM has two objectives:
  - separating the two classes with the maximum margin
  - minimizing the number of misclassifications

- **Effects**: Margin is maximized with low total misclassification error

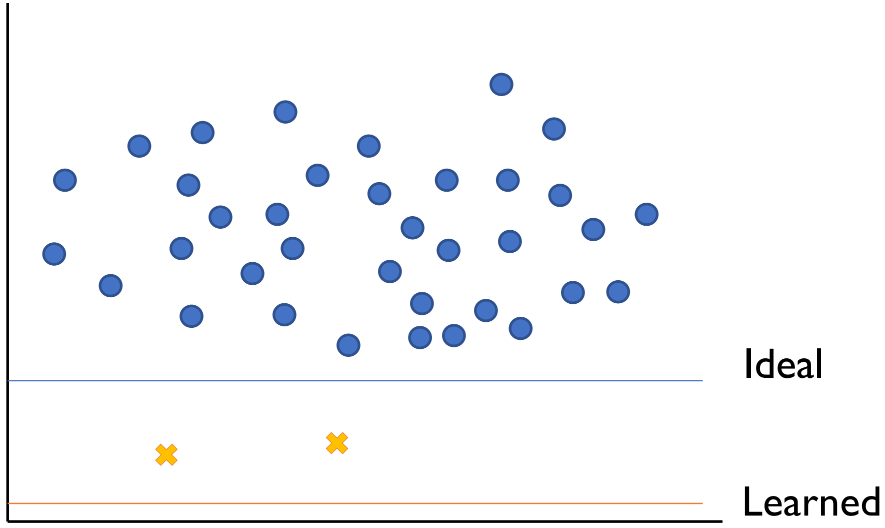


Fig. Effects of Imbalance on SVM

Outline

- Background
- Literature
- Proposed Method
- Future Work

Literature

- The existing solutions of class imbalance classification can be roughly divided into two types:
  - Algorithm-level methods: Modify learners, such as adding misclassification cost or modified loss function
    - Adaboost[1], inducing the misclassification cost of the minority class to build an ensemble of weak learners
  - Data-level methods: Modify data distribution, create balanced dataset
    - Bootstrap[2], sampling dataset with replacement in each iteration
    - Advantages:
      - more **universal** as they do not rely on any specific learner
      - more **flexible** combined with other techniques in machine learning.

1. P. Thanathamathee and C. Lursinsap, "Handling imbalanced data sets with synthetic boundary data generation using bootstrap re-sampling and adaboost techniques," PRL, 2013.
2. J. Han, J. Pei, and M. Kamber, Data mining: concepts and techniques. Elsevier, 2011.

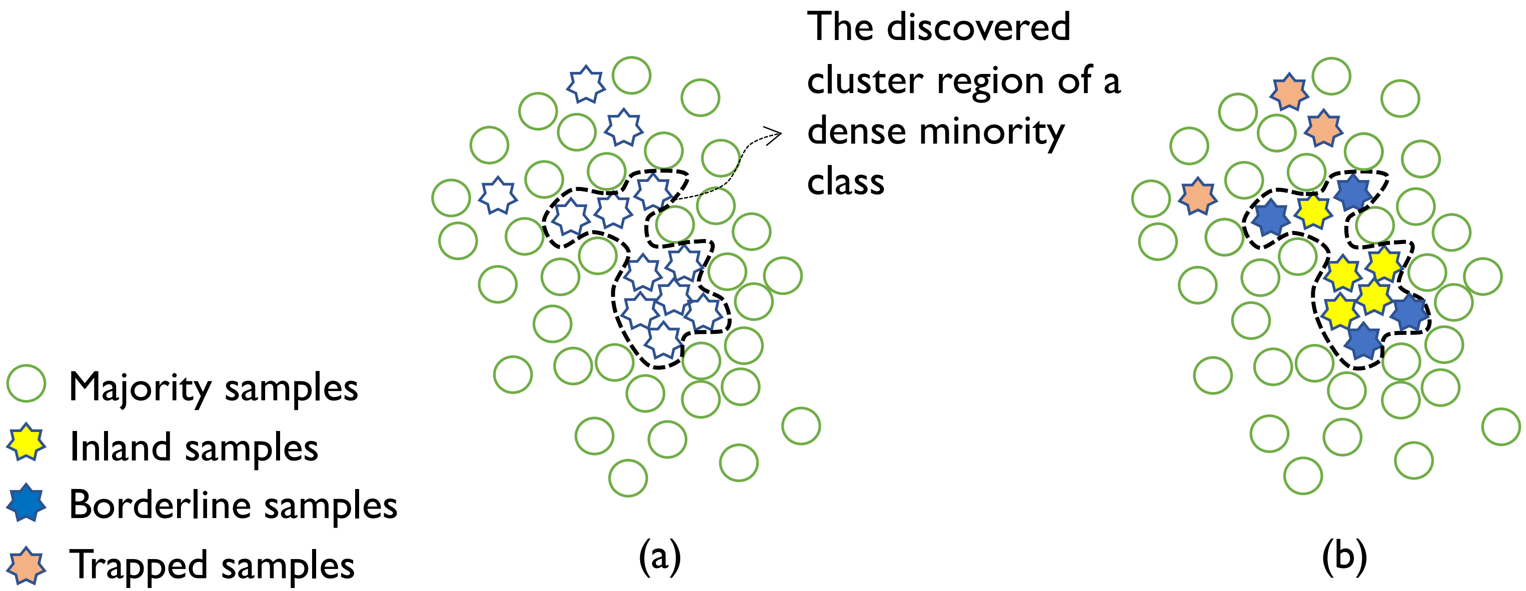
Literature: Data-level

- **Under-sampling**: sample part of majority data
  - **Issues**: probably will lose useful information by discarding the majority instances.
- **Over-sampling**: generate more minority data
  - **Issues**: more frequently used in data-level methods, as do not discard examples and would not lose useful information for classifier.
- **Cluster-based sampling**: integrate clustering algorithms with over-sampling.
  - Not only classify **between classes**, also handle samples **within classes**.
  - E.g. Cluster-SMOTE[1] (K-means), DBSMOTE[2] (DBSCAN)
  - **Issues**: introducing additional noise because of improper and non-robust clustering methods.

1. Nekooeimehr and S. K. Lai-Yuen, "Adaptive semi-unsupervised weighted oversampling (a-suwo) for imbalanced datasets," Expert Systems with Applications, 2016.
2. D. A. Cieslak, N. V. Chawla, and A. Striegel, "Combating imbalance in network intrusion datasets." in GrC, 2006, pp. 732–737.

Preliminary

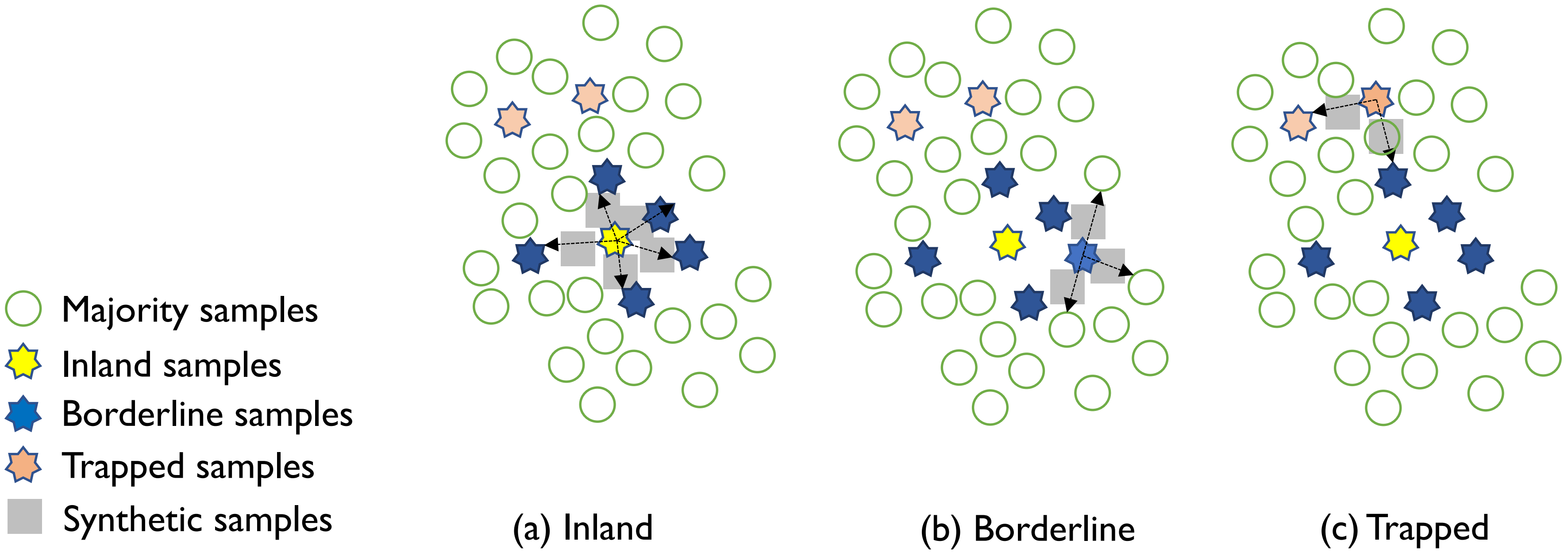
- Recently, a position characteristic-aware interpolation over-sampling algorithm (PAIO) has been proposed to re-balance data sets.
- There are two main phases of this work:
  - 1. **Cluster** the minority examples and identify them into inland, borderline, trapped points.



Reference: T. Zhu, Y. Lin, and Y. Liu, "Improving interpolation-based oversampling for imbalanced data learning," Knowledge-Based Systems, 2020.

Preliminary

- Recently, a position characteristic-aware interpolation over-sampling algorithm (PAIO) has been proposed to re-balance data sets.
- There are two main phases of this work:
  - 2. **Generate** synthetic examples accordingly.



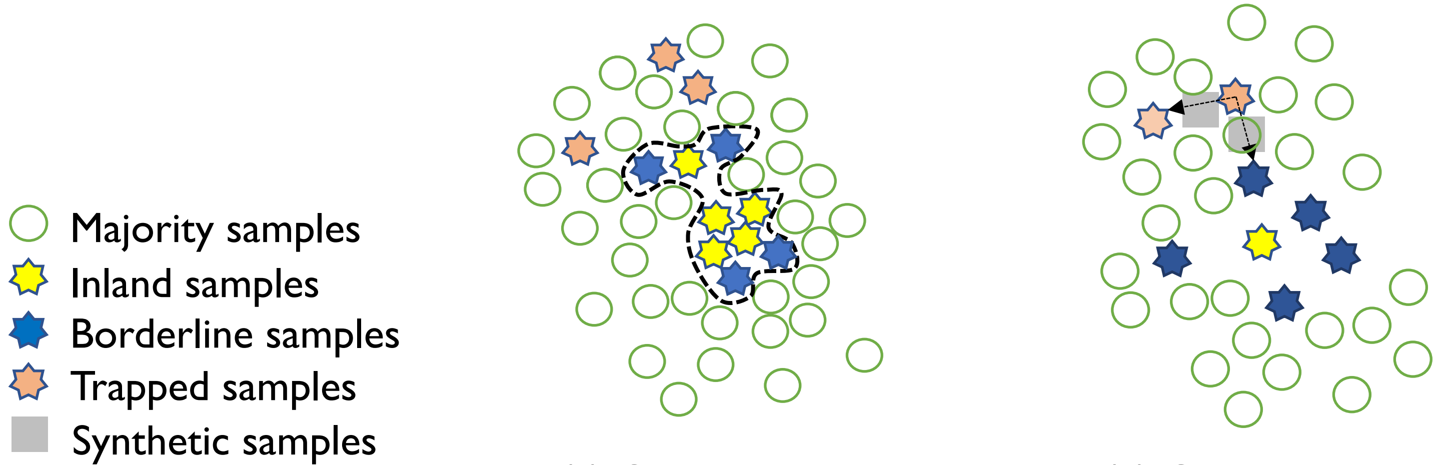
Reference: T. Zhu, Y. Lin, and Y. Liu, "Improving interpolation-based oversampling for imbalanced data learning," Knowledge-Based Systems, 2020.

Outline

- **Background**
- **Literature**
- **Proposed Method**
  - Motivation
  - Formulation
  - Experiment
- **Future Work**

Motivation

- **Issues of PAIO**
  - **Clustering Issue:** PAIO tends to group two dense minority samples into one cluster.
    - Leads the synthetic samples locate in majority sample.
  - **Generation Issue:** PAIO tries to generate synthetic points for trapped samples according to k-nearest neighbors
    - Easily causes the points close to majority samples.



(a) Clustering Issue (b) Generation Issue

Reference: T. Zhu, Y. Lin, and Y. Liu, "Improving interpolation-based oversampling for imbalanced data learning," Knowledge-Based Systems, 2020.

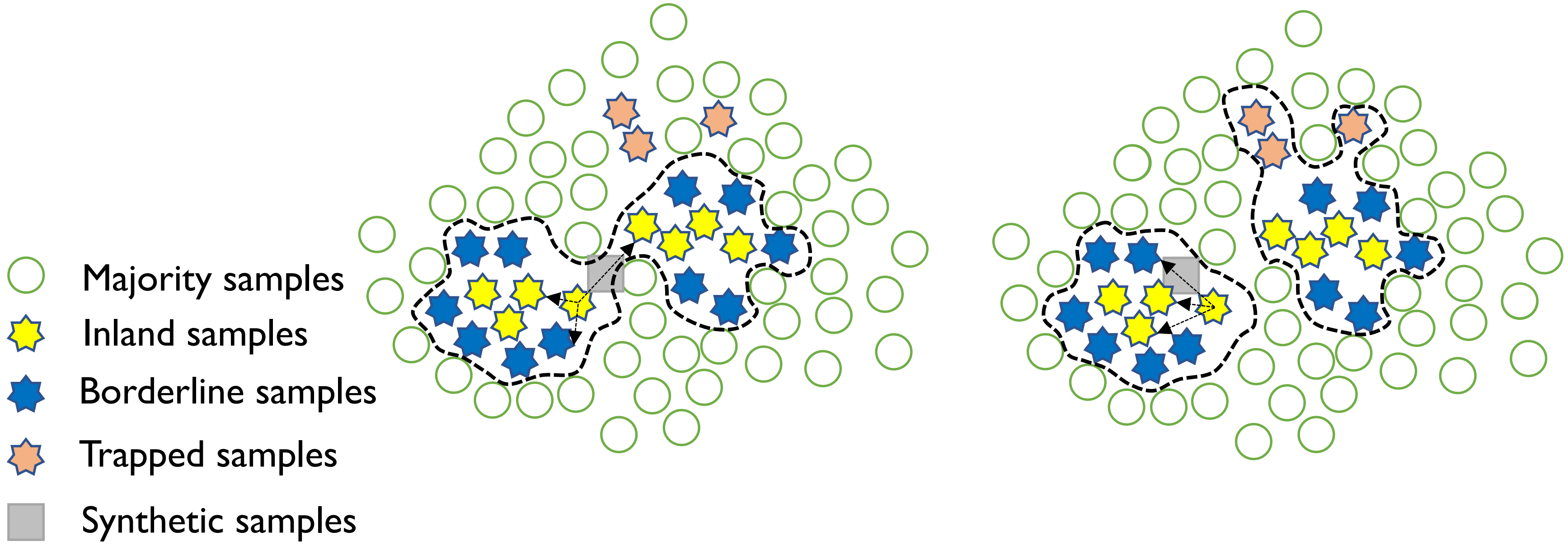
Outline

- **Background**
- **Literature**
- **Proposed Method**
  - Motivation
  - Formulation
  - Experiment
- **Future Work**

Proposed Method

Clustering

- We employ one advanced clustering algorithm, CFSFDP, which is able to handle such scenario.



(a) Clustering adopted by PAIO[1]   (b) CFSFDP[2] clustering

1. T. Zhu, Y. Lin, and Y. Liu, "Improving interpolation-based oversampling for imbalanced data learning," Knowledge-Based Systems, 2020.
2. A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, 2014.

Proposed Method

Clustering

- Given a distance matrix $D = [d_{ij}]_{n*n}$, where d_{ij} denotes the distance between the minority samples x_i and x_j.
- For each minority sample x_i compute:
  - $\rho_i = \sum_{j:j\neq i} e^{-(\frac{d_{ij}}{d_c})^2}$ (local density of minority points within a distance d_c)
  - $\delta_i = min_{j:\rho_j>\rho_i}(d_{ij})$ (distance to the closest minority point with higher density)

- CFSFDP assumes that cluster centers are defined by **a high local density** $\boldsymbol{\rho}$ within **a relatively distance** between centers.

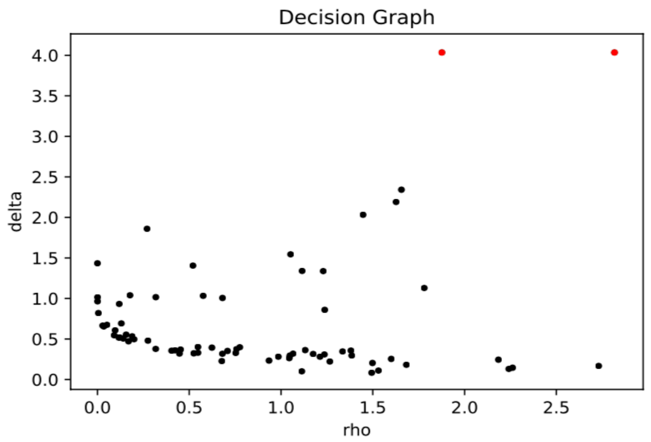


Fig. Decision Graph the imbalance dataset:Vowel dataset

Proposed Method

Division

- Given a set of minority examples $X = \{x_1, \cdots, x_n\}$ and a set of clusters from the clustering $L = \{L_1, \cdots, L_{n_L}\}$, where $L_c \subset X, L_i \cap L_j = \emptyset$ for any $i \neq j$.
- For each minority example x_i compute its local density within its m-nearest neighbors $N_m(x_i)$:
  - $\kappa(x_i) = |N_m^*(x_i)|/m,$
  - where $N_m^*(x_i) = \{x_j | x_j \in N_m(x_i) \cap L_c, x_i \in L_c, i \neq j\}$

Proposed Method

Division

- After clustering the minority examples, classify them into **inland, borderline, trapped examples.**
    - Inland: $\kappa(x_i) > \rho TH$
    - Borderline: $\kappa(x_i) \leq \rho TH$ and an inland sample exist in its $N_m(x_i)$
    - Trapped: $\kappa(x_i) \leq \rho TH$ and not a single inland sample exist in its $N_m(x_i)$

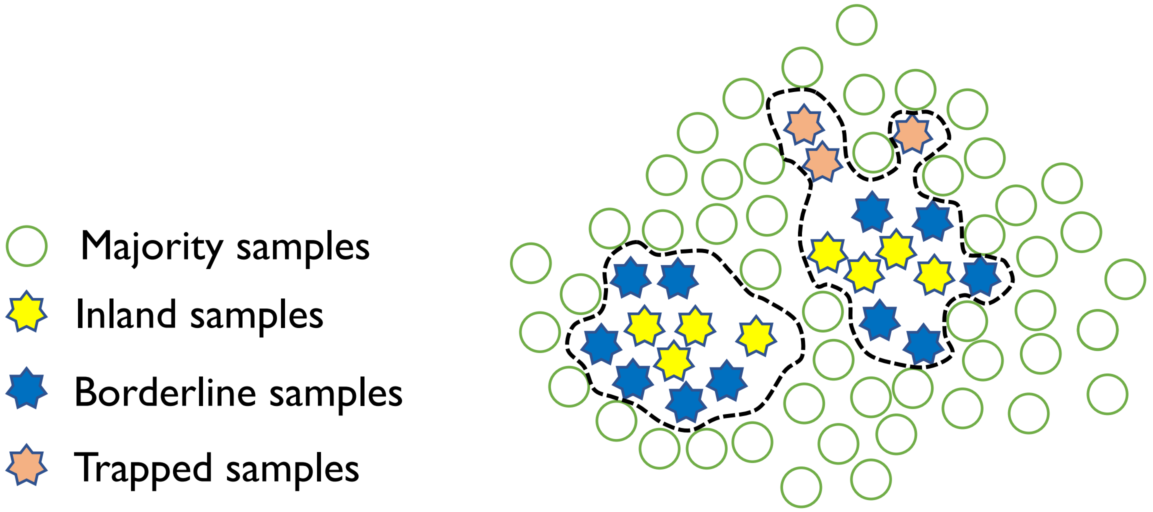


Fig. Our proposed classifying minority examples based on clustering

Proposed Method

Generation

- After grouping the minority examples, generate synthetic points for inland, borderline and trapped samples, respectively.
  - We follow the interpolation-based method for **inland** and **borderline**, the same with PAIO.
    - Given a point x_i and its candidate point x_j, the synthetic point s is calculated as:

$$s = x_i + \gamma\left(x_j - x_i\right), \text{ where } \gamma \text{ is a constant vector.}$$

  - We propose a new method for **tapped points** to reduce noise.

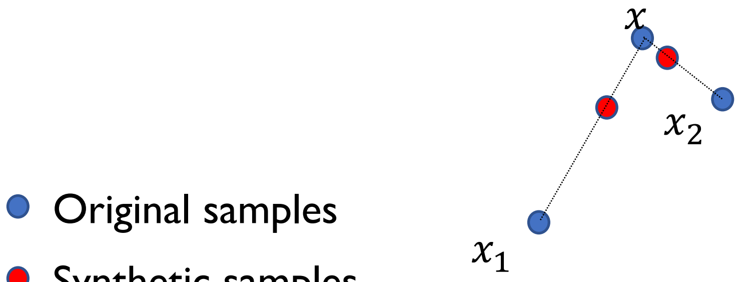


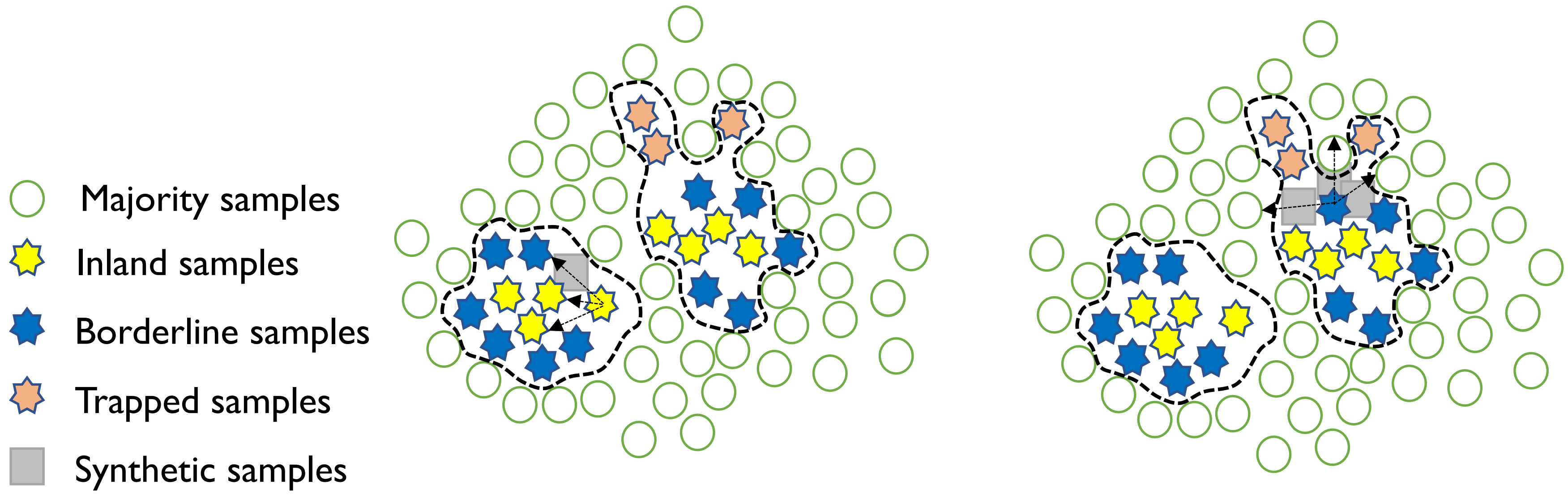
Fig. Interpolation-based Method

Reference: T. Zhu, Y. Lin, and Y. Liu, “Improving interpolation-based oversampling for imbalanced data learning,” Knowledge-Based Systems, 2020.

Proposed Method

Generation for inland and borderline

- For inland minority example x_i, its candidate point x_j chosen from the same cluster $\boldsymbol{L_c} \backslash \boldsymbol{x_i}$, where $x_i \in L_c$.
- For borderline minority example x_i, its candidate point x_j chosen from k_{maj} nearest majority neighbors $\boldsymbol{N_{maj}(x_i)}$.



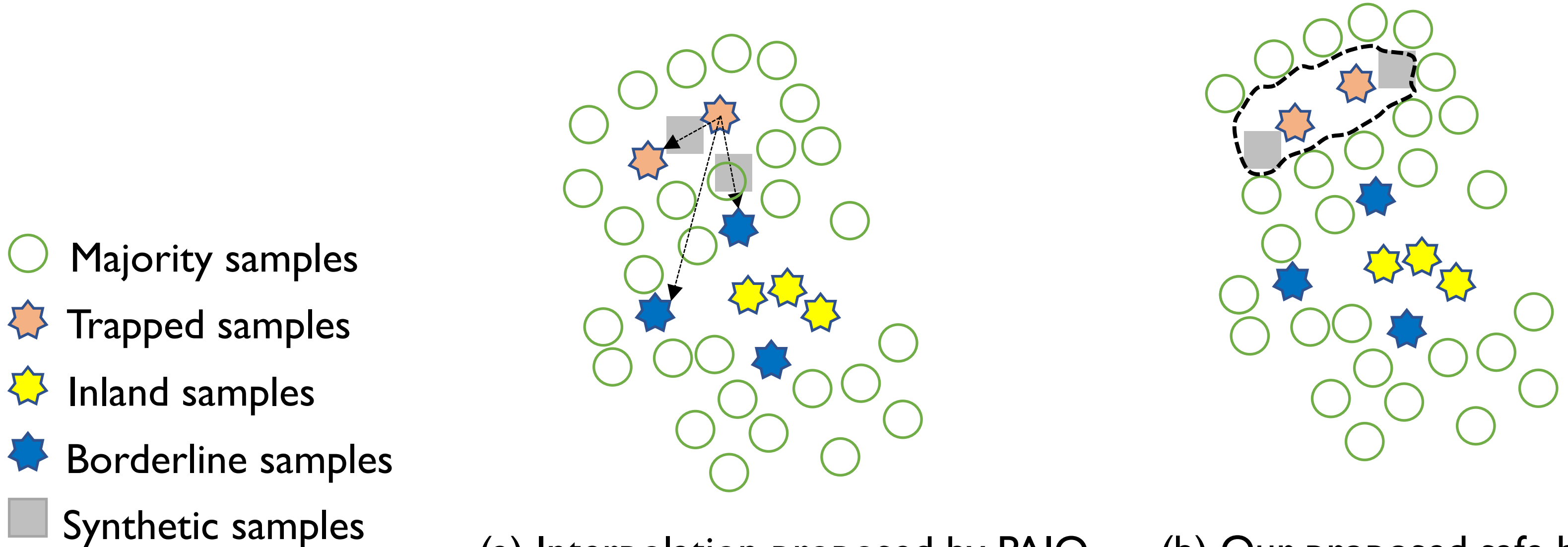
(a) Interpolation for Inland

(b) Interpolation for Borderline

Proposed Method

Generation for trapped

- We propose to learn the **safe boundary** of **trapped** samples to generate synthetic points with following hypothesis:
    - The embedding vectors of synthetic instances should be more similar to its corresponding trapped instance, than to any other majority instance.



(a) Interpolation proposed by PAIO (b) Our proposed safe boundary

Reference: T. Zhu, Y. Lin, and Y. Liu, "Improving interpolation-based oversampling for imbalanced data learning," Knowledge-Based Systems, 2020.

Proposed Method

Generation

- We propose to learn the **safe boundary** of **trapped** samples to generate synthetic points with following hypothesis:
  - We define a partial loss l for a synthetic instance s as follows:

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

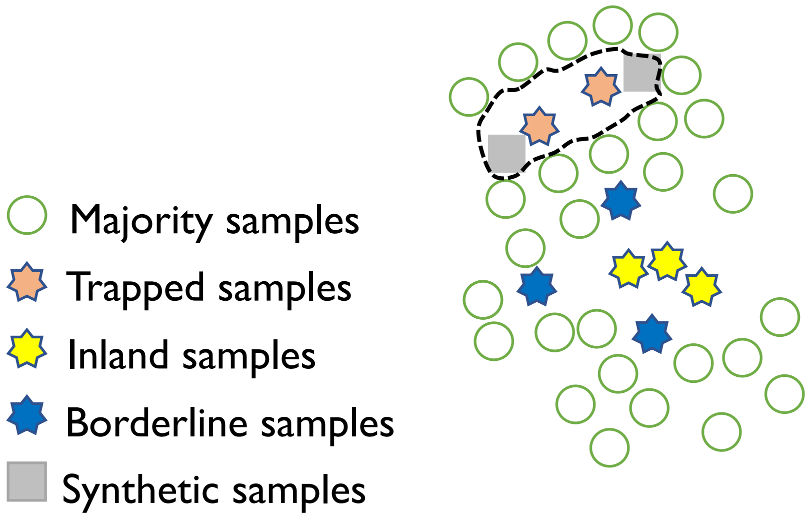
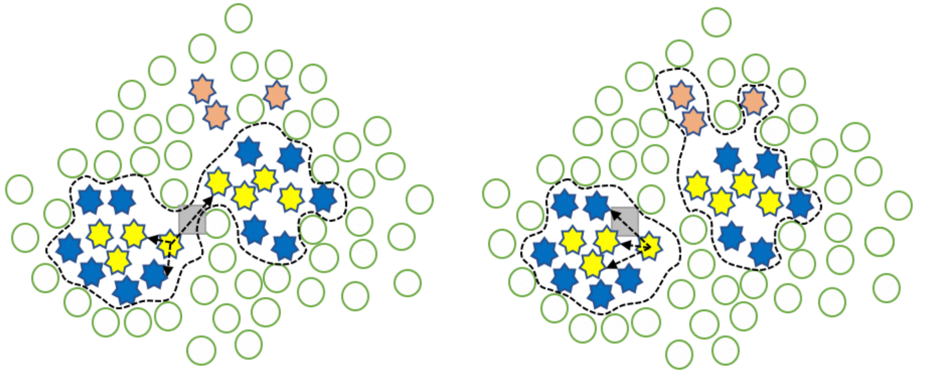


Fig. Our proposed safe boundary

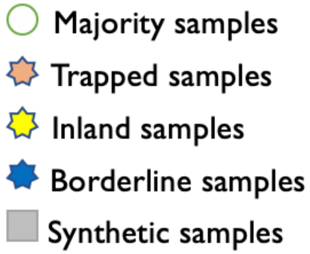
Proposed Method

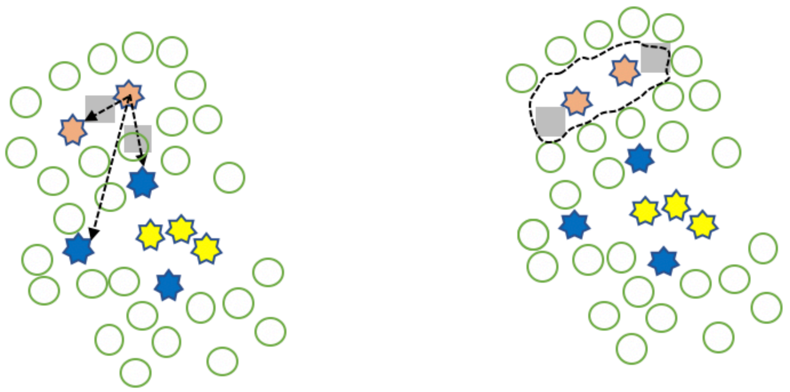
Contribution

- **Clustering**
  - We employ CFSFDP clustering algorithm to alleviate improper clustering in PAIO.
- **Generation**
  - We propose to learn the **safe boundary** of **trapped** samples to avoid noises.



(a) Clustering adopted by PAIO (b) CFSFDP clustering





(a) Interpolation proposed by PAIO (b) Our proposed safe boundary

1. T. Zhu, Y. Lin, and Y. Liu, "Improving interpolation-based oversampling for imbalanced data learning," Knowledge-Based Systems, 2020.
2. A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, 2014.

Outline

- **Background**
- **Literature**
- **Proposed Method**
  - Motivation
  - Formulation
  - Experiment
- **Future Work**

Experiment

- Data sets:
  - Five classical Imbalanced data sets from UCI repository
- Compared with other nine oversampling algorithms:
  - ROS
  - SMOTE[1] and its variant: safe-SMOTE[2]
  - MWMO[3]
  - SMOM[4]
  - INOS[5]
  - MDO[6]
  - RACOG[7]
  - PAIO[8]
- Use two classical classifiers: Linear SVM and C4.5 decision tree

Experiment

Data sets:

Data	# Min class	# Maj class	# Min examples	# Maj examples	# numeric features	Imbalance ratio
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

- Description of inland, borderline and trapped and generated synthetic samples

Data	# I	# B	# T	# S_I	# S_B	# S_T
Pima	135	12	9	135	57	76
Ecoli	43	12	9	69	68	761
Vowel	88	2	0	580	302	0
Yeast	10	14	57	72	180	1012
AB1	0	0	99	0	0	3930

Experiment

Metrics

- Precision and Recall
  - Precision = TP/(TP+FP)
  - Recall = TP/ (TP+FN)
- F1-score: harmonic mean between precision and recall
  - F1-score = $\frac{2*precision*recall}{precision+recall}$
- G-mean: balance between the classification performance on both the majority and minority samples
  - G-mean = $\sqrt{\frac{\mathrm{TP}}{TP+FN} * \frac{TN}{TN+FP}}$
- AUC: area under the receiver operating characteristics curve

Experiment

- F1-score, G-mean, and AUC values 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**

Experiment

- F1-score, G-mean, and AUC values of all the oversampling methods on each numerical imbalanced dataset using **C4.5 decision tree**.

Data	Metrics	None	ROS	SMOTE	Safe-SMOTE	MWMO	SMOM	INOS	MDO	RACOG	PAIO	PABIO
pima	F1-score	0.6061	0.6188	0.6378	0.635	0.6396	0.6407	0.6206	0.6397	0.5562	0.6414	**0.6968**
	G-mean	0.6879	0.7002	0.713	0.7139	0.7166	0.716	0.7021	**0.7174**	0.6452	0.7157	0.7097
	AUC	0.7619	0.7676	0.7708	0.7728	0.7644	0.7722	0.7632	**0.785**	0.7071	0.7689	0.7121
Ecoli	F1-score	0.648	0.6432	0.6847	0.6592	0.6698	0.6879	0.694	0.6657	0.5724	0.6725	**0.7539**
	G-mean	0.7453	0.776	0.8222	0.7925	0.8202	0.8275	0.8286	0.7637	0.7319	0.8239	**0.8606**
	AUC	0.8745	0.8961	0.9015	0.896	0.8832	0.9052	0.898	0.857	0.8076	0.8952	**0.9107**
vowel	F1-score	0.7943	0.783	0.7866	0.7934	0.8043	0.794	0.7589	0.7602	0.5113	0.7979	**0.8477**
	G-mean	0.8681	0.884	0.9061	0.9022	0.901	0.9144	0.9399	0.9333	0.8312	0.9241	**0.9763**
	AUC	0.9369	0.9487	0.954	0.9541	0.9496	0.9566	0.9687	0.9648	0.9002	0.9615	**0.9764**
Yeast	F1-score	0.3047	0.3552	0.3345	0.359	0.3142	0.3601	0.387	0.3977	0.3021	0.3637	**0.4286**
	G-mean	0.4403	0.629	0.6989	0.5956	0.7178	0.7531	0.6949	0.5627	0.8068	0.7296	**0.8314**
	AUC	0.7504	0.8433	0.833	0.8221	0.8333	0.851	0.8674	0.8125	0.8339	0.8523	**0.8926**
Abalone	F1-score	0.0769	0.0961	0.1331	0.1154	0.1251	0.1333	0.1253	0.1157	0.0833	0.135	**0.1753**
	G-mean	0.0045	0.3172	0.5021	0.3054	0.5009	0.5614	0.46	0.0109	0.6508	0.4816	**0.7062**
	AUC	0.5093	0.7783	0.7832	0.7688	0.7825	0.79	0.7852	0.5742	0.6946	0.787	**0.8351**

Experiment

Summary

- In terms of **F1-score**, our PABIO achieves the best results of all five data sets, either classified by linear SVM or decision tree.
- In terms of **G-mean**, our PABIO outperforms most of the five data sets.
  - Both high precision and recall
- In terms **robustness**:
  - Vowel dataset: <u>no trapped example</u>
    - Our proposed PABIO discovers more dense minority groups, which generates synthetic inland samples safely.
  - Abalone dataset: <u>only has trapped examples</u>
    - Our proposed PABIO learns safe boundary of interpolation, which can expand the minority region effectively and not introduce additional noise points.

Experiment

Hyperparameters

- To compare our proposed algorithm with PAIO, we adopt the recommended values of the common parameters in it.
  - The number of nearest neighbors $\mathrm{m} = 8$
  - The density threshold to divide minorities $\rho Th = 0.5$
  - The number of majority nearest neighbors to generate synthetic borderline samples $k_{maj} = 0.5$

Experiment

Hyperparameters

- Our proposed PABIO oversampling depends on the clustering of minorities, thus the cut-off distance: d_c of the clustering algorithm is crucial.
  - The value range of d_c is from 4% to 10%.

- **Findings**: Most of the five datasets have several the same F1-score, as if d_c falls into an appropriate range, it would result in the same clustering result, further, the same F1-score.

d_c \ Data	Pima	Ecoli	Vowel	Yeast	Abalone
4%	**0.6667**	0.7107	**0.5385**	**0.4715**	0.1544
6%	**0.6667**	**0.7710**	**0.5385**	**0.4715**	0.1851
8%	0.5991	0.6721	0.5008	0.3003	**0.2286**
10%	0.5991	0.6721	0.5008	0.3003	**0.2286**

Table. F1-score of Proposed PAIO Classified By Linear-SVM Varying Cut-off Distance.

Outline

- Introduction and background
- Literatures
- Proposed Method
- Future Work

Future Work

- Integrate proposed oversampling e.g. PABIO **with data cleaning, additional classifiers, or classifier ensembles** etc.
    - The majority class has the main concept of data, may also includes noise examples.
        - Integrate oversampling with the existing under-sampling.
        - Propose cluster-based under-sampling to identify overlapping borderline examples.
- Extend proposed oversampling on **data sets mixed with numerical and categorical variables**.
    - Distance metrics between categorical variables.
    - Interpolate meaningful synthetic categorical variables.
- Evaluate proposed oversampling on biological datasets, which usually have extreme high imbalance ratio, such as 10,000: 1.

Reference

1. N.V. Chawla, K.W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: synthetic minority over-sampling technique. J. Artif. Int. Res. (JAIR), 16:321–357, 2002.
2. C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap. Safe-level-SMOTE: Safe level synthetic over-sampling technique for handling the class imbalanced problem. In Proc. PAKDD 2009, volume 5476 of Springer LNAI, pages 475–482, 2009.
3. S. Barua, M. M. Islam, X. Yao, and K. Murase, "Mwmote–majority weighted minority oversampling technique for imbalanced data set learning," IEEE TKDE, vol. 26, no. 2, pp. 405–425, 2012.
4. T. Zhu, Y. Lin, and Y. Liu, "Synthetic minority oversampling technique for multiclass imbalance problems," Pattern Recognition, vol. 72, pp. 327–340, 2017.
5. L. Abdi and S. Hashemi, "To combat multi-class imbalanced problems by means of over-sampling techniques," IEEE TKDE, vol. 28, no. 1,pp. 238–251, 2015.
6. H. Cao, X.-L. Li, D.Y.-K. Woon, and S.-K. Ng, "Integrated oversampling for imbalanced time series classification," IEEE Transactions on
7. Knowledge and Data Engineering, vol. 25, no. 12, pp. 2809–2822, 2013. [34] B. Das, N. C. Krishnan, and D. J. Cook, "Racog and wracog: Two probabilistic oversampling techniques," IEEE transactions on knowledge and data engineering, vol. 27, no. 1, pp. 222–234, 2014.
8. T. Zhu, Y. Lin, and Y. Liu, "Improving interpolation-based oversampling for imbalanced data learning," Knowledge-Based Systems, vol. 187, p. 104826, 2020.

Thank you!