Synthetic oversampling with the majority class

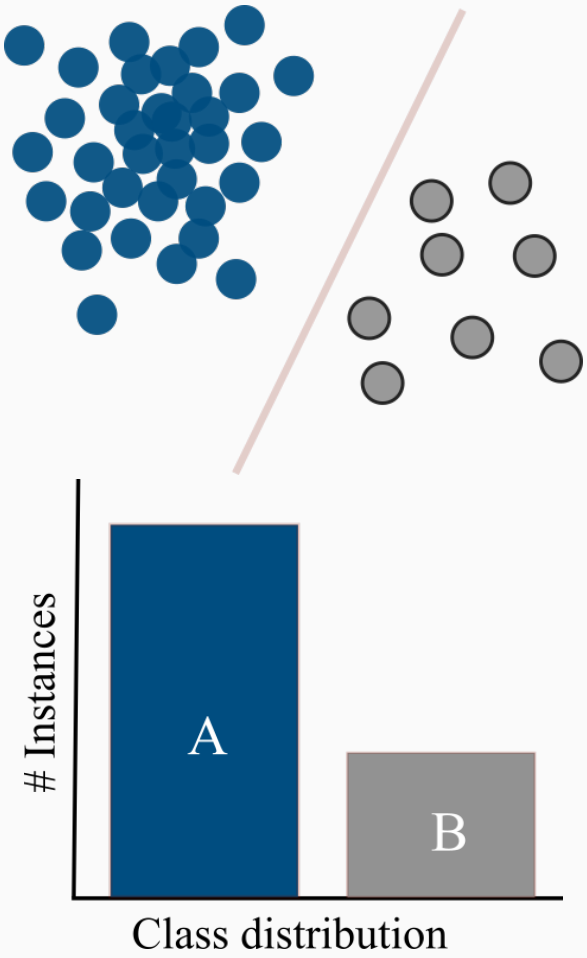
A new perspective on handling extreme imbalance

Colin Bellinger

National Research Council of Canada
Ottawa, Canada
Colin.Bellinger@nrc-cnrc.gc.ca
https://web.cs.dal.ca/~bellinger

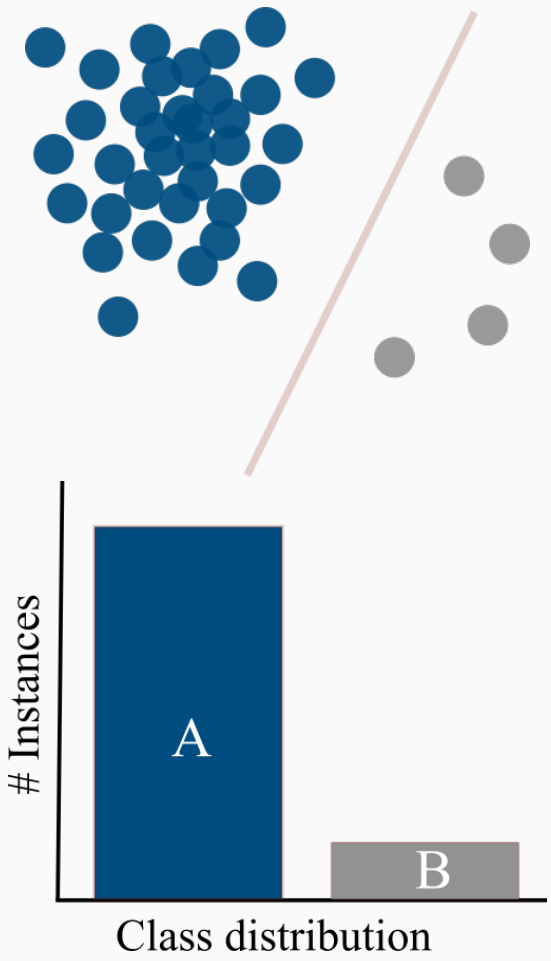
Class Imbalance

- Focus $\rightarrow$ classification problems
  - Experiments are binary
  - Can generalize to multi-class
- Standard class imbalance (not extreme):
  - Known to negatively impact classifier performance
  - Impact managed via the standard toolbox
    - Resampling
    - Cost adjustment



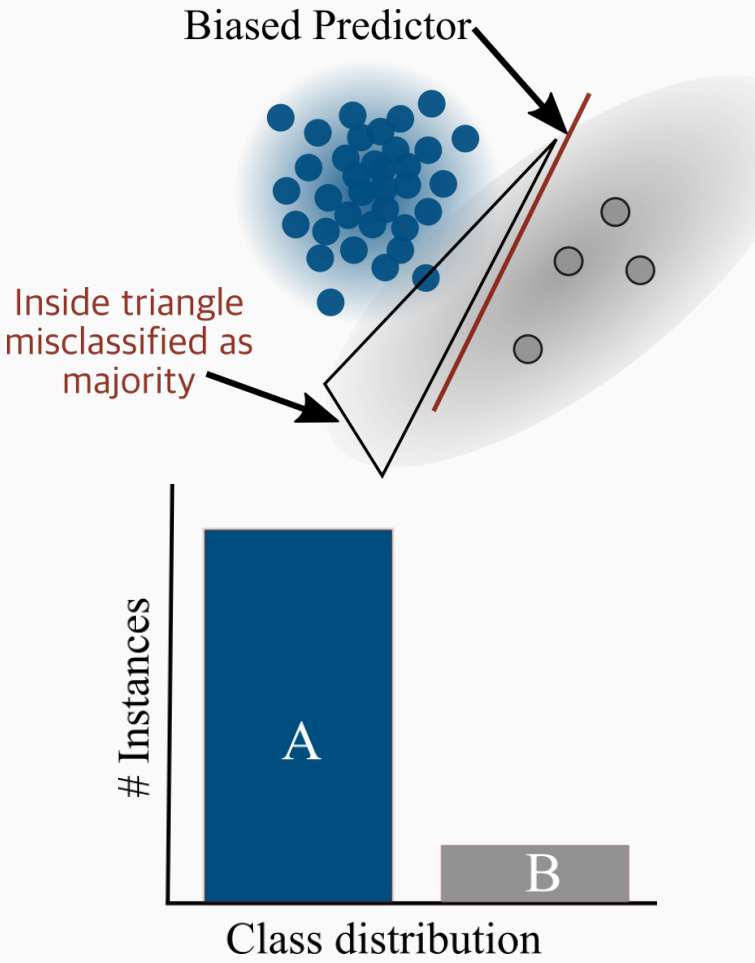
Extreme Class Imbalance Definition

- Extreme imbalance:
  - High imabalance ratio between the large class(es) and the small class(es) ($> 1:100$)
  - Very low absolute number of minority class instances (< 15).
- Application areas:
  - Fault detection, disease classification, software failures and customer churn prediction, *etc.*
- Radioactive threat detection
  - Classify rare but dangerous occurences of specific isotopes
  - Sign of potential security threath, problem at nuclear powerplant, *etc.*
  - Ratio: $< 1:10,000$

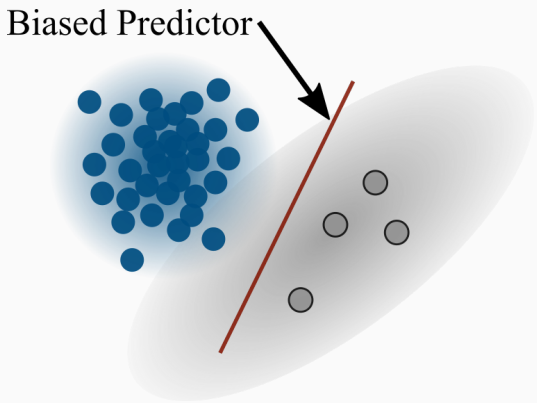


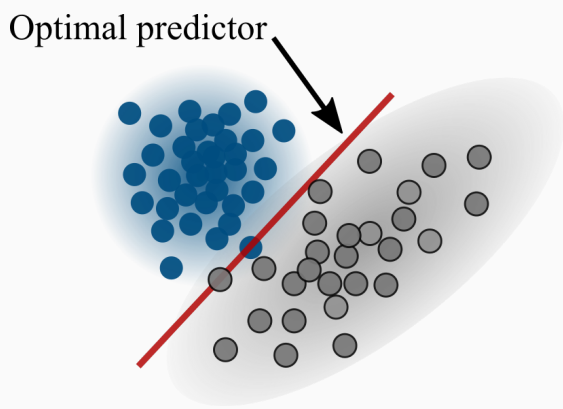
Extreme Class Imbalance Affect

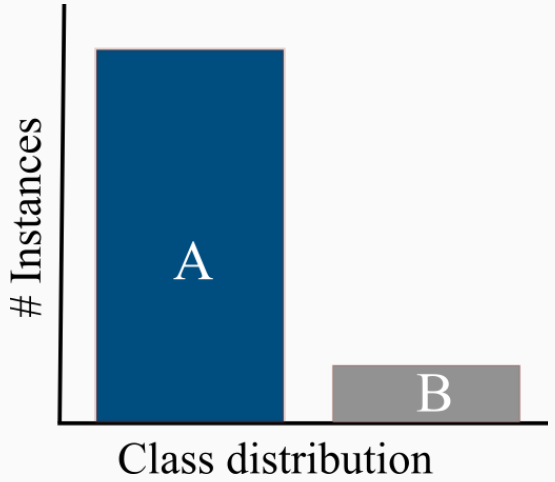
- Impact of extremely imbalanced
  - Highly biased predictors
  - Erroneously biased in favour of majority class predictions



Extreme Class Imbalance Affect







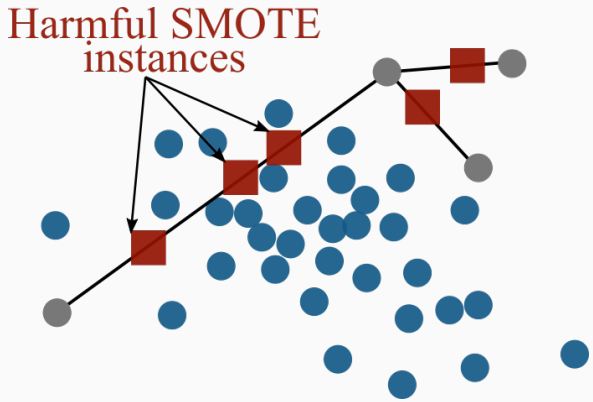
- Representative dataset
  - Very different optimal desicion boundary
    - Relative position
    - Angle
    - Shape
- Goal $\rightarrow$ use resampling to shift biased predictor towards optimal predictor

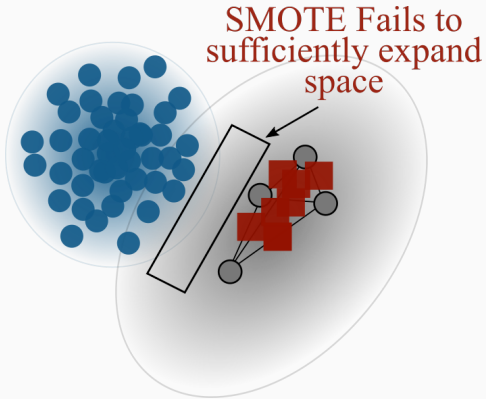
Limitations of Standard Methods

- Standard methods fail on extreme imbalance
  - Resampling: undersampling, oversampling, synthetic oversampling
  - Cost adjustment
- SMOTE: Synthetic Minority Oversampling TEchnique is the standard approach
- Numerous variations proposed to mitigate weaknesses
  - Posthoc cleaning
  - Borderline SMOTE
  - Adaptive synthetic oversampling
  - Manifold-based synthetic oversampling
- Atleast a minimum number of minority samples required for success
  - No solutions for extreme imbalance!

Limitations of SMOTE

- SMOTE algorithm
  - Interpolated synthetic instances between k-nearest neighbours in minority class
  - Populates convex-hull
    - Can be too small or too large
- Extreme imbalance $\rightarrow$ potentially large distance between neighbours
  - Synthetic instances in low-probability regions
  - Synthetic instances in harmful regions
- Extreme imbalance $\rightarrow$ potentially clustered in small unrepresentative region
  - Synthetic instances reinforce overfitting





Research Question

Research Question

Is there an effective methodology for synthetic oversampling the minority class in domains that exhibit extreme imbalance?

Approach

- SWIM: Sampling WIth the Majority
  - Utilizes the distribution of the majority class to inform the generation of synthetic istances
  - Alternative resampling methods laregly ignore the majority class
  - Alternative anomaly detection and one-class classifiers ignore informative labels altogeher!
- Intuition behind SWIM
  - Synthesize instances in regions with similar densities w.r.t. majority class as minority class seeds
  - Synthesize instances in regions that neighbour minority class seeds.

SWIM Algorithm Details

SWIM Algorith:

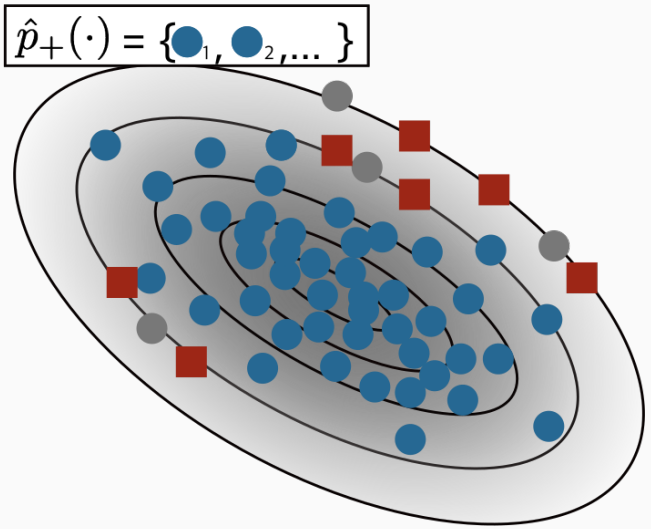
1) Estimate the PDF $\hat{p}_+(\cdot)$ of the majority class.

2) Repeat *n* times:

   2.1) Randomly select a minority class seed x_-

   2.2) Synthesize a minority instance x' from x_- such that r shifts x_- to a neigbouring region with same density

   - $x' = x_- + r$
   - $\hat{p}_+(x') \approx \hat{p}_+(x_-)$



SWIM Algorithm Implementation

- Utilized Mahalanobis Distance to estimate density
  - Pro: fast and effienient
  - Con: distribution assumption
  - Works well in practice
- Alternatives include GMM and RBF

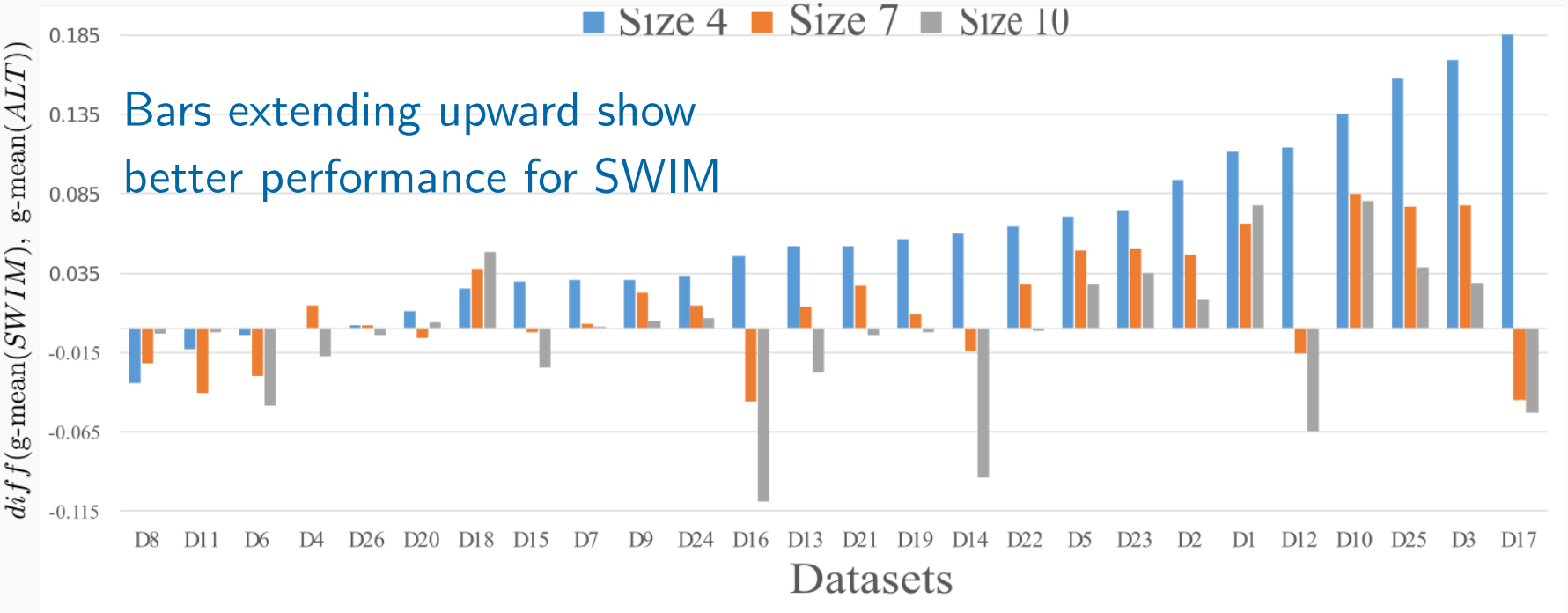
Methodology

- Resampling methods: SWIM versus SMOTE, SMOTE with one-sided selection, SMOTE with the removal of Tomek links, and borderline SMOTE
- Classification methods: Naive Bayes, kNN, Decision Trees, Multilayer Perception,SVM
- Evaluation:
  - 26 Benchmark UCI datasets
  - 30 independent trials
  - Average g-mean$= \sqrt{TPR \times TNR}$
  - Statistical significance test via Bayesian signed test

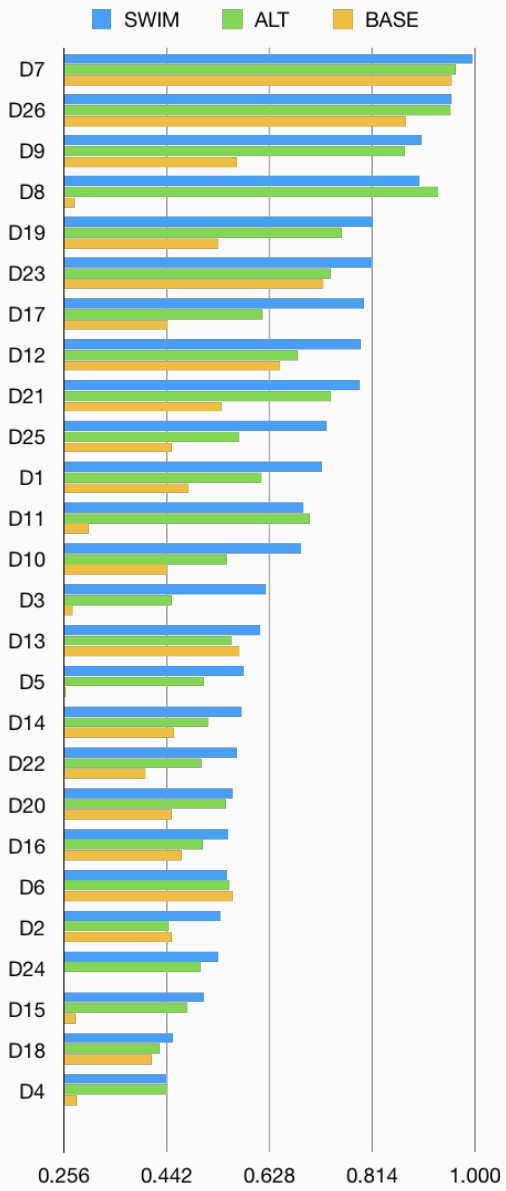
Dataset	Name	Dim.	Maj. Size	R4	R7	R10
D1	Abalone 9-18	8	689	1:173	1:99	1:69
D2	Ada Agnostic	48	3430	1:858	1:490	1:343
D3	Alphabets	15	3077	1:770	1:440	1:308
D4	Analcat Data DMFT	7	642	1:161	1:92	1:65
D5	Diabetes	8	500	1:125	1:72	1:50
D6	Forest Cover	54	2970	1:743	1:425	1:297
D7	KDD Synthetic Control	61	500	1:125	1:72	1:50
D8	Mfeat Karhunen	64	1800	1:450	1:258	1:180
D9	Delft pump AR	160	531	1:133	1:76	1:54
D10	Spambase spam	57	2788	1:697	1:399	1:279
D11	Waveform 0	21	600	1:150	1:86	1:60
D12	Page Blocks	10	4913	1:1229	1:702	1:492
D13	PC4	37	1280	1:320	1:183	1:128
D14	Piechart	37	644	1:161	1:92	1:65
D15	Pima Indians	8	500	1:125	1:72	1:50
D16	Pizza Cutter	37	609	1:153	1:87	1:61
D17	Ring Norm	20	3736	1:934	1:534	1:374
D18	Thoracic Surgery	37	400	1:100	1:58	1:40
D19	Vehicle 0	18	647	1:162	1:93	1:65
D20	Vehicle 1	18	629	1:158	1:90	1:63
D21	Vehicle 2	18	628	1:157	1:90	1:63
D22	Vehicle 3	18	634	1:159	1:91	1:64
D23	Vowel 10	13	898	1:225	1:129	1:90
D24	Wine Quality Red 4	11	1546	1:387	1:221	1:155
D25	Wine Quality White 3 vs 7	11	880	1:220	1:126	1:88
D26	Wisconsin	9	444	1:111	1:64	1:45

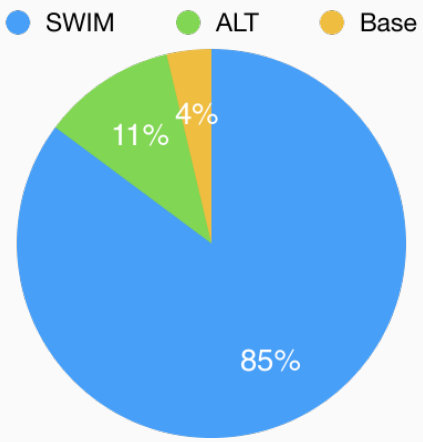
Results - Minority Size 4, 7, and 10

- Comparison of relative performane
  - Minority size 4,7 and 10
  - Relative performance $diff(SWIM(D_i) - ALT(D_i))$



Detailed Results - Minority Size 4





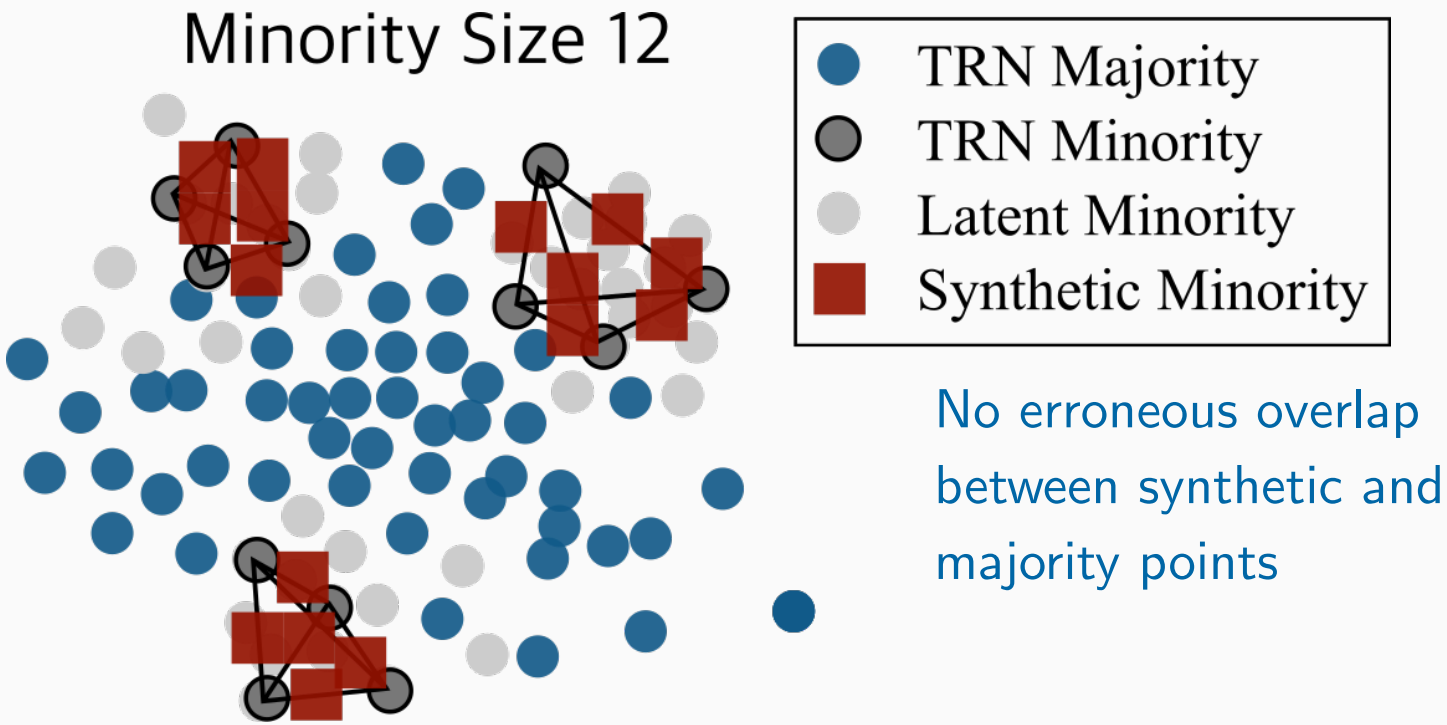
Result

- SWIM outperforms the best alternative methods on the vast majority of exteme imbalanced datasets
- Statistically verified with Bayesian signed test
- SWIM $>$ ALT on size 7 and 15 as well

Discussion

Impact of Minority Class Size

- Relative advantage of SWIM inversely related to minority class size
  - Likelihood of nearby *kNN* in SMOTE increases with minority class size

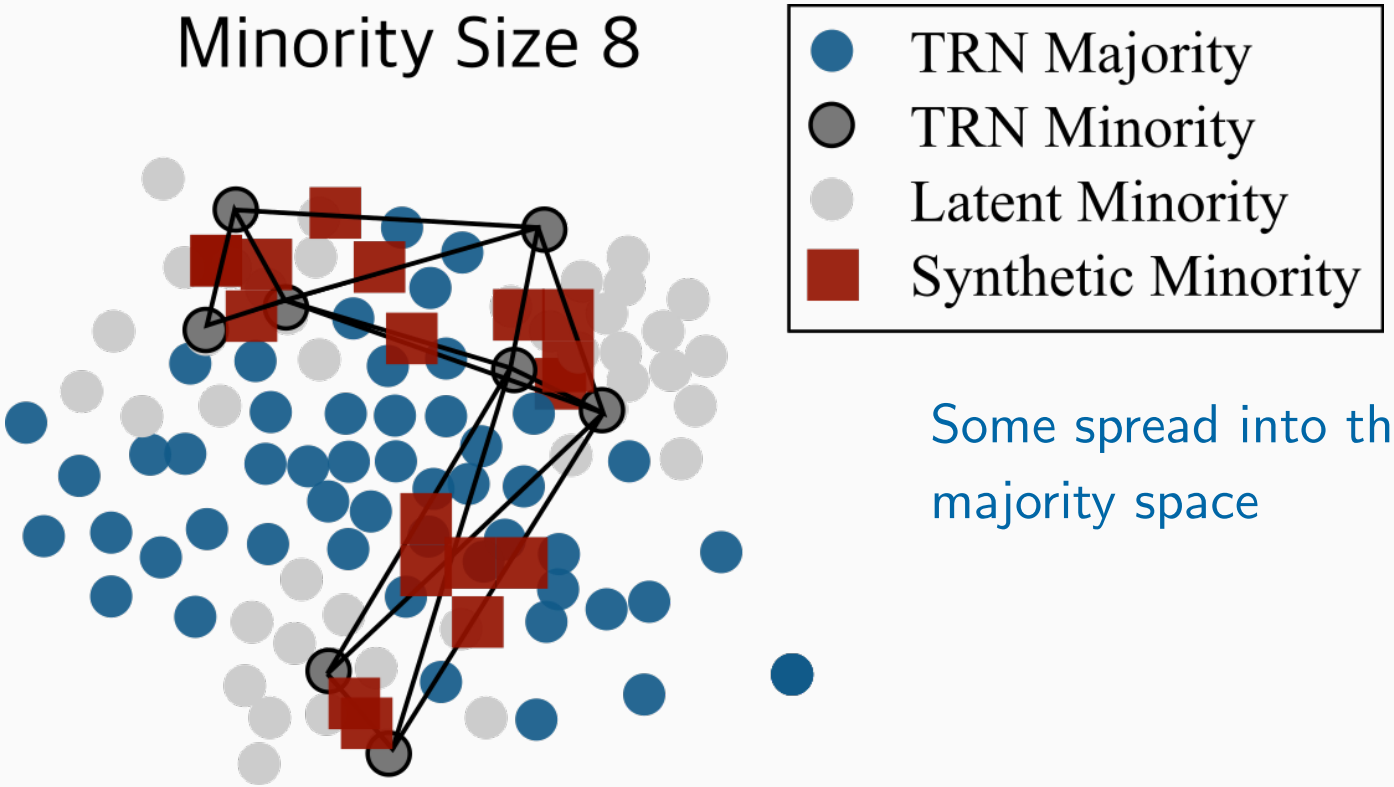


No erroneous overlap between synthetic and majority points

Discussion

Impact of Minority Class Size

- Relative advantage of SWIM inversely related to minority class size
  - Likelihood of nearby *kNN* in SMOTE increases with minority class size

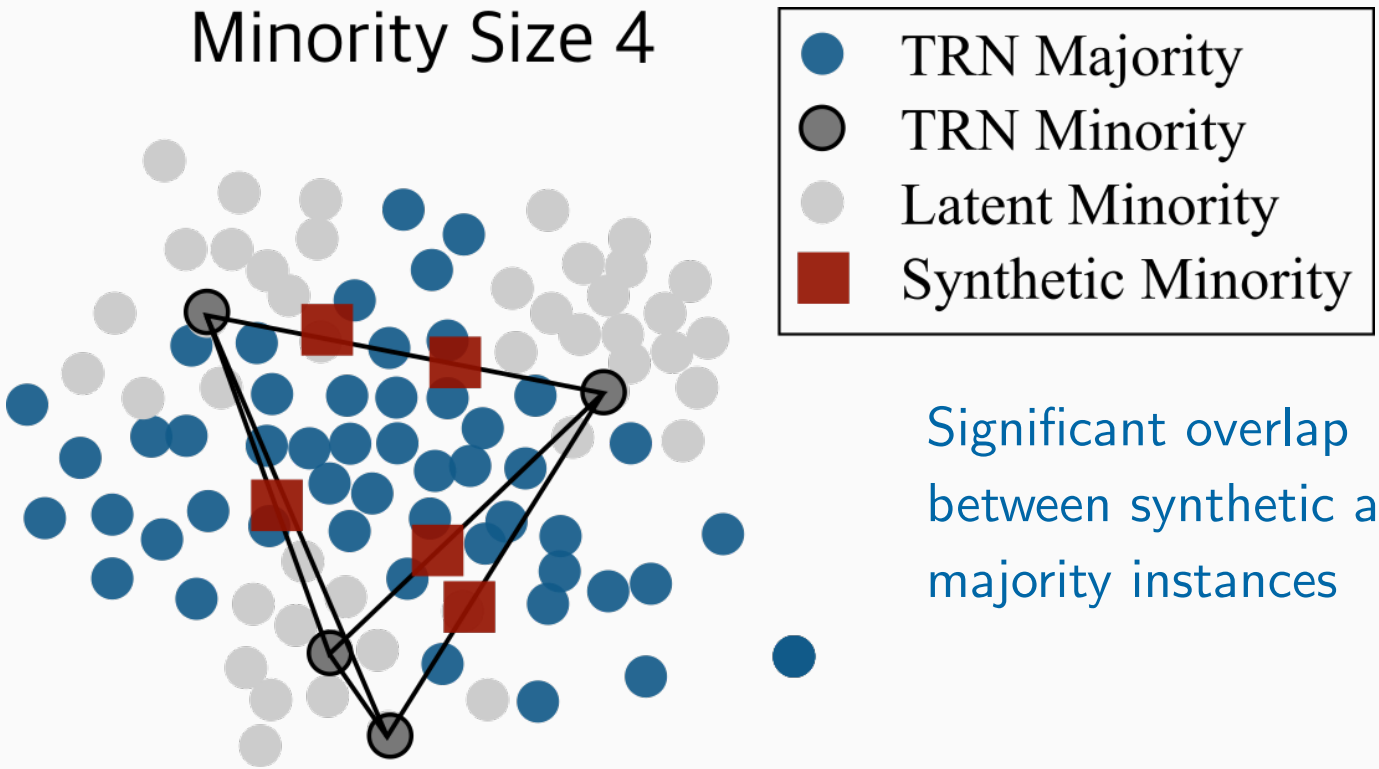


Some spread into the majority space

Discussion

Impact of Minority Class Size

- Relative advantage of SWIM inversely related to minority class size
  - Likelihood of nearby *kNN* in SMOTE increases with minority class size



Significant overlap between synthetic and majority instances

Conclusion

- Standard methods of synthetic oversampling fail on extreme imbalance
- A majority focused synthetic oversampling method is required in these cases
  - Use the distribution of the majority class
  - Synthesize instances based on minority seeds density w.r.t. majority class
  - Synthetic samples in neighbouring regions with similar density
- SWIM $\rightarrow$ simple and efficient to implenent apply
- SWIM $\rightarrow$ greater advantage on more extreme imbalance

Questions?

Thank you!

colin.bellinger@nrc-cnrc.gc.ca
https://web.cs.dal.ca/ bellinger/

Backup slides

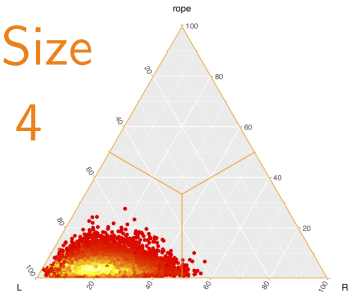
Mahalanobis distance implementation of SWIM

- Calculate sample mean and covariance on the majority class training data A: $\bar{\mu}_a = mean(A)$, $\bar{\Sigma}_a = \text{cov(A)}$
- Centre the majority and minority training data A and B with respect to A: $A_c = A - \bar{\mu}_a$, $B_c = B - \bar{\mu}_a$
- Whiten the minority class with respect to the covariance of A: $B_w = B_c \bar{\Sigma}_a^{\frac{1}{2}}$
- For each feature f in B_w calculate feature bounds as the upper u_f and lower l_f range of values of x:
  $u_f = \bar{\mu}_f + \alpha\sigma_f$
  $l_f = \bar{\mu}_f - \alpha\sigma_f$
- Given a minority reference point x in the whitened space, generate new sample s in whitened space by:
  - for each feature f, generate a random number between u_f and l_f
  - transform s to have the same Euclidean norm as: $s_{norm} = s\frac{||x||_2}{||s||_2}$
  - Scale s_{norm} to the feature space of A $s_{new} = \left(\Sigma_A^{-\frac{1}{2}}\right) s_{norm}$

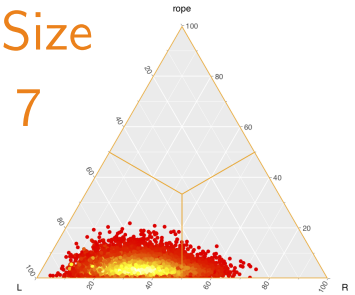
Backup slides

- Is SWIM better than ALT?
- Bayesian signed test based Dirichlet process
  - Parameter of the Dirichlet as $s = 0.5$ and $z_0 = 0$
- Skew cluster of points shows significant difference for size 4 and 7

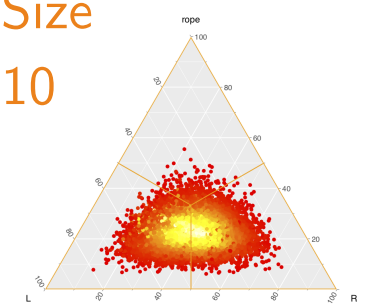
Size 4



Size 7



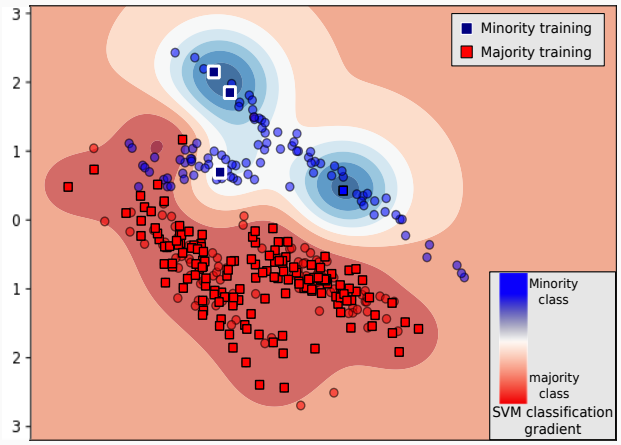
Size 10



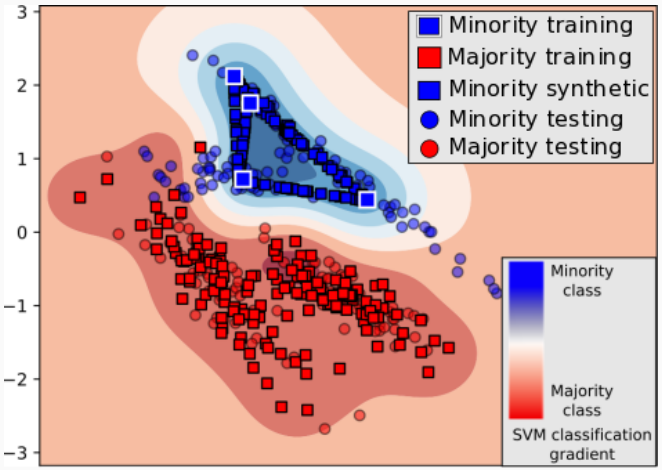
Backup slides

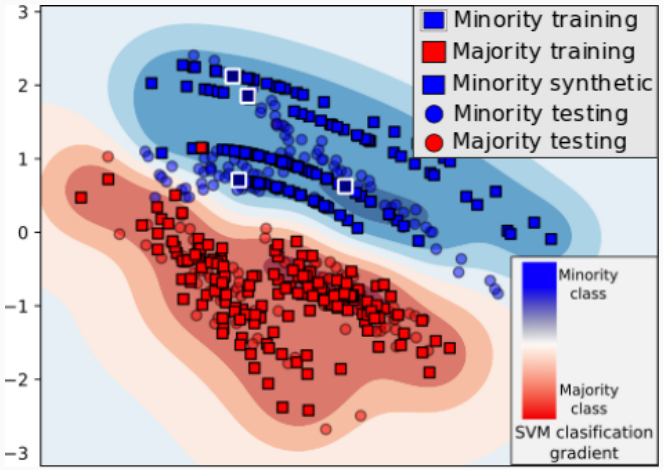
Example

SVM with RBF kernel applied 2D toy dataset with 3 minority class training instances (baseline (top), SMOTE (lower left), and SWIM (lower right))



SWIM spreads outside the convex-hull leading to better generalization





Backup slides

Dataset	Baseline	ALT	SWIM	Dataset	Baseline	ALT	SWIM
D1	0.481	0.612	**0.723**	D14	0.455	0.516	**0.576**
D2	0.451	0.445	**0.539**	D15	0.276	0.479	**0.509**
D3	0.27	0.451	**0.620**	D16	0.468	0.506	**0.552**
D4	0.279	0.440	**0.440**	D17	0.442	0.614	**0.799**
D5	0.259	0.509	**0.580**	D18	0.414	0.428	**0.453**
D6	0.561	**0.554**	0.550	D19	0.534	0.758	**0.814**
D7	0.958	0.965	**0.996**	D20	0.450	0.549	**0.560**
D8	0.274	**0.933**	0.899	D21	0.541	0.739	**0.791**
D9	0.569	0.872	**0.903**	D22	0.402	0.505	**0.569**
D10	0.440	0.550	**0.685**	D23	0.724	0.738	**0.812**
D11	0.301	**0.701**	0.688	D24	0.224	0.502	**0.535**
D12	0.647	0.679	**0.793**	D25	0.451	0.572	**0.730**
D13	0.572	0.559	**0.611**	D26	0.874	0.956	**0.958**

Backup slides

Dataset	Name	Dim.	Maj. Size	R4	R7	R10
D1	Abalone 9-18	8	689	1:173	1:99	1:69
D2	Ada Agnostic	48	3430	1:858	1:490	1:343
D3	Alphabets	15	3077	1:770	1:440	1:308
D4	Analcat Data DMFT	7	642	1:161	1:92	1:65
D5	Diabetes	8	500	1:125	1:72	1:50
D6	Forest Cover	54	2970	1:743	1:425	1:297
D7	KDD Synthetic Control	61	500	1:125	1:72	1:50
D8	Mfeat Karhunen	64	1800	1:450	1:258	1:180
D9	Delft pump AR	160	531	1:133	1:76	1:54
D10	Spambase spam	57	2788	1:697	1:399	1:279
D11	Waveform 0	21	600	1:150	1:86	1:60
D12	Page Blocks	10	4913	1:1229	1:702	1:492
D13	PC4	37	1280	1:320	1:183	1:128
D14	Piechart	37	644	1:161	1:92	1:65
D15	Pima Indians	8	500	1:125	1:72	1:50
D16	Pizza Cutter	37	609	1:153	1:87	1:61
D17	Ring Norm	20	3736	1:934	1:534	1:374
D18	Thoracic Surgery	37	400	1:100	1:58	1:40
D19	Vehicle 0	18	647	1:162	1:93	1:65
D20	Vehicle 1	18	629	1:158	1:90	1:63
D21	Vehicle 2	18	628	1:157	1:90	1:63
D22	Vehicle 3	18	634	1:159	1:91	1:64
D23	Vowel 10	13	898	1:225	1:129	1:90
D24	Wine Quality Red 4	11	1546	1:387	1:221	1:155
D25	Wine Quality White 3 vs 7	11	880	1:220	1:126	1:88
D26	Wisconsin	9	444	1:111	1:64	1:45