Novel Approaches to Mitigate Abnormal Instances in Imbalanced Datasets

for Improved Classification Performance

Surani Matharaarachchi

PhD Candidate, Department of Statistics, University of Manitoba

Data Scientist, Government of Manitoba

matharas@myumanitoba.ca

Joint work with Dr. Saman Muthukumarana, and Dr. Mike Domaratzki

University of Manitoba

The WNAR/IMS/Graybill 2024 Conference
June 9-12, 2024
Colorado State University, Fort Collins.

Outline

- Introduction
- Existing methods
- Challenges
- Novel approaches for imbalanced data
- Simulation & Application Results
- Conclusion

Introduction: Class Imbalance

- Occurs when the number of instances in different classes is significantly disproportionate.
- Examples:
  - Spam Detection
  - Fraud Detection
  - Medical Diagnosis
  - Churn Prediction
- Issue:
  - Leads to biased models
  - Decreases predictive accuracy

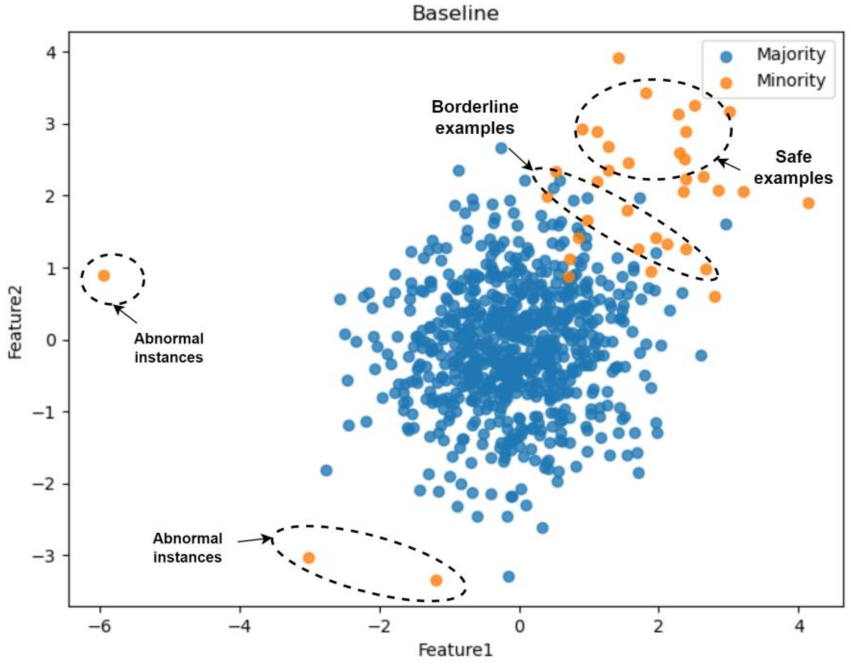


Figure: Class imbalance with outliers in the minority class

Synthetic Minority Oversampling Technique (SMOTE)

- Balancing the Dataset:
  - Strategy:
    - Create new samples for the minority class to help balance the dataset.
  - Technique:
    - Interpolate between randomly chosen minority class samples and their nearest neighbors.
    - $p_{new} = p_0 + \alpha(p_3 - p_0)$

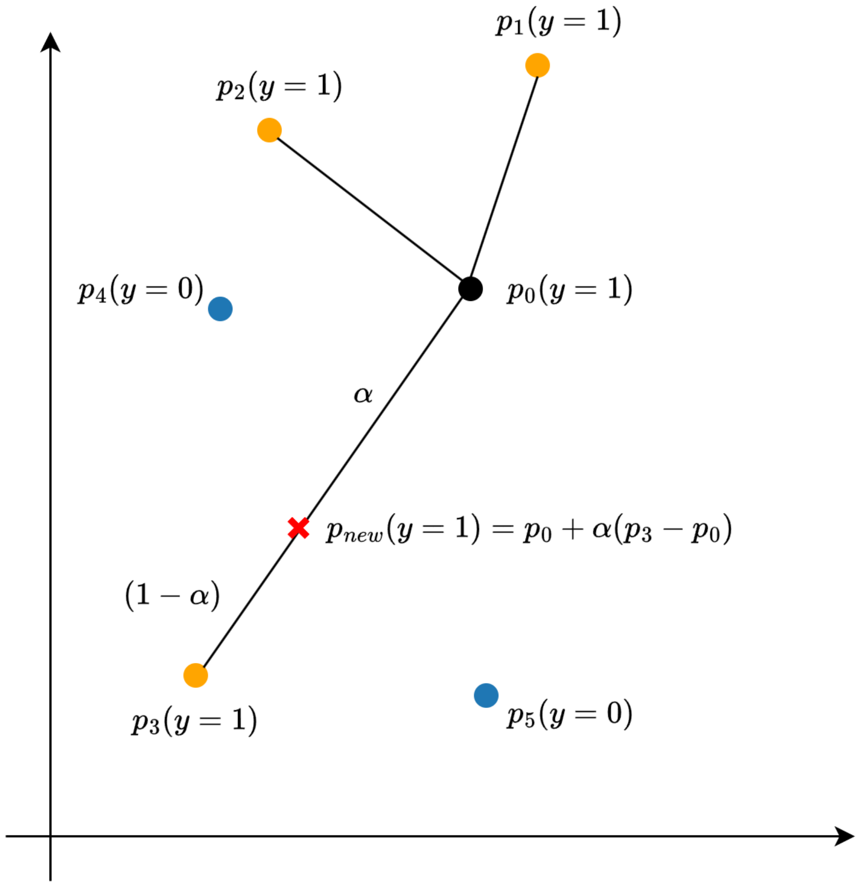


Figure: SMOTE data generation

Challenges of SMOTE

- SMOTE is challenged by outliers within the minority class.

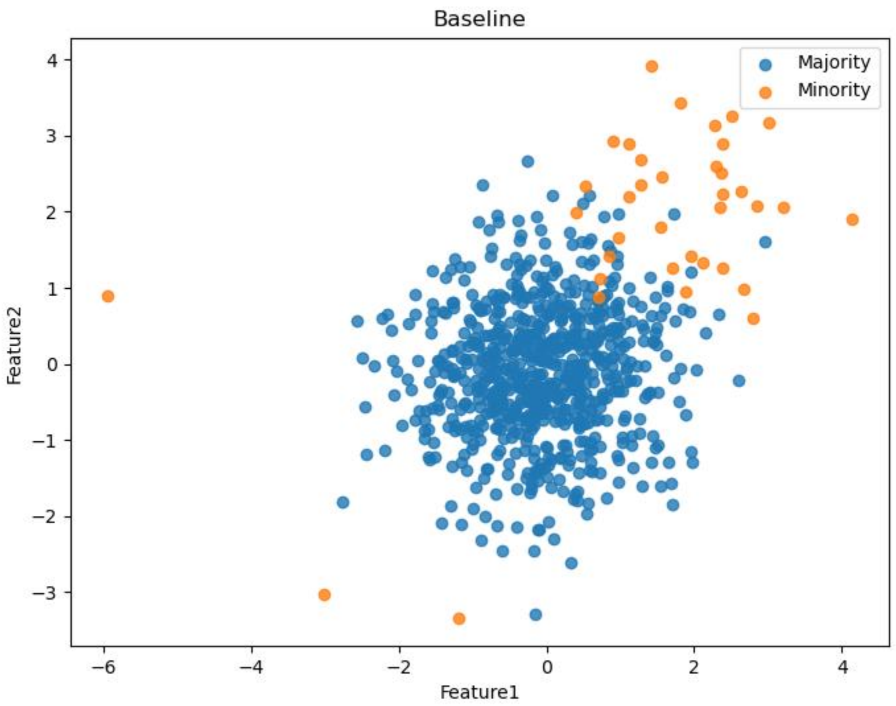


Figure: Original Data

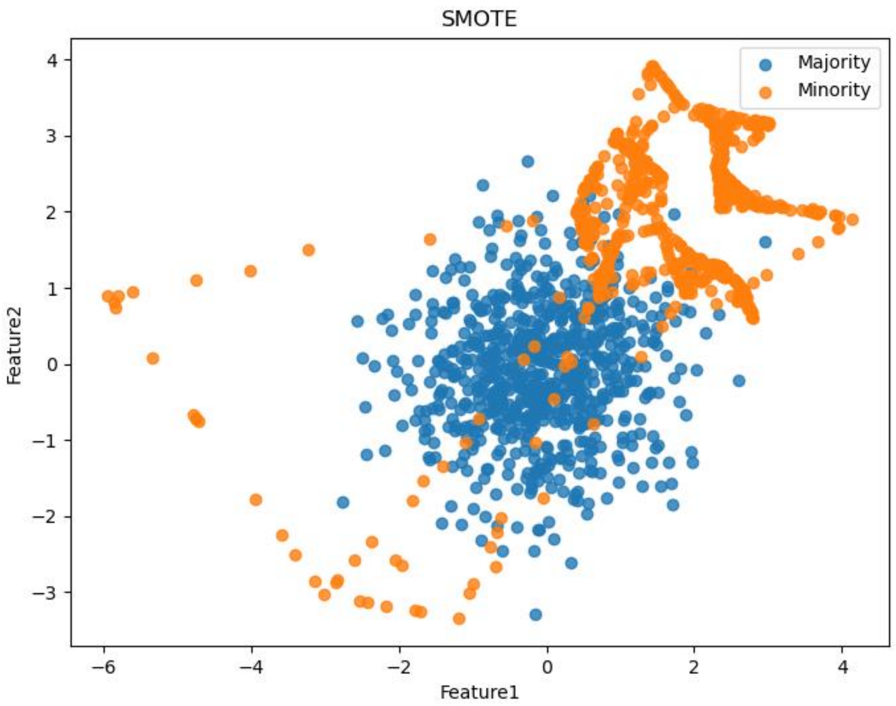


Figure: Re-sampled data with SMOTE

Novel Methods

- Technique:
  - Use a weighted average of neighbouring instances.
  - $p_{new} = \frac{\sum_{j=1}^{k}(w_j \times p_j)}{\sum_{j=1}^{k} w_j}, j = 1, \ldots, k$
  - Improve robustness against outliers and noisy data.
  - Learn from a more extensive set of nearest neighbours
- Challenge:
  - Selecting suitable weights to enhance resilience to outliers and noisy data.

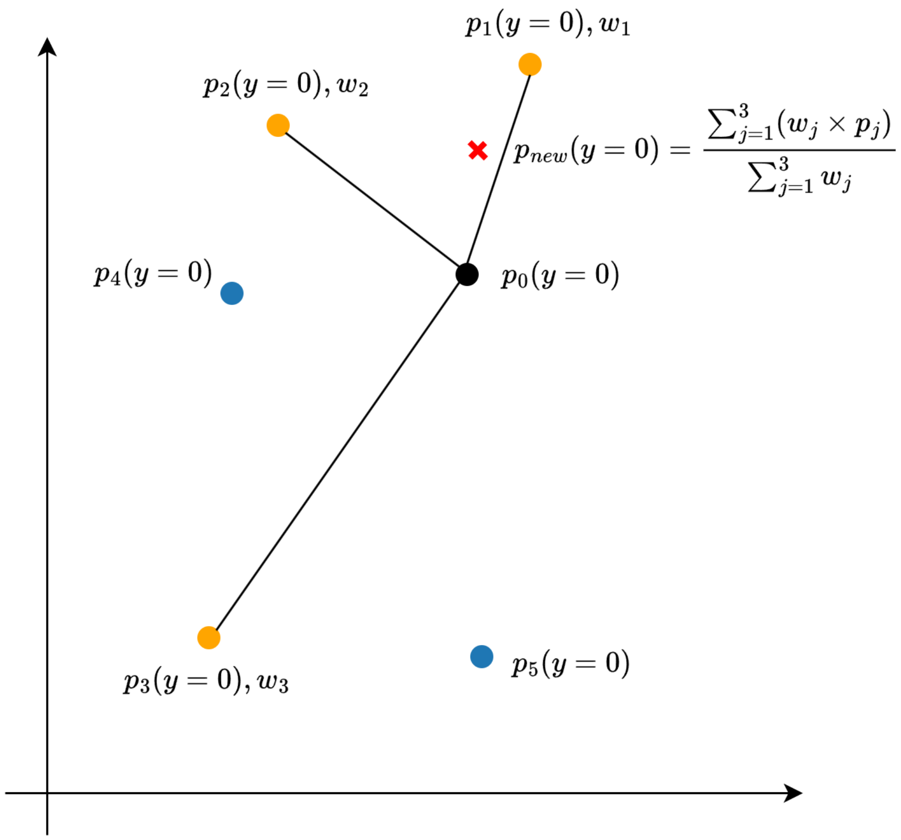


Figure: Proposed method data generation

Developing new SMOTE extensions

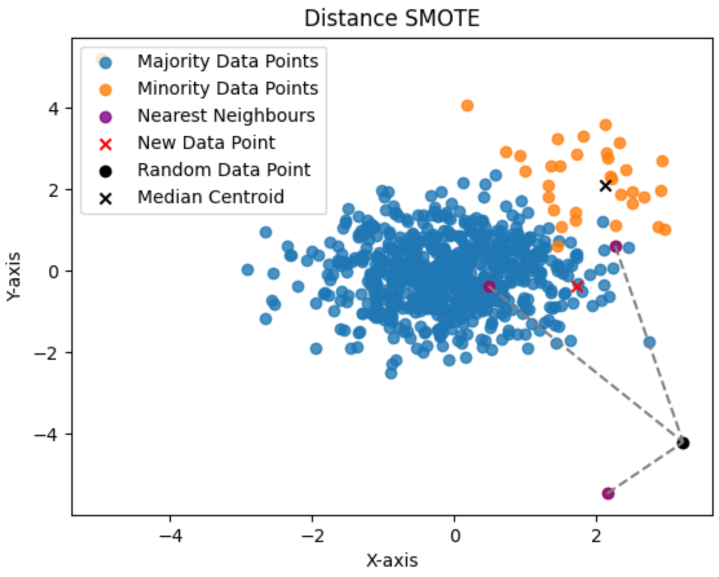
- Solution:
  - **Use inverse distance to the median centroid of the minority class**.
  - Higher weights for closer instances in feature space.

1. Distance extSMOTE
2. Dirichlet extSMOTE
   I. Uniform Random Vector
   II. Uniform Vector
   III. Inverse Distance
3. FCRP SMOTE - SMOTE with Finite Chinese Restaurant Process Idea
4. BGMM SMOTE - SMOTE with Bayesian Gaussian Mixture Model
   I. with Dirichlet prior
   II. with Dirichlet Process prior

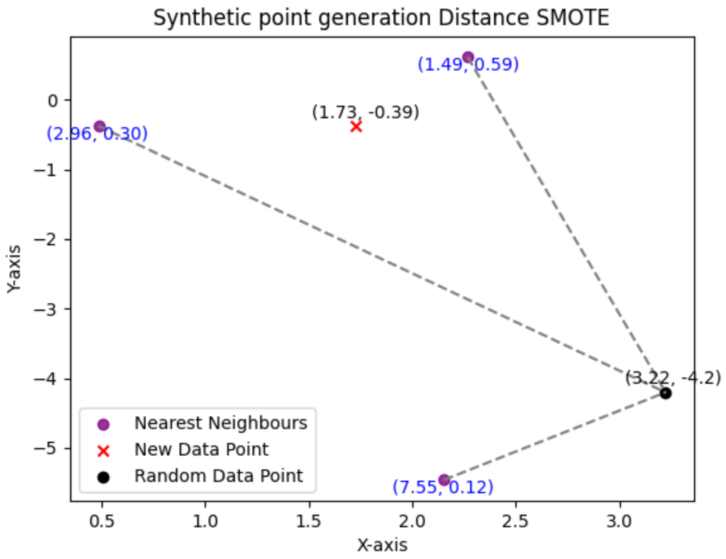
Distance extSMOTE

- $d_j \in \mathbb{R}$ is the Euclidean distance between the median centroid of the minority class and the nearest neighbours
- $w_j = d_{j,norm}^{-1}$ = Normalized inverse distance

An example of creating a sample - Distance extSMOTE



(a). This scenario occurs when an outlier is chosen as a neighbouring point.

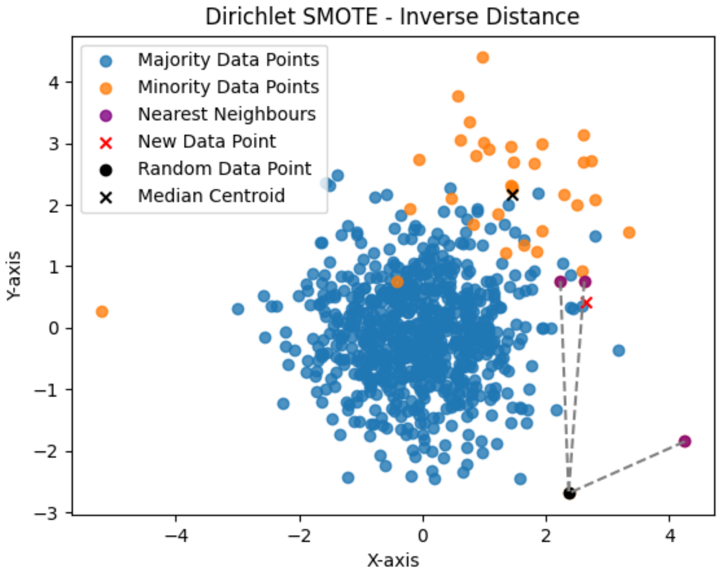


(b). The values within parentheses indicate (d_j, w_j).

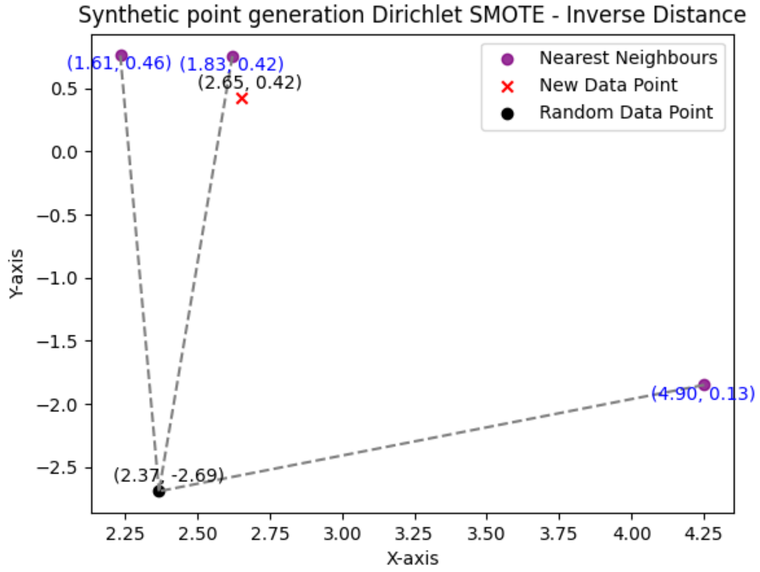
Dirichlet extSMOTE (Inverse Distance)

- $w_j = Dir(\alpha)_j$
- $\alpha = m.\boldsymbol{D}^{-1}, \boldsymbol{D} = [d_1, \dots, d_k], \boldsymbol{D}^{-1} = [\frac{1}{d_1}, \dots, \frac{1}{d_k}]$

An example of creating a sample - Dirichlet extSMOTE

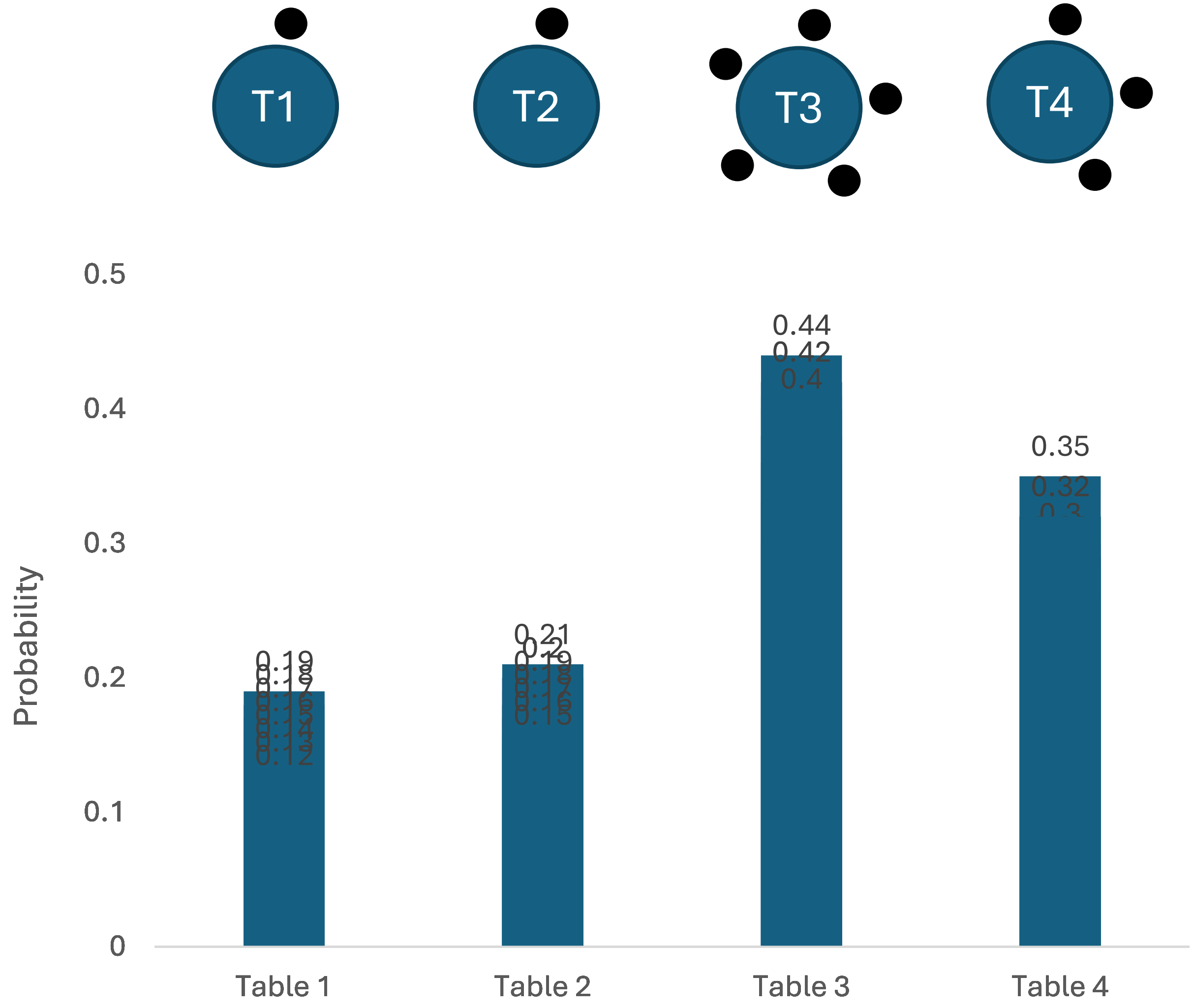


(a). This scenario occurs when an outlier is chosen as a neighbouring point.



(b). The values within parentheses indicate (d_j, w_j).

FCRP SMOTE

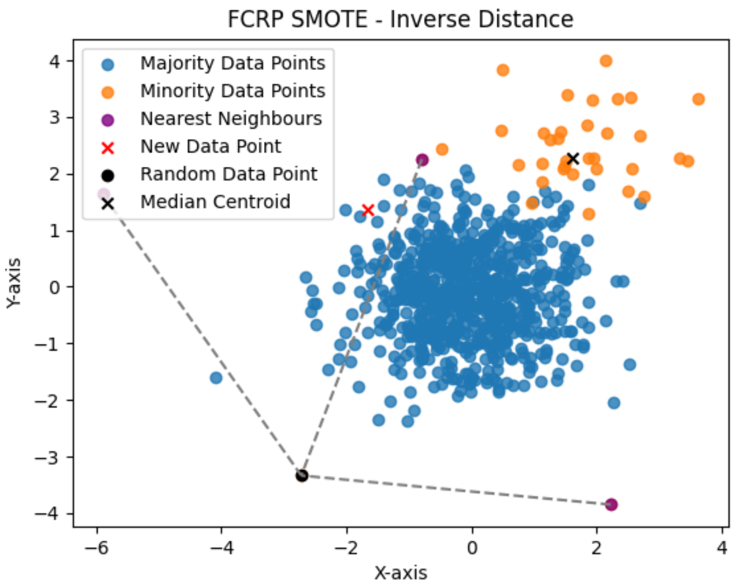


Showcasing the weight selection of FCRP SMOTE using Finite Chinese Restaurant Process with scaling parameter $\alpha = 0.1$

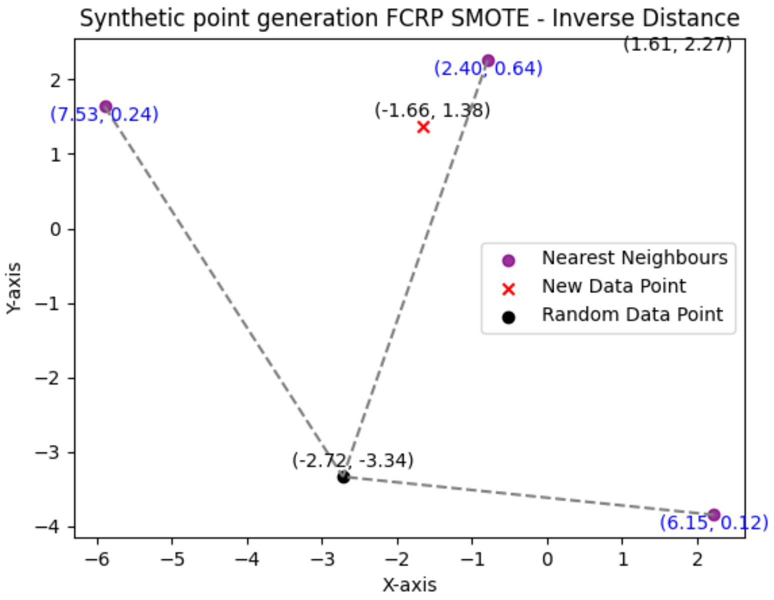
FCRP SMOTE

- Initial preferences $= d_{norm}^{-1}$
- w_j = Final allocation probabilities

An example of creating a sample – FCRP SMOTE



(a). This scenario occurs when an outlier is chosen as a neighbouring point.



(b). The values within parentheses indicate (d_j, w_j).

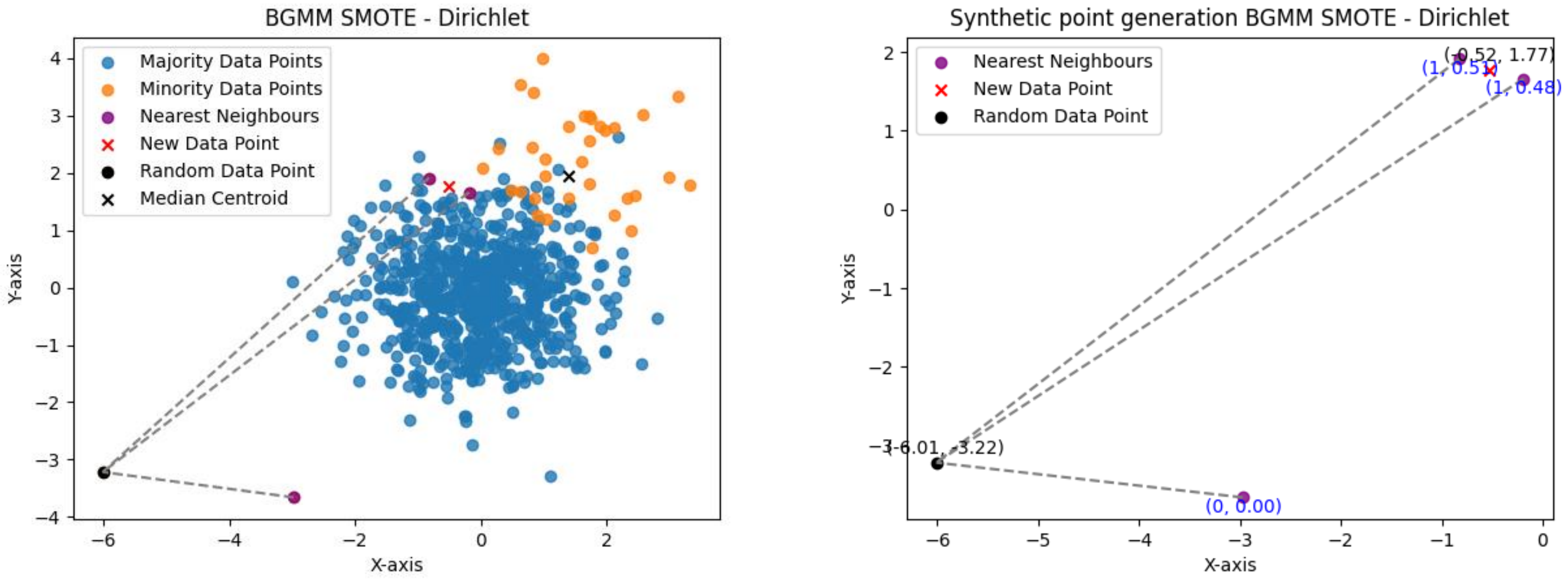
BGMM SMOTE

- A probabilistic model used for clustering
- Cluster Assignment
    1. Expectation Maximization:
        - Expectation (E-step): For each data point, the model calculates the probability of the point belonging to each cluster
        - Maximization (M-step): Update the parameters of the model by maximizing the expected log-likelihood
    2. Cluster Assignment: Probabilistically assigns data points to clusters based on the calculated probabilities.
    3. Soft Assignments: This does not definitively allocate a point to a single cluster.

BGMM SMOTE

- C_j = Cluster assignment of the J^{th} nearest neighbour
- w_j = Normalized cluster probability of the cluster which the median centroid belongs

An example of creating a sample – BGMM SMOTE



(a). This scenario occurs when an outlier is chosen as a neighbouring point.

(b). The values within parentheses indicate (d_j, w_j).

Simulation Results

- $\boldsymbol{X}_{minority-outliers} \sim \mathcal{N}(\mu_1, \Sigma_1)$
- $\boldsymbol{X}_{majority} \sim \mathcal{N}(\mu_2, \Sigma_2)$
- $\boldsymbol{X}_{outliers} \sim Uniform(-10,10)$

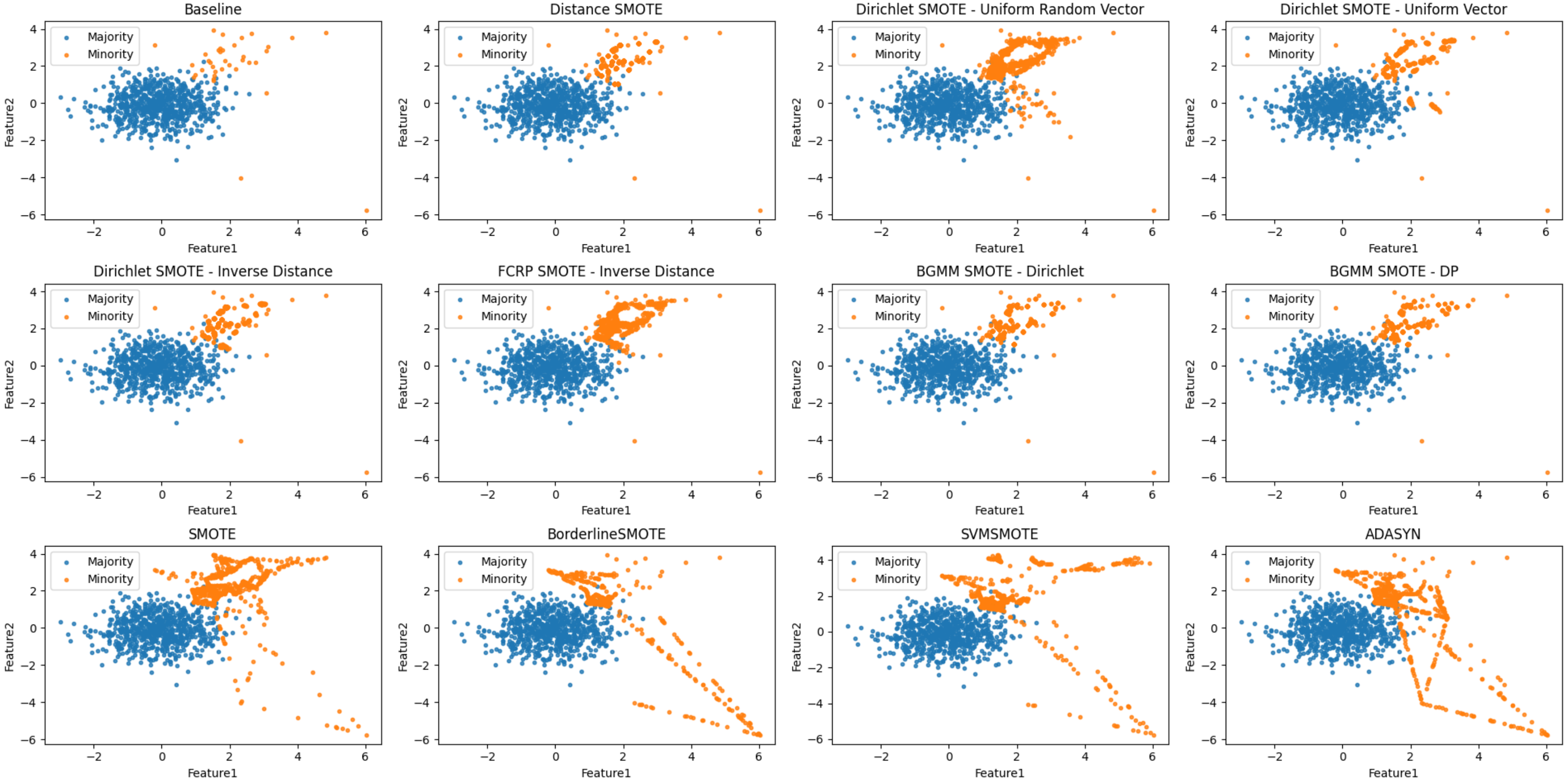


Figure: Comparison of resampled data

Simulation Results

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

$$= \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

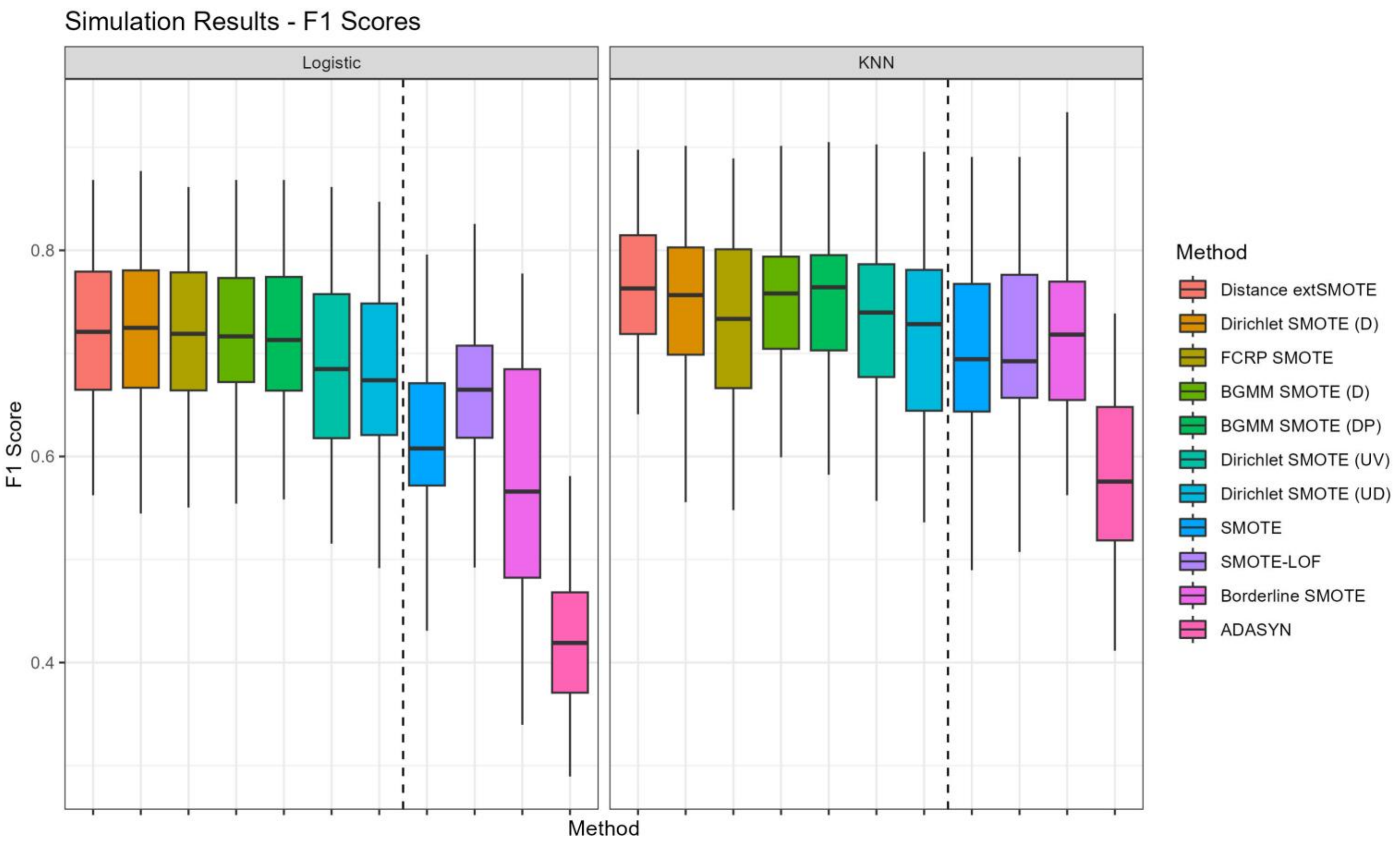


Figure: F1 Scores for 100 simulated datasets with 5-fold cross-validation.

Application Results

- Used 11 imbalanced datasets from the UCI repository

	Name	Target	Ratio	#S	#F
1	mammographic_masses	malignant	2.2:1	961	5
2	Breast_cancer	malignant	2.7:1	569	30
3	Diabetes	Diagnosis: yes	2.9:1	768	8
4	Ecoli	imU	8.6:1	336	7
5	Spectrometer	>=44	11:1	531	93
6	Isolet	A, B	12:1	7797	617
7	Car_eval_34	Good, v good	12:1	1728	21
8	Us_crime	>0.65	12:1	1994	100
9	Thyroid_sick	Sick	15:1	3772	52
10	Oil	Minority	22:1	937	49
11	Abalone19	Age 19	130.5:1	4177	8

Application Results

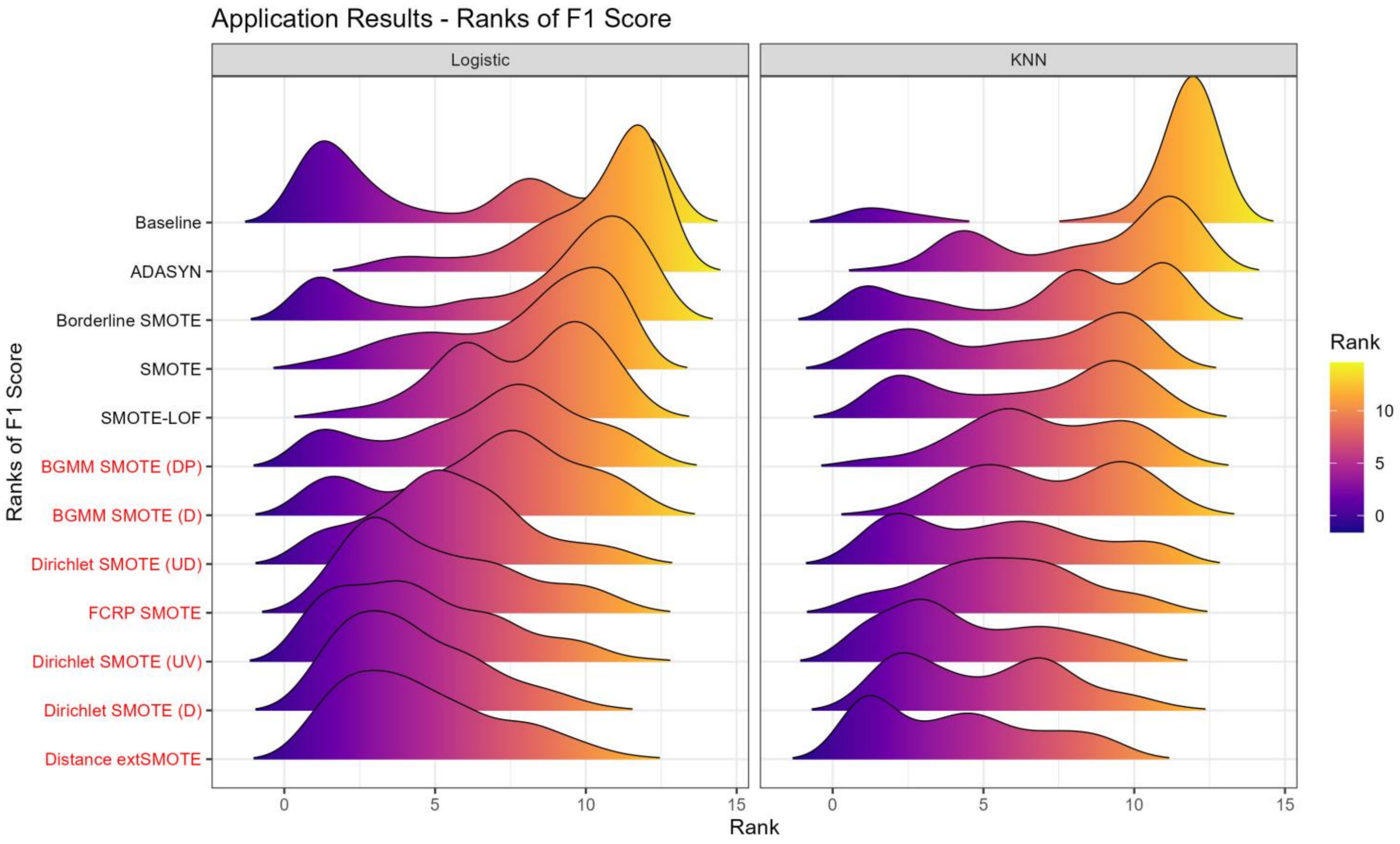


Figure: F1 Score Ranks for the datasets with 30 x 5-fold cross-validation.

Conclusion

- Class imbalance is a significant problem in classification.
- Novel methods are advancing imbalanced classification within machine learning.
- Effectively incorporate measures to minimize outlier effects.
- Creating more **accurate and reliable predictive models.**
- Across diverse domains, including fraud detection, medical diagnosis, and churn prediction, where imbalanced datasets with outliers are prevalent.
- The manuscript related to this work is in review.

References

1. Blei, D. M. and Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. Bayesian analysis, 1(1).
2. Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. The Journal of Artificial Intelligence Research, 16:321–357.
3. Hawkins, D. (1980). Identification of Outliers. Monographs on Applied Probability and Statistics. Springer Netherlands, Dordrecht.
4. Newman, C. B. D. and Merz, C. (1998). UCI repository of machine learning databases
5. Roberts, S., Husmeier, D., Rezek, I., and Penny, W. (1998). Bayesian approaches to Gaussian mixture modelling. IEEE Transactions on Pattern Analysis and machine intelligence, 20(11):1133–1142.
6. Matharaarachchi, S., Domaratzki, M., and Muthukumarana, S. (2021). Assessing feature selection method performance with class imbalance data. Machine learning with applications, 6:100170

Acknowledgement

- I would like to express my special thanks of gratitude to,
  - My supervisors, Dr. Saman Muthukumarana and Dr. Mike Domaratzki, for their excellent guidance.
  - The Department of Statistics and the staff for funding and resources.
  - My family and friends for their continuous support.

Surani Matharaarachchi, UoM

Thank You!

matharas@myumanitoba.ca