The SMICT algorithm for enhancing fairness in Dynamic Datasets

Research Project under the topic of Dynamic Algorithmic Fairness.
Bogdan Badale - <B.Badale@student.tudelft.nl>

Responsible Professor And Supervisor:
Anna Lukina - <A.Lukina@tudelft.nl>

1 - Introduction

Background

- The increasing need for fairness-aware programming [1]
- SMOTE used to increase fairness [2]

Research Gap

- Work mainly focused on detection of unfairness rather than dynamic correction.
- Very little research on SMOTE for dynamic fairness, and even less for it's variants.

Proposed Solution - SMICT - Synthetic MInority Cross-sampling Technique

- SMOTE supplemented by samples from other classes.

Research Question: Can SMICT be used to increase fairness in dynamic datasets?

2 - What is SMOTE?

Synthetic Minority Oversampling Technique [4]

- Frequently used alongside **Machine Learning** algorithms to increase the accuracy of predictions for a minority class.
- Creates **Synthetic data points** between existing data points rather than adding weights or duplicating data.
- **Nearest Neighbors** - For every element in the minority class, distances to every other element are calculated. Synthetic samples are generated between neighboring points

3 - Methodology

- **Implement** SMOTE and SMICT for the chosen "Folktables" dataset [3].
- **Train** simple Logistic Regression Algorithm on the modified data.
- **Test** on Unmodified Data.
- **Compare** Performance and Fairness evaluation: *Accuracy, Equal Opportunity, Demographic Parity*
- **Evaluate** the performance of SMICT compared to SMOTE and the no-modification baseline.

4 - The SMICT algorithm

Synthetic Minority Cross-Sampling Technique

- Oversamples Minority class by interpolating features with those of members of all other classes. **Cross-Samples** are less Prone to underrepresentation bias in the minority class.
- Increased focus on **Fairness**, minority class features become more similar to those of majority classes.
- **Dynamic** - Unlike SMOTE, SMICT uses random choice rather than Nearest Neighbors, significantly reducing the runtime.

Ideal Datasets for SMICT:

- SMICT, in theory, performs best when the **True Distributions** of classes can be assumed to have at least some **overlap** (Figure 1)

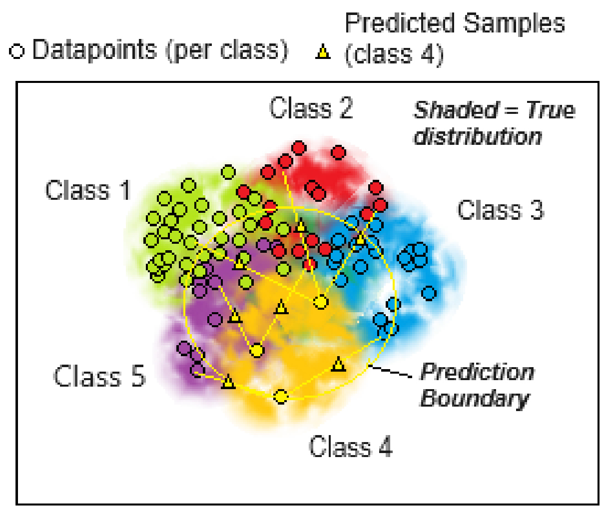

Figure 1: A visualization of SMICT for a dataset with heavy overlapping true distributions and imbalanced class sizes.

5 - Experimentation and Results

- **Metrics Used - Calculated from a confusion Matrix (Figure 2):**
  - **Accuracy: (TP + TN) / (TP + FP + TN + FN)**
  - **Equality Of Opportunity: Equalized True Positive Rate (TP/TP+FN)**
  - **Demographic Parity: Equalized Positive Prediction Rate ((TP+FP) / (TP+FP+TN+FN))**
- **EQ Opportunity and Dem Parity are measured as error rates. - The lower the better.**

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Figure 2: Confusion Matrix

SMICT and SMOTE were run on 102 total data subsets from the Employment Dataset. This comprises US census data for the years 2017, 2018. - Labeled true/false based on whether a person was employed at the time. This data contained 9 classes with 16 features each.

- **Baseline Average (No Oversampling)**
  - **Accuracy: 0.76958**
  - **MSE EQ-Opp: 0.0347**
  - **MSE Dem Parity: 0.017**

The logistic regression algorithm was used on unmodified data first. All following data displays the difference to the baseline average

- **Average Accuracy Increase**
  - **SMOTE: -0.00103 (0.1% lower accuracy)**
  - **SMICT: -0.0058 (0.6% lower accuracy)**

For this dataset, applying both SMICT and SMOTE resulted in marginally lower accuracy.

- **Average Time Taken (Seconds)**
  - **SMOTE: 107.71888**
  - **SMICT: 0.543988**
  - **Highest difference: 2197.518**

When running the experiments, SMOTE ended up being the main bottleneck, particularly for the larger data subsets.

- **Average Dem Parity Error Increase**
  - **SMOTE: 0.00048 (Increased fairness error)**
  - **SMICT: -0.00051 (Decreased Fairness error**

Again, SMICT outperformed SMOTE on average, with a lower Demographic Parity error

- **Average EQ Opportunity Error Increase**
  - **SMOTE: 0.00040 (Increased fairness error)**
  - **SMICT: -0.00160 (Decreased fairness error)**

SMICT performed better than SMOTE and overall on average, increased Equality of Opportunity fairness.

Analysis - For this dataset, SMICT, on average performed worse for accuracy, but better for Equality of Opportunity and Demographic Parity than SMOTE. It also did this a lot faster.

- Notably, this is an average. SMICT has also increased accuracy in **39/102** instances. In **11/102** data subsets, SMICT outperformed SMOTE in ALL categories.
- Accuracy, EQOpportunity, and Dem Parity performance can differ from dataset to dataset, based on the underlying distribution

6 - Conclusions

Research Question

- SMICT can be used to increase fairness, as shown in the experiments.
- Accuracy of SMICT as well as performance is dependent on the underlying distribution of the data. (In this case accuracy was lowered)
- Runtime cost is minimal, allowing it to run in a dynamic setting.

Future Work

- Improvements upon SMICT, more evaluation on more varied datasets. Analysis of the variance of SMICT.
- SMICT could be a start towards more research on active dynamic fairness balancing measures. As well as other ideas for transferring static Machine learning balancing solutions to a dynamic fairness context. (Such as Tomek links for example)

References

[1] Albarghouthi, A., Vinitsky, S., University of Wisconsin–Madison, & University of Wisconsin–Madison. (2019). Fairness-Aware programming. In Conference on Fairness, Accountability, and Transparency (p. 9) [Conference-proceeding]. https://pages.cs.wisc.edu/~aws/papers/fat19.pdf

[2] Lucentia, & De Alicante Departamento De Lenguajes Y Sistemas Informáticos, U. (2022, April 25). A Methodology based on Rebalancing Techniques to measure and improve Fairness in Artificial Intelligence algorithms. https://rua.ua.es/dspace/handle/10045/123225

[3] Ding, F., Hardt, M., Miller, J., & Schmidt, L. (2021, August 10). Retiring Adult: New datasets for fair machine Learning. arXiv.org. https://arxiv.org/abs/2108.04884

[4] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357. https://doi.org/10.1613/jair.953