The SMICT algorithm for enhancing fairness in Dynamic Datasets

Research Project under the topic of Dynamic Algorithmic Fairness.

Bogdan Badale

Supervisor(s): Anna Lukina

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 8, 2023

Name of the student: Bogdan Badale
Final project course: CSE3000 Research Project
Thesis committee: Anna Lukina, Cristoph Lofi

Abstract

As machine learning algorithms become more and more prevalent, so do the inherent risks of unfair classification of disadvantaged or underrepresented groups. Additionally, in a dynamic context, the underlying distributions can shift over time, so corrective measures that can work in a static context may end up being detrimental in the long run. In this paper, we propose a new algorithm with the aim to improve fairness in datasets, by modifying the commonly used SMOTE algorithm in a way to work better in a dynamic context, with an added focus on fairness criteria. The results in this paper indicate that this modification, labelled as the SMICT algorithm, can be a promising approach to improving fairness, albeit with limitations and challenges that need to be considered whenever the algorithm is used.

1. Introduction

Over the past years, the quantity of data has rapidly increased, and with it, there has also been a rise in automated decision-making algorithms to process that data. Increasingly, these algorithms, with their rising scale and complexity, have become difficult to manually monitor for fairness and discriminatory practices, making the automation of these tasks vital to ensure the long-term wellbeing of those involved.

Fairness however is not static. The distribution and features of classes can shift over time. As such, implementing fairness counterbalancing in a way that works in a static context, may prove ineffective or even harmful in a dynamic context (D'Amour et al. 2020)(Liu et al. 2018).

Nevertheless, more traditional machine learning techniques can still have a positive impact on fairness. In a recent work by Zhou et Al. (2023b) the performance of the popular SMOTE algorithm (Chawla et al. 2002) for the purpose of fairness is evaluated. The SMOTE algorithm was originally designed to decrease numerical class imbalance in a dataset, by generating samples for under-represented classes in a synthetic way, by making new samples that are similar, but not identical to previously existing samples of that class. In the work by Zhou et Al, the SMOTE algorithm is analyzed both theoretically and empirically in the context of fairness, showing that both theoretically and empirically, SMOTE is a promising approach to fairness in AI models, while avoiding a significant loss in prediction accuracy.

In this paper we present a new SMOTE variant called SMICT – Synthetic Minority Cross-Sampling Technique. This variant is most similar to Borderline-SMOTE (Han et al., 2005), which generates samples for a minority class by also considering the samples from the majority class that are closest to samples in the minority class (on the border). SMICT also utilizes samples from external classes (cross-samples), with a few key differences.

Throughout this research, we have found very few works that consider oversampling techniques like SMOTE and even fewer considering Borderline SMOTE in the context of fairness. This might be because of the unsuitability of these algorithms for large datasets, as both techniques are algorithmically expensive to run due to the cost of the Nearest Neighbors operation, as will be explained in **Section 2.**

Monitoring fairness in machine learning is important, as the complexity of algorithms increases, and the impact of those algorithms becomes difficult to predict manually. Work like that of Albarghouthi et al. (2019) focuses on defining checks that alert a programmer whenever a fairness criterion is breached. This is a valid approach to long term fairness monitoring, however we felt that, while it could detect the underlying unfairness of a classification, it took no steps towards fixing it in a dynamic manner. As such, we started thinking of a way to dynamically reduce unfairness, rather than detecting it, and as such, we landed upon the SMOTE algorithm, and by extension designed the SMICT algorithm.

In our research, we evaluate whether the SMICT algorithm can increase fairness in automated decision making in a more algorithmically cost-effective way than SMOTE or Borderline-SMOTE.

The results of this evaluation indicate that, in the right context, SMICT can indeed increase fairness in a more time-efficient manner. The accuracy and fairness of the predictions however do fluctuate, having cases when accuracy or fairness can be either higher or lower as a result of applying SMICT, on average performing slightly worse in terms of accuracy, but better in all fairness categories. Thus, SMICT, while not a perfect solution, can be a good step forward for dynamic algorithmic fairness.

In **Section 2** we discuss the preliminaries of the paper and relevant related work on the SMOTE algorithm in **Section 3**. **Section 4** then presents the SMICT algorithm and its formal definition, followed by **Section 5,** which discusses the contributions of this paper to dynamic fairness methodology. In **Section 6** we present an overview of the experimental setup and results, and finally in **Section 6, 7, 8,** we present our findings and wrap the research up with our analysis and conclusions on the topic as well as future research and work that can be done.

2. Preliminaries

To better understand the SMICT algorithm, it is important to first understand the SMOTE algorithm, as well as the Borderline-SMOTE algorithm.

The SMOTE algorithm by Chawla et al. (2002), operates as follows: First, it calculates the nearest neighbors for each member of the minority class. This implies comparing every member of the class to every other member using a predefined distance metric. Then, given an oversampling ratio O, and a value K, the algorithm uses, for each sample, the K nearest neighbors to generate new synthetic samples, by generating a new datapoint on the line between a chosen neighbor and the specified sample. This is done until the ratio between original and synthetic samples corresponds to the value O.

For Borderline-SMOTE (Han et al., 2005), the process uses only the edge values of the minority class. For a given distance metric, the datapoints in the minority class that are furthest from the center of the class are selected, then the nearest neighbor operation is performed between the selected samples and the elements of the negative class (i.e. those that do not belong to the minority class). The members of the negative class that are closest to the border are then used to generate new synthetic samples in combination with the edge values of the minority class. While this is initially defined only for a two class case, it can be trivially expanded to a multi-class algorithm, by defining the "negative class" as all datapoints that are not part of the minority class we are considering.

In this paper, we also use a simple Logistic Regression classifier (LaValley, M. P. 2008) offered by Sklearn in python (*LogisticRegression*, n.d.). This model is trained on the data modified by SMICT, as well as SMOTE for comparison, and aim to predict the labels of the unmodified data. The accuracy of this serves to represent the utility of the oversampling algorithm.

The performance of machine learning algorithms is typically evaluated with the use of a confusion matrix (Figure 1):

Table 1: Confusion Matrix

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

This confusion matrix determines how many samples were correctly classified, i.e. if the predicted label corresponds to the actual label.

For all instances where the Logistic Regression classifier is used, the predictive rates for each class in the data set will be calculated, then further used to calculate, then compare Accuracy and fairness between classes and further between different subsets of the data.

In this paper, the commonly used definition for predictive accuracy will be used: Accuracy = (TP + TN) / (TP + FP + TN + FN), to evaluate the performance of SMICT as an oversampling algorithm. We chose to use this simpler definition of accuracy over alternatives like the F-Score, as it also takes into account the True Negative rate.

To evaluate fairness, two metrics are used. The first is the Equal Opportunity Measure, which aims to equalize the True Positive Rate (TP/ TP+FN) amongst all classes.

The second measure that is Demographic Parity, (also known as Statistical Parity) which can be satisfied by equalizing, for every class, the rate at which a positive prediction is made ((TP+FP) / (TP+FP+TN+FN))

To evaluate Equality of Opportunity and Demographic parity we took the Mean Squared Error of both. For Equality of Opportunity, the difference between each TPR (true positive rate) and the Mean TPR for all classes was taken, and squared, then we took the average squared error for all classes as an indicator of how different each true positive rate is from the dataset average, and hence how Unequal the true positive rates of the dataset are. The same is done for Demographic parity, as both measures aim to equalize a metric between all classes, with the Error rate in the MSE technique, indicating the unfairness in the dataset.

The open-source Folktables database was used as a baseline, (Ding et al., 2021). The Folktables database was chosen as it is a more modern remade version of the original 1994 UCI Adult dataset that has been used for many works on machine learning algorithms due to its inherent class imbalance. Folktables provides several in-built prediction tasks for income, employment, health, transportation, and housing. For the scope of this research, we will be focusing primarily on employment status.

The final conclusions regarding the SMICT algorithm are reached while taking into account Equal Opportunity, Demographic Parity, as well as the resulting Accuracy of the algorithm, when compared to the equivalent evaluation of SMOTE, as well as the baseline Fairness and accuracy measures of the Logistic Regression model on the unaltered dataset.

3. Related Work

As mentioned before, there are not many articles that directly evaluate SMOTE or borderline-SMOTE in the context of fairness.

The previously referenced work by Zhou et Al. (2023b) goes in depth into the application of SMOTE and provides both a theoretical and empirical justification for the success of SMOTE in fairness balancing. The main critique of the authors was that, while there are works that use SMOTE and empirically prove its effectiveness, little study was previously made looking into the theoretical aspects of why exactly SMOTE is effective.

SMOTE however is not equally effective for all datasets. For instance, in a separate study, (Lucentia & De Alicante Departamento De Lenguajes Y Sistemas Informáticos, 2022) SMOTE performed worse than oversampling and under sampling, and the best performing technique for that dataset was to remove the sensitive attribute entirely. This could imply that, depending on the original bias and shape of the data, smote can end up accentuating the bias in the original distribution, causing a decrease in fairness.

Lastly, another relevant article by Sha et Al. (2022) discusses the effectiveness of class-balancing techniques on both predictive accuracy and fairness in the context of education. It compares 4 under sampling, 4 oversampling and 3 hybrid techniques of class balancing, amongst which borderline-SMOTE is also evaluated. However, as this work is more general and is not focused on a single algorithm or type of algorithm, there is no direct evaluation or consideration given to borderline-SMOTE. In their work, they conclude that there does not necessarily need to be a trade-off between accuracy and fairness, and that, in some cases, increasing fairness directly leads to increased accuracy.

4. SMICT – Problem Definition

As an oversampling technique, SMICT aims to solve several issues. The first being that of a low variety of base samples in the minority class – leading to a flawed prediction of the class true distribution. SMICT solves this by sampling from other classes.

The second issue is that of the function cost. In Borderline-SMOTE, the nearest-neighbors from other classes are used. For large datasets, this operation is very computationally expensive, something that should be minimized in a dynamic context. SMICT solves this through random choice, in both the minority and majority class.

Lastly, SMICT also is designed to have a greater focus on fairness. By using samples from existing majority classes for minority class predictions, a classifier would learn to treat a minority class more similarly to a majority class, reducing class prediction imbalance.

In a multi-class setting, SMICT interpolates existing minority class samples with samples from all the classes that are more heavily represented, in proportion to their relative size.
As such, take the following assumptions:

Assumption A*: For a given dataset A, and a known set of classes C in A with their true distributions in C', for any class c' in C' there exists at least one element in each of the other classes in C' which are similar to at least one element in class c'.*

Assumption B: *The more represented a class c is in the dataset, the closer class c is to the true distribution c'. For any class c with true distribution c' and n samples, if for any class f with true distribution f' and m samples, m > n then f' – f <= c' – c.*

Intuitively Assumption A would imply that there are no classes in the dataset with true distributions that are entirely disjoint from one another. This assumption is difficult to verify or enforce, as it relies on the true distribution of the classes. However, it can still prove useful when deciding whether to use SMICT or not if one has knowledge of the dataset and believes the assumption could reasonably hold.

Assumption B requires you to be able to trust the source of your data and that all new samples that are added are added to a class do in fact belong to that class. Adding new real samples should never increase the difference to the true distribution.

If both assumptions hold, then we can make certain guarantees about the performance of SMICT as a predictor algorithm, as, it would mean that at least one element in each other class can be chosen randomly and perform at least as well as a member of the minority class when predicting samples for the minority class (Assumption A). Additionally, following Assumption B, the classes with highest accuracy with respect to their true distributions should be prioritized for sampling, hence the procedure of taking samples from other classes based on class size difference, with the largest classes generating more samples.

While the minimal fulfilment of these assumptions provides only a minimal guarantee of accuracy, it still provides a boundary beyond which the SMICT algorithm is applicable. The more heavily Assumption A is met (the more the classes are inter-connected) the better SMICT will perform, however, SMICT does not need 100% overlap to offer a better prediction, or to perform better than SMOTE for a given dataset.

Finally, a theoretical example of a dataset for which SMICT would be optimal, would be a dataset with heavily overlapping true distributions for the various classes, for which, however, the known distributions are numerically heavily imbalanced, leading to improper classification of underrepresented classes.

5. Contributions – Methodology

In this paper, we present and evaluate a new variant of SMOTE, called SMICT for the purpose of algorithmic fairness in a dynamic context. We outline the theoretical advantages as well as evaluate empirical results.

This is done by comparing the implemented SMICT algorithm to a simple implementation of SMOTE on the open source "Folktables" Database (Ding et al., 2021), using an existing Logistic Regression algorithm from Sklearn (*LogisticRegression*, n.d.). We compare the predictive accuracy when oversampling with SMOTE and SMICT respectively, as well as contrasting those results with the baseline accuracy of the model on the dataset without any oversampling.

We also compare the resulting fairness of the predictions, using the metrics of Equality of Opportunity and Demographic Parity as defined in **Section 2.**

From the Folktables Database, the employment subsection of the database was chosen under the hypothesis that it would be able to satisfy Assumption A of the SMICT algorithm as defined in **Section 4** of having some overlap between the true distributions of each class. As the status of whether someone is employed or not should not be unequal to the extent that there exists absolutely no overlap between different groups.

The findings of our research can be summarized as follows:

Out of the 102 aggregated statistical performance measurements, SMICT on average resulted in an increase of both Equality of Opportunity (55% of the time) and Demographic Parity (47% of the time but larger increase) when compared to both SMOTE and the classification on the dataset without modifications. SMICT however has overall lowered the accuracy of the algorithm in the majority of data instances (Increased accuracy in only 39% of instances). There were however instances, about 11%, where SMICT performed the best in all categories: Accuracy, EQOpportunity and DEMParity.

6. Experimental Setup and Results.

6.1 – Experimental setup

From the Folktables database, the Employment data was used to test the ability of the SMICT algorithm. Elements in this dataset used 16 features and 9 different classes. Each datapoint was also labelled either true, or false, depending on whether the person was employed at the time.

SMOTE was modified slightly in implementation to account for this structure. To generate both true and false instances for employment synthetically, the algorithm was run twice, once with only existing true samples, to generate true samples, and once with only existing false samples to generate false samples. Additionally, due to the size of some of the datasets, SMOTE had to be limited to only be performed on minority classes rather than ALL classes, as performing the Nearest Neighbors operation on 20,000+ would not be realistic for testing, Additionally, for all tests, an oversampling ratio of 6 was used for SMOTE.

SMICT similarly was divided into true and false generated instances, however, due to the innate random choice, it was able to be performed on all classes without any additional time investment. Additionally, SMICT generates samples based on the difference in class size between the target class and the largest class, and as such does not use an oversampling ratio, and, similarly to the implemented SMOTE, affected minority classes a lot more than majority classes.

A random seed was used for all operations involving randomness, and the implementation of the SMICT algorithm within the Folktables database framework can be found at *https://gitlab.ewi.tudelft.nl/cse3000/2023-2024-q4/Lukina/bbadale-Dynamic-Algorithmic-Fairness-in-Machine-Learning*. That is also where you can find all the results and statistics generated by the SMICT algorithm evaluation mentioned in the following section **6.2**.

Of note, of the 9 classes one or two of them do not appear in every single state. Only the classes present in each state are considered for evaluation.

6.2 – Results

The two algorithms SMOTE and SMICT were run over a total of 102 data subsets, once for each of the 50 US States (and the District of Columbia) over the years 2017 and 2018 respectively.

For each state and year, the confusion matrix for each demographic group was calculated as described in **Section 2.** The resulting matrix was then used to calculate the Accuracy for each group, and hence the accuracy of the data subset, as well as the overall Eqality of Opportunity and Demographic parity between the groups for each state.

The average dataset accuracy, Equality of Opportunity and Demographic Parity is then calculated for all 102 data samples.

- **Baseline Average(No Oversampling) (1)**
  - **Accuracy: 0.76958**
  - **MSE EQ-Opp: 0.0347**
  - **MSE Dem Parity: 0.017**

Before being able to properly compare SMICT and SMOTE, it is important to establish what the baseline accuracy and fairness of the predictive model is when the data is unmodified **(1)**.

The accuracy is measured as a percentage, in this case 76.9% accuracy baseline.

For equal opportunity and demographic parity, the measure indicates the mean square error in each fairness metric respectively. As such, the lower the better, wherein an MSE of 0 would imply 100% fairness.

- **Average ACCURACY Difference (2)**
  - **SMOTE: -0.00103**
  - **SMICT: -0.0058**

On average, both SMOTE and SMICT had lower accuracy than the baseline (as shown in statistic **(2)**). SMOTE being 0.1% less accurate, and SMICT being 0.6% less accurate.

- **Average MSE EQ Opp Difference (3)**
  - **SMOTE: 0.00040**
  - **SMICT: -0.00160**

For Equal Opportunity **(3)**, SMOTE decreased fairness on average, indicated by a positive value (0.00040) difference between the average SMOTE equal opportunity and the baseline equal opportunity.

SMICT performed significantly better than SMOTE, as well as overall increasing Equality Of Opportunity. The negative value (-0.0016) indicating a decrease in the Equality of Opportunity Mean Square Error compared to the baseline.

- **Average MSE DP Difference (4)**
  - **SMOTE: 0.00048**
  - **SMICT: -0.00051**

Similarly, for Demographic Parity **(4)**, the error, compared to the baseline was higher for SMOTE, whereas the fairness error for SMICT decreased.

- **Average Time Taken (Seconds) (5)**
  - **SMOTE: 107.7188s**
  - **SMICT: 0.54398s**
  - **Highest difference: 2197.51s**

Importantly, throughout the testing, it became clear that SMOTE heavily bottlenecked the runtime. In some of the more populous states, (specifically Texas and California), the difference was most apparent, the highest difference being in California 2017 where SMOTE was 36 minutes slower than SMICT.

Additionally, as SMICT does not use the Nearest Neighbors operation, it was able to keep a consistently low runtime at an average of 0.544 seconds. With the highest value being California 2017 with 2.6 seconds runtime compared to the 2200 seconds taken for SMOTE.

Incidentally in that example, SMICT also performed better than SMOTE in all fairness categories, with SMOTE having slightly higher accuracy.

- **California 2017 – (Green marks best performance) (7)**

	Baseline	SMOTE	SMICT
Accuracy	**0.759426**	**0.760070**	**0.759907**
MSE EQ-OP	**0.001835**	**0.002005**	**0.001753**
MSE DP	**0.006650**	**0.007030**	**0.006798**

For this example **(7)**, SMICT performed better than the baseline in accuracy, and the best for Equal Opportunity. For Demographic Parity, it was only a slight decrease compared to the baseline. SMOTE had the highest accuracy, but the worst fairness metric, and, as mentioned before, took almost a thousand times more time than SMICT to run.

- **SMICT outperformed SMOTE (8)**
    - **Accuracy: 35/102 Files**
    - **EQ-Opp: 50/102 Files**
    - **DP: 52/102 Files**

Significantly, while SMICT outperformed SMOTE in both Equal Opportunity and Demographic parity on average, SMICT performed better in slightly fewer instances than SMOTE for Equal Opportunity, but overall, the numerical increase was more significant for SMICT resulting in a higher overall average as shown in the **(3)** statistic earlier this section.

- **SMICT outperformed No modification (9)**
    - **Accuracy: 39/102**
    - **EQ-Opp: 55/102**
    - **DP: 47/102**

Lastly, we can see that compared to the baseline, SMICT had a positive impact in accuracy approximately 39% of the time, and in more than half of the data instances, the Equal Opportunity was increased. Demographic Parity was increased in slightly fewer instances, but more significantly, as indicated by statistic **4**.

In 11/102 files, SMICT performed better across every category: accuracy, Equality of Opportunity and Demographic parity.
More specifically those instances are the following: (10)

- **2017:**
    - **Kentucky – Massachusetts – Michigan – Oklahoma – South Carolina - Utah**
- **2018:**
    - **Texas – Iowa – Maryland – Missouri – South Carolina**

Figure 1 is a visualization of the SMICT algorithm prediction for Maryland 2018. After performing a basic Principle Component Analysis to condense the 16 features of the datapoints into two principle components. The proximity of the Real Points (Green) to the Synthetic Points (Red) can be a measure of how accurately the synthetic data captures the distribution of the class.

However, this does not necessarily translate to predictive accuracy, as a predicted distribution with a different shape from the original can still, when trained on by a classifier, can still perform accurate predictions for new data. For instance, in **Figure 1**, numerically, SMICT was more accurate only for classes 5, 7, 8.

Class 8 appears to have almost full overlap between predicted and actual datapoints, whereas class 5 has had the highest increase in accuracy, 72% -> 77% despite the predicted distribution having a completely different shape.

For all 102 instances of data, a similar plot was generated for both SMOTE and SMICT, to see whether there is a visible trend between the predictive strategy and the accuracy of the classification.

6.3 – SMICT as a predicter

Overall, in terms of Accuracy, SMICT performed worse than predicted given the dataset that was used. While there were instances where SMICT performed better than SMOTE, and instances in which SMICT improved accuracy significantly, overall running the SMICT algorithm has decreased the accuracy of the predictions.

An interesting trend in the data that we weren't quite able to capture visually, shows that, even for datasets where SMICT is on average more accurate, most of the classes are individually less accurately predicted, with one or two exceptions balancing out the predictions.

In practice, using cross-samples to generate minority samples tends to create quite similar predictions for every class. As seen in **Figure 1**, this is sometimes beneficial, corresponding to the actual data, but this "averaging" almost of the data points is more often detrimental in terms of accuracy for other distributions. This "averaging" effect was not as noticeable for classes that were already large, as they required less data samples to be generated (as seen in class 1 and 2 of **Figure 1**)

6.4 – SMICT for fairness

As a fairness balancing measure, SMICT overall has shown positive results. Increasing Equality of Opportunity in over half of the analyzed distributions. Despite only increasing Demographic Parity in 47/102 class distributions, Demographic Parity was more noticeable when it was increased than when it was decreased.

This increase in fairness can likely be attributed to the use of cross-samples, as classes that share a lot of similarities are likely to be classified in an equal way more often.

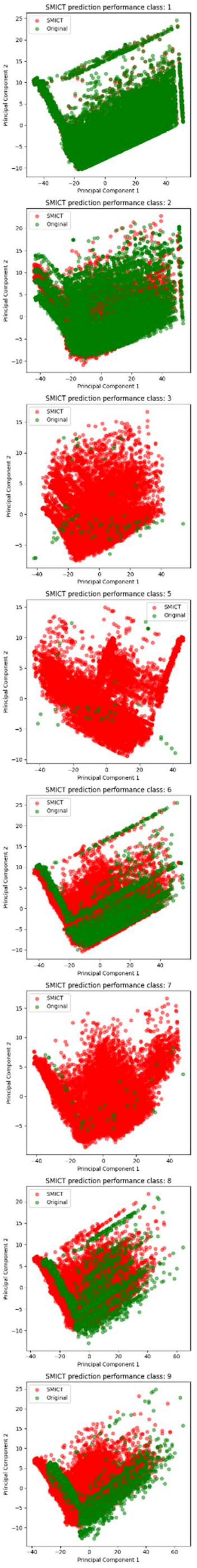

Figure 1- Maryland 2018 - SMICT performance PCA Per Class

In regards to dynamic fairness, the SMICT algorithm in many ways is a static fairness measure, however, the low runtime and processing requirements compared to SMOTE allows SMICT to also be effective in a dynamic environment, as, regardless of the size of the data, it can be run at any time to increase fairness over the currently observed distribution.

7. Discussion

7.1 – Reflection on the research approach.

Overall, the Folktables database, and particularly the subset used for these experiments, provided a good indication of the capabilities of SMICT as both a predicter algorithm and a Fairness measure. The belief that it satisfied the assumptions laid out in **Section 4** was to some extent proven by the performance of SMICT which, while not optimal, was overall beneficial in terms of fairness.

However, the database proved to not be fully suited for the SMOTE algorithm. In our comparison to SMICT, SMOTE ended up performing a lot worse than expected for the given dataset, while also doing so a lot slower, bottlenecking the experiment.

Regarding the accuracy of the predictions, while SMICT did on average lower accuracy, the increase overall increase in fairness is still relevant and significant. The assumptions laid out in **Section 4** only provided minimal guarantees towards the accuracy of SMICT, and as such, for a database more suited to the use of SMICT, we can safely assume that the accuracy will also be increased alongside the fairness criteria.

7.2 – Limitations

SMICT is very much dependent on the type of data it is used on. Throughout my testing, I have also performed an analysis on a dataset that could reasonably be assumed to **not** satisfy the class overlap assumption laid out in **Section 4**. The Income database from Folktables has true and false values corresponding to whether a person had an income of over 500,000 that year. Intuitively, the underlying true distribution for this is much more likely to be heavily imbalanced across different societal groups.

As such, when running SMOTE and SMICT on the income database, SMICT performed worse than SMOTE and No modification in all categories (apart from time taken), whereas SMOTE, at the cost of 0.05% accuracy, was able to increase both Equal Opportunity and Demographic Parity. (These statistics can be found and generated in the same Gitlab repository)

Lastly it is worth noting that the SMICT algorithm could be heavily impacted by the random selection it uses for its operation. Further testing is required to be able to fully conclude how much of an impact random selection has on the overall performance of the algorithm.

As such, the conclusions of this work are not a conclusive evaluation of the SMICT algorithm as a technique. Nevertheless, we believe that, while not conclusive, the observations made over 102 different data distributions can still provide a good indication towards whether or not SMICT shows any promise for the purpose of dynamically improving fairness.

8. Conclusions and Future Work

8.1 Conclusion

The original research question asked whether SMICT could be used to improve fairness in a dynamic setting, and while in this case it came at the cost of some accuracy, it was able to overall perform better in both the “Equality of Opportunity” and “Demographic Parity” fairness measures, and do so with low computational time requirements, allowing SMICT to be run in a dynamic setting.

In conclusion, while the SMICT algorithm would need more testing to get a full grasp of its capabilities, the research done shows that the techniques employed by SMICT may be a promising approach towards dynamic algorithmic fairness correction in datasets.

8.2 Use Cases

SMICT is situational. It can not and should not be applied to every dataset used for machine learning. Its advantage however is that it’s easy to implement and test on existing data. As new data of the same type is often similar to older data, SMICT would continue to perform well and ensure a degree of fairness even if the accuracy of the changing distribution may decrease over time.

Additionally, while not the intended purpose, SMICT can also be used to gain insight into the true distribution imbalance of a dataset. As SMICT relies on two assumptions to ensure accuracy (outlined in **Section 4**) which both relate to the true distributions of the classes in the dataset. A decrease in the accuracy of SMICT could potentially indicate an increase in the true class imbalance.

8.3 Future Work

Currently there is little research being done towards active dynamic fairness balancing. The SMICT algorithm could be a starting point for other attempts to adapt static fairness measures to dynamic contexts.

SMICT itself can be improved and evaluated more, by running it on different and more varied datasets, and performing a more in-depth analysis of exactly what type of datasets SMICT could be used for. Additionally, more research on the impact of the random variance of SMICT on its performance would also give a more accurate evaluation of SMICT for the general use case.

9. Responsible Research

Responsible research is necessary for any academic work. In the context of this paper, the main issue addressed was that of reproducibility of results, as the methods used make use of random sampling as well as custom-built code.

The issue of randomness is solved by using random seeds, such that any researchers can, when running the same code, reproduce the same results as those displayed and referenced in the Experimental Results section. The code itself is open source at *https://gitlab.ewi.tudelft.nl/cse3000/2023-2024-q4/Lukina/bbadale-Dynamic-Algorithmic-Fairness-in-Machine-Learning*.

For transparency, it is worth noting that ChatGPT 4.0 was used to assist in writing the code used to parse the data generated and aggregate the statistics displayed in **Section 6**. It was also used to help with unfamiliar syntax and was not used at any point in the design and creation of the SMICT algorithm, which is the original work of this paper.

All analysis, conclusions and experiments shown in this paper are our own, and any use of work that belongs to others is clearly indicated and referenced in the text and the references at the end of this paper.

References

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., & Hardt, M. (2018). Delayed impact of fair machine learning. In University of California at Berkeley, *Proceedings of the 35th International Conference on Machine Learning*. https://proceedings.mlr.press/v80/liu18c/liu18c.pdf

D'Amour, A., Srinivasan, H., Atwood, J., Baljekar, P., Sculley, D., Halpern, Y., & Google Research. (2020). Fairness is not static: deeper understanding of long term fairness via simulation studies. *Conference on Fairness, Accountability, and Transparency (FAT* '20)*, 12. https://doi.org/10.1145/3351095.3372878

Chawla, N. V., Bowyer, K. W., Hall, L., & Kegelmeyer, W. P. (2002b). SMOTE: Synthetic Minority Over-sampling technique. *Journal of Artificial Intelligence Research/ the Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Zhou, Y., Kantarcıoğlu, M., & Clifton, C. (2023b). On Improving Fairness of AI Models with Synthetic Minority Oversampling Techniques. In *Society for Industrial and Applied Mathematics eBooks* (pp. 874–882). https://doi.org/10.1137/1.9781611977653.ch98

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

Han, H., Wang, W., & Mao, B. (2005). Borderline-SMOTE: a new Over-Sampling method in imbalanced Data sets learning. In *Lecture notes in computer science* (pp. 878–887). https://doi.org/10.1007/11538059_91

Sha, L., Rakovic, M., Das, A., Gasevic, D., & Chen, G. (2022). Leveraging Class Balancing Techniques to Alleviate Algorithmic Bias for Predictive Tasks in Education. *IEEE Transactions on Learning Technologies*, *15*(4), 481-492. https://doi.org/10.1109/tlt.2022.3196278

LaValley, M. P. (2008). Logistic Regression. *Circulation*, *117*(18), 2395-2399. https://doi.org/10.1161/circulationaha.106.682658

LogisticRegression. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html#sklearn.linear_model.LogisticRegression

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

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