Dynamic Algorithmic Fairness Monitoring in Machine Learning

The Effect of Ageing of Datasets in Long Term Fairness

Jorden van Schijndel[1]

Supervisor(s): Anna Lukina[1]

[1]**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 23, 2024

Name of the student: Jorden van Schijndel
Final project course: CSE3000 Research Project
Thesis committee: Anna Lukina, Christoph Lofi

Abstract

Recent scandals like the dutch Toeslagenaffaire have shown the importance of fairness monitoring of machine learning models. When not careful, automated decision making models can unfairly favor groups of people and discriminate other groups. The results can be devastating for the people involved. It has been recognised that this problem requires proper research. However, most of the already conducted research looks at the problem in a static context, while almost all the real life applications are a dynamic process. Datasets are constantly increasing, and an automated decision process can have an effect on the newer entries on this dataset. A new problem then arises, when looking at these prediction tasks in a dynamic context, is the older data just as relevant as the new entries? This question can be answered by the use of fading algorithms. Fading algorithms use different methods to prioritise new data and forget old data. This paper investigates the effect of these fading algorithms on the fairness of a model. The different methods researched are an abrupt fading algorithm, a gradual fading of weight algorithm and a gradual fading of amount of data algorithm. This research resulted in the showing of importance of looking at the data in a dynamic context, observing a significant improvement on the equality of opportunity, at the cost of the efficiency of the model.

1 Introduction

The problem of algorithmic fairness has become more significant over the last few years with the increasing use of machine learning models. It is important fairness is properly researched to prevent the potential harming of a certain group of individuals. The majority of fairness monitoring thus far has been static, but we should take into consideration that the size of most datasets is constantly increasing, and that entries can become outdated, causing a loss in accuracy in prediction tasks[13]. Other research has also shown how important it is that fairness monitoring should be looked at dynamically[10; 1; 8]. Wagner[13] has already worked on ageing and fading of datasets on a multinomial naive bayes classifiers, and in this research we will try to couple it with the work of D'Amour[4] and Calder and Verwer[3] to see the effect of these fading algorithms on the fairness.

The ageing of datasets and outdated entries has been thoroughly researched in various fields [11; 9; 12]. These papers show that data can quickly become outdated and the age of an entry should be taken into consideration when using data in prediction tasks. In this paper we find out if the Income prediction task for the Adult dataset[5] is influenced by outdated data entries.

The research question we aim to answer at in this paper is: What is the effect of ageing of datasets in long term fairness? Using baseline models from other research papers, we can test the effect of ageing and fading of datasets on the fairness. We aim to find the relevance of ageing and fading to algorithmic fairness. The sub questions identified are:

- How do you measure long term fairness?
- When is an instance no longer important?
- How do you modify a dataset to make newer instances more relevant?
- Is it feasible to modify datasets for fairness in machine learning models?

In this paper Multinomial Naïve Bayes classifiers are trained with different fading algorithms on the Adult dataset. More specifically, in order to measure the fairness as accurately as possible, the datasets from Ding[5] are used. These contain predefined prediction tasks explicitly created for the measurement of fairness. The accuracy and the equality of opportunity[7] of the different models were tested on these prediction tasks to find out if the fading algorithms have any influence on the fairness of a classifier.
The equality of opportunity is a common metric in fairness monitoring representing how a group is disadvantaged. We observed a notable improvement on the equality of opportunity when using an algorithm which takes the age of the data in consideration. This however came at a cost of a slight decrease in the accuracy of the model, and with significant increase in the time it takes to train a model.

The rest of the paper is organised as followed: In section 2 we will go over the problem definition. In section 3 we will discuss similar papers in the Related Work section. In section 4 we will explain the preliminaries for the research, followed by the main contributions of the paper in section 5. The experimental results will be shown in section 6. The ethical issues regarding the paper will be described in the Responsible Research part, namely section 7. Lastly, we have the discussion in section 8 and to finish the paper we draw the conclusions and future work in section 9.

2 Problem Definition

Papers have already been published showing the effect of fading of datasets on the accuracy of a model [13], however the influence of these changes have not yet been researched in regard to the fairness of the classifier. In this research we aim to answer the question whether these previously mentioned fading algorithms influence the long term fairness. This includes finding out whether instances are no longer important and how one should measure long term fairness.

To clarify why it is needed to look at the fairness monitoring with a dynamic approach, see Figure 1[10]. In this figure you see 2 lines, representing a machine learning model each. The orange line is a model trained on data from the same year as the test set. The blue line is a model trained on data from 2014. We can see that the model trained with more relevant data outperforms the model from 2014 data almost every year. This figure only shows the data from 2015-2019, but if we consider data from even further ago, the difference

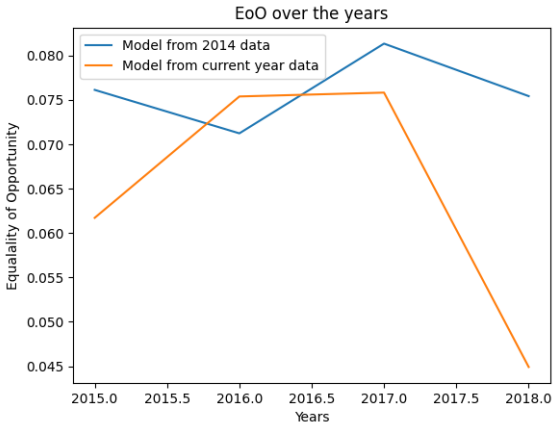

Figure 1: Importance of training model on relevant data

will become even more significant. This means that a model trained on an older dataset will become outdated, which might effect the fairness of the classifier.

3 Related Work

There are two research fields that are relevant to this paper, namely the field of dynamic machine learning fading algorithms, and the field of algorithmic fairness. In this section some of the papers researching these fields are discussed.

Starting with the algorithmic fairness field, there are two directions in this field applicable for the research. Dynamic algorithmic fairness and static algorithmic fairness. Dynamic fairness is a newer research field proven to be more in accordance to real life [4; 10; 1]. This field also includes research into which datasets to use for fairness monitoring[5]. However because there has been vastly more research regarding static machine learning, some papers researching interesting algorithms have been considered for this research, e.g. improved Naïve Bayes classifiers [3].
As a metric, the equality of opportunity[7] is used. EoO is the true positive rate of a majority group minus the true positive rate of a minority group. This shows the difference of opportunity of these groups. EoO is often used as a metric in fairness monitoring. This makes it possible for us to compare the performance of the proposed fading algorithms with other fairness-aware classifiers.

Secondly, dynamic machine learning fading algorithms. The main focus of this paper discusses the fading algorithms on a Multinomial Naïve Bayes classifier for a stream of Twitter data [13]. There is however a significant difference between a data stream of Twitter data and the Adult dataset. Thus, inspiration for the fading algorithms are taken from this paper, but modified accordingly to fit the Adult dataset. More about this paper in the preliminaries section. Another relevant paper in this field discusses multiple forgetting algorithms, including a gradual and an abrupt algorithm tested by Gama[6]. Again, these algorithms were not tested on the Adult dataset, but rather on artificially created datasets[2] with a concept drift. Both these papers focus on the accuracy of the model, while our proposed algorithms focus mainly on the fairness of the model.

4 Preliminaries

In order to understand the algorithms researched in this paper, some specific fading algorithms are explained in this section. Starting with the fading algorithm for opinionated data streams [13]. It is a modified Multinomial Bayes model which takes into temporal information into consideration when estimating the class prior $\hat{P(c)}$, see Equation 1, and the class conditional word probability $\hat{P(w_i|c)}$, see Equation 2. These equations are then used to predict which class a word belongs to.

$$\hat{P}^t(\mathbf{c}) = \frac{N_c^t \cdot e^{-\lambda \cdot (t - t_{lo}^c)}}{|S|^t} \tag{1}$$

$$\hat{P}^t(w_i \mid c) = \frac{N_{ic}^t \cdot e^{-\lambda \cdot (t - t_{lo}^{(w_i,c)})}}{\sum_{j=1}^{|V^t|} N_{jc}^t \cdot e^{-\lambda \cdot (t - t_{lo}^{(w_j,c)})}} \tag{2}$$

In Equation 1, N_c^t is the amount of documents in the data stream at timepoint t classified as class s, $|S^t|$ is the total amount of documents at timepoint t, t_{lo}^c is the timepoint from most recent observation from class c and λ is the decay rate at which the data should become less relevant.
In Equation 2, $|V^t|$ is the total amount of distinct words found in the documents, the other variables are the same as in Equation 1, with $t_{lo}^{(w_i,c)}$ meaning the timepoint of the most recent observation of word i belonging to class c.
It is important to note that this algorithm measures the amount of time it takes for a word belonging to a class to appear again. Since we use the Adult dataset, there are no identical entries and thus we utilize a different time measure. The different time measure used for this research is more in accordance with abrupt and gradual forgetting algorithms[6].
The abrupt algorithm makes use of a sliding window with a *first-in-first-out* (FIFO) method. This means that older data is completely disregarded at some point. All the data inside of the sliding window has equal weight, and the weight of all the data outside of the window is zero. The gradual forgetting algorithm on the other hand, takes all the data up to a timepoint into consideration, but prioritising newer data by giving it a higher weight than older data.

5 Measuring fairness with fading algorithms

As explained in previous sections, fading algorithms[13; 6] and fairness monitoring [3; 4; 10] have both been extensively researched. However the effect of these fading algorithms on the fairness is still not fully clear. In this section the fairness of a model with different forgetting algorithms is measured on the ACSIncome[5] prediction task. This predicts whether an individual has an income greater than $50.000 a month.

We consider three different approaches to forgetting data; Abrupt, gradual weight and gradual amount fading. The goal of the experiment is to find out which data we can discard, and which data we should prioritise. Testing these three approaches against each other and the baselines will show us exactly that. To make sure the results are unbiased, we test the algorithms with different training-test dataset splits and calculate the mean and standard deviation.

5.1 Baselines

First, as baseline, the fairness is measured with a no fading static model. A Multinomial Naïve Bayes modifier is trained on data from 2008, and then tasked with predicting the salaries of entries from 2008 to 2013. This is considered the baseline. The model is then compared to a no fading dynamic model, an MNB which continuously learns over the years from 2008 to 2013, with all years given the same weight.

5.2 Abrupt fading

Moving on to the forgetting algorithms, the first one tested is the abrupt fading algorithm. An MNB is again trained on 2008 data, and then continuously trains and tests on data from 2008 to 2013. The difference between this and the no fading dynamic algorithm is that there is a sliding window, and data outside of this window is not taken into consideration. It is important to note that the data does not have a timestamp, only the year in which it was collected. Therefore limiting the different sizes of sliding windows that can be tested. Any sliding window of size 6 or larger will display the same results as the no fading dynamic model and a sliding window of size 1 will display the same results as training and testing a model on data of one year. In this paper we only test a sliding window with size 2 because of these limitations and the difference in performance between models with sliding windows of size 2 to 6 or larger were negligible.

5.3 Gradual weight fading

Secondly, the gradual algorithm. In this model the newer data is prioritised by a factor over the older data. There are multiple ways to do this, e.g. simply scaling the weight of the samples uniformly or making use of an exponential factor $e^{-\lambda * S}$ [4]. The uniform approach is simple as the data is sorted by years, thus making it easy to increase the weight of samples every year by a constant factor c, see Equation 3. Here the constant factor c is equal to $w_i - w_{i+1}$. y_i is the current year, y_0 is the first year of which data has been obtained and n is the total amount of years of which data has been obtained. For the exponential method, the λ can be changed to discover the best decay ratio.

$$w_i = \frac{y_i - (y_0 - 1)}{n + 1} \tag{3}$$

5.4 Gradual amount fading

Finally, forgetting algorithm based on the amount of data. A model is trained with more data from newer years and less data from older years. All entries are treated with equal weight, however the model is trained with only a portion of the total amount of data. The data from a year i is randomly split by a factor of w, see Equation 3. The same equation from the uniform gradual algorithm can be used for this algorithm. That portion of the data is then used to train the model.
The average EOO and accuracy of the models are tested against each other to detect the influence of their fading algorithms on the fairness and precision of the models.

6 Experimental Results

Now that the main contribution and methodology of the paper have been explained, we can move on to the results and analyze them. The results will be summarized and analyzed in this section. For each algorithm the advantages and the disadvantages will be revealed.

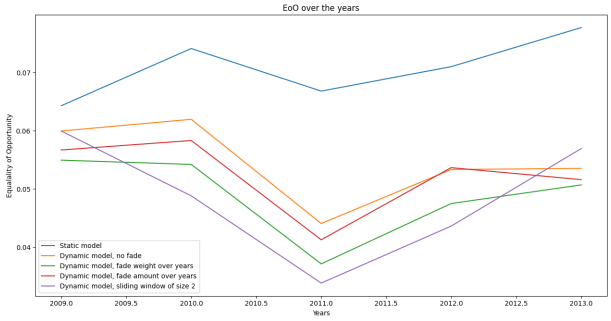

Figure 2: Comparison of EoO the fading algorithms

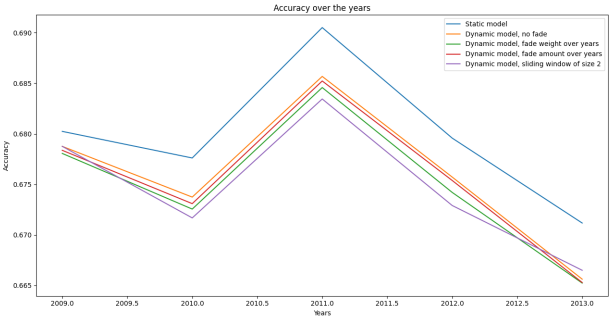

Figure 3: Comparison of accuracy the fading algorithms

Like explained before, the baseline is a static MNB model trained on data from 2008. This is the fastest algorithm as it is only trained once on data from 2008. However, the paper is based around the fact that the models should be trained continuously and not statically. This is shown again when looking at the equality of opportunity scores from this model, see Figure 2. Figures 2 and 3 show a simulation of the algorithms' performance over the years on the Adult dataset. The lines represent the different algorithms and the values on the y-axis show the equality of opportunity and the accuracy the algorithm obtains on the test data from that year.
We can clearly see that the static baseline is significantly

outperformed by all the dynamic approaches on the equality of opportunity. On the other hand, the accuracy of the model is consistently slightly above the fading algorithms, see Figure 3. This paper focuses mainly on the fairness properties, thus this slight accuracy decrease is deemed affordable.

We also observe a trend in the data, with all the algorithms performing the best on data from 2011. This likely means that the data from 2011 consists of less outliers, causing all algorithms to perform the best on this year. This trend is not only observable in 2011, we can see that all lines have a similar shape through 2009 to 2013. This shape just shows that some years are more similar to the data from previous years than others. This trend is not relevant for the research, as we focus on the performance of the different fading algorithms. The comparison of the lines against each other is what we should analyze.

Moving on to the dynamic algorithms. In this single simulation we see that the sliding window performs exceptionally well from 2010 to 2012, better than the other fading algorithms. However it is outperformed in 2009 and 2013 by all the other algorithms. This shows that the sliding window is heavily influenced by the similarity of the data of 2 consecutive years. Meaning the abrupt fading algorithm is more sensitive to outliers. This is again shown by its standard deviation of the equality of opportunity in Table 1. The mean of the abrupt fading algorithm is very close to lowest and best mean of the equality of opportunity, however the standard deviation is considerably larger compared to the other algorithms. Thus indicating that the abrupt fading algorithm is not consistent.

	Mean EoO	STD EoO
Static	0.069839312	0.008136377
Dynamic no fade	0.049905245	0.00683245
Abrupt	0.046902824	0.011646918
Weight	0.046846232	0.006602121
Amount	0.048661459	0.006823865

Table 1: Mean and standard deviation of the equality of opportunity

On the other hand, the gradual weight fading algorithm is much more consistent, outperforming the other algorithms with the lowest mean and standard deviation for the equality of opportunity, see Table 1. However, the algorithm comes with a cost, as this algorithm also takes the longest to train. This introduces a trade-off between the equality of opportunity and the amount it takes to train a model. In this trade-off we disregard the accuracy of the different models as we barely observe a difference in the accuracy of the models, see Table 2.

The trade-off has 2 edge cases; the static model, which has very little training time as it is only trained once with 2008 data, but with a high equality opportunity, and the gradual weight fading algorithm, which trains continuously but with a much lower equality of opportunity.

	Mean score	STD score
Static	0.67821964	0.006672671
Dynamic no fade	0.673552594	0.006878765
Abrupt	0.6728605	0.005736844
Weight	0.672950157	0.006772536
Amount	0.67333697	0.006847099

Table 2: Mean and standard deviation of the score

In the middle of the trade-off we have gradual amount fading algorithm, which still has a significantly lower equality of opportunity compared to the static model while taking less time to train than the gradual weight fading algorithm.

7 Responsible Research

As this research deals with ethical and reproducible issues, it is important these are addressed. In this section both points are discussed.
The data used for the research are the files from the American Community Survey (ACS) Public Use Microdata Sample (PUMS) managed by the US Census Bureau. Each entry does represent an individual, making the data sensitive to ethical problems. However, the data is used responsibly through official sources and only used for their original purpose.
All the results are reproducible as the source code is made public on the gitlab page[1]. The random seeds used to split data are mentioned in the source code, as well as which data and models are used when training and testing the data.

8 Discussion

As can be seen in the results section, the fading algorithms certainly result in fairer models compared to static models. However, these algorithms also have a few downsides. In this section all the upsides and downsides are shown in a broader context.

The mail goal of the fading algorithms was showing that data should not be looked at in a static context for fairness. This goal was fulfilled as can be seen by the results of the fading algorithms in Table 1. Especially the fading of weight of data entries over the years shows a significant improvement in the equality of opportunity on the Adult dataset. However this comes with a cost. The accuracy of this algorithm has also decreased slightly. As the focus was on fairness monitoring, the decrease was acceptable.
But that is not the only downside. Another downside that should be taking into consideration is the efficiency. The baseline model is trained on data from 1 year, which is notably less than a model trained from data over 6 years. This works in favor of the model gradually increasing the amount of data over the years. This algorithm has a slight increase in equality of opportunity, a very small decrease in accuracy and takes significantly less time to train than the other dynamic algorithms.

[1] https://gitlab.ewi.tudelft.nl/cse3000/2023-2024-q4/Lukina/jjvanschijndel-Dynamic-Algorithmic-Fairness-in-Machine-Learning.git

Another limitation regarding this research, is that the amount of data was previously known. The range of years is used in the formulas to calculate the weight of the data. This means that implementing the models on actual datasets increasing each year is difficult and the formulas should be modified for actual continuous data.
Additionally, it should be noted that the adult dataset entries do not contain timestamps. All entries are only stored by year, making more fine-grained algorithms possible for other datasets.

9 Conclusions and Future Work

The main goal of the research was to find the effect of modifying the training datasets with fading algorithms on the fairness of a model. It can be concluded that there is a considerate effect of looking at the data dynamically instead of statically by the results shown in the previous sections. Fading algorithms show significant improvement on the equality of opportunity of a model on the Adult dataset. This however comes with a trade-off on the accuracy and the efficiency of a model.

There a various ways to modify a dataset to prioritise newer data, and when not taking anything into consideration apart from the equality of opportunity, the best way to do this is to gradually increase the weight of newer data. However, like mentioned before, the gradual weight fading isn't simply better than the other algorithms, as it takes the most amount of time to train.
We can conclude that when performing prediction tasks on the Adult dataset, we should make use of a dynamic fading algorithm in order to obtain fairer model. Which algorithm to use, depends on the specifics of the prediction task.

This side of algorithmic fairness monitoring can be and should be researched more extensively. For example, the fading algorithms can be tested on other datasets and tweaked to that datasets for better results. We can also merge the different fading algorithms, e.g. the abrupt fading with the gradual weight algorithm, or the gradual amount with the gradual weight algorithm.
Additionally, a few limitations of this research were mentioned in the previous section. The most interesting future research would include more fine-grained formulas as the current uniform formula for the gradual algorithms is quite simple. There is plenty of room for improvement, however showing the significant effect of ageing of the adult dataset over the years is a good starting point.

References

[1] Aws Albarghouthi and Samuel Vinitsky. Fairness-aware programming. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 211–219, New York, NY, USA, 2019. Association for Computing Machinery.

[2] Albert Bifet, Geoff Holmes, Bernhard Pfahringer, Jesse Read, Philipp Kranen, Hardy Kremer, Timm Jansen, and Thomas Seidl. Moa: A real-time analytics open source framework. In Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 617–620, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[3] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

[4] Alexander D'Amour, Hansa Srinivasan, James Atwood, Pallavi Baljekar, D. Sculley, and Yoni Halpern. Fairness is not static: deeper understanding of long term fairness via simulation studies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 525–534, New York, NY, USA, 2020. Association for Computing Machinery.

[5] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[6] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):1–37, 2014.

[7] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning, 2016.

[8] Zhimeng Jiang, Xiaotian Han, Hongye Jin, Guanchu Wang, Rui Chen, Na Zou, and Xia Hu. Chasing fairness under distribution shift: A model weight perturbation approach, 2023.

[9] Sanjit Kaul, Roy Yates, and Marco Gruteser. Real-time status: How often should one update? In *2012 Proceedings IEEE INFOCOM*, pages 2731–2735, 2012.

[10] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. *CoRR*, abs/1803.04383, 2018.

[11] Qingyu Liu, Chengzhang Li, Y. Thomas Hou, Wenjing Lou, Jeffrey H. Reed, and Sastry Kompella. Ao2i: Minimizing age of outdated information to improve freshness in data collection. In *IEEE INFOCOM 2022 - IEEE Conference on Computer Communications*, pages 1359–1368, 2022.

[12] Diomidis Spinellis. The decay and failures of web references. *Communications of the ACM*, 46(1):71–77, 2003.

[13] Sebastian Wagner, Max Zimmermann, Eirini Ntoutsi, and Myra Spiliopoulou. Ageing-based multinomial naive bayes classifiers over opinionated data streams. volume 9284, 11 2015.