Impact of the IBM AI Fairness 360 toolkit for outliers

on the example of gender classification

for people with Down syndrome

Selected Topics of Artificial Intelligence

Presented by

Patrick Brenner

Niklas Janssen

pb055, nj015

February 24, 2019

at the Stuttgart Media University

Supervisor: Prof. Dr. Gottfried Zimmermann

Abstract

Machine learning models are getting increasingly used for real-life applications. The decisions they are making are getting more impactful on people's lives. A big problem in these systems is bias transferred from society into the algorithms. Either unknowingly (or knowingly) through the developer or indirectly enclosed in the data, used for training the model.

One possible solution for this problem is the IBM AI Fairness 360 toolkit. It can either change the weights of the data used to make it fairer or equalize the decisions that an algorithm makes to remove the gap between privileged and less privileged groups. This paper evaluates if it is already applicable for disabled people on the example of people with Down syndrome in a system that determines gender based on face images. We found that there is a slight improvement when using the toolkit in contrast to not using it.

Table of contents

Introduction

As AI systems are getting increasingly used for real-life applications it is getting more important to find solutions for a fair treatment for everyone affected including minorities. With more far-reaching consequences of decisions machine-learned models have to take, it must be ensured that everyone is treated fairly and that existing inequalities in our society are not translated into a learnt model.

Many of the equality related problems that are occurring in machine learning can be traced back to existing social problems. Data often already contains an unwanted bias either through over-/under-sampling of a group or through the prejudices of the person labelling it [1]. Under-sampling in the case of disabilities is difficult to avoid, as there is not enough data available in most datasets. Because of the data problem other approaches must be taken. The IBM Fairness 360 kit aims to be one of those approaches. It works with inequitable distributions of data and tries to remove bias without changing the data itself. A short overview over the tool is given in 1.2.

In the following it is then evaluated if the IBM AI Fairness 360 kit can already make a difference for disabled people without solving the data problem. As an application the gender classification of face images is used which is available on the IBM website. The experiment is restricted on the Down syndrome as an outlier characteristic.

1 AI Fairness

1.1 Fairness in AI systems

The main problem with AI systems in the field of image processing is how vastly different a human's characteristics can be. It is very difficult to train a model in a way that it considers all the assorted attributes a human can have.

Furthermore, there is often not enough data of so-called "outliers". These can for example be people with disabilities like the Down syndrome. Even though about 0.1% of newborn children are affected by it [2] datasets seldom represent that number. The reasons for that are wide-ranging. Many times, people with a disability keep it for themselves in fear of any negative societal effects a disability might induce. In addition, differing life situations can impede their entry in a data set.

Datasets that include enough people with disabilities can still be problematic. For example, a statistic of a health insurance company will contain many entries that link high costs to people with a disability. A model that learns from such a dataset will automatically rate someone with a certain disability as a high cost factor. Even though this might not be the case for this person.

Datasets can often contain bias and patterns of existing exclusion in our society [1]. People with a migration background e.g. in general often earn less than their more privileged counterparts. An algorithm that evaluates their credit score might give them a lower score just because they have a certain attribute that the algorithm connects with a lower income.

Therefore, some classes for fairness in machine learning have been established over the last years. Especially three formal definitions have emerged [3].

1. Anti-classification: Not taking protected characteristics like race or disability into account.
2. Classification parity: Equal classification rates for privileged and non-privileged groups.
3. Calibration: Actual risk outcomes should be independent from certain protected attributes when compared with the estimated value.

[3] proves however that these three rules alone are still not enough to reveal discriminatory algorithms. Using them to create algorithms can even harm the goals that want to be achieved [3]. For example, would rule number 1 still not avoid bias already contained in the dataset.

This shows that designing AI systems in a way that they act non-discriminatory is even harder than one might initially expect. Other solutions must be found to ensure that a machine learned model can still be fair even though its preconditions weren't ideal. That's why the IBM AI Fairness 360 kit has been developed.

1.2 IBM AI Fairness 360

The IBM AI Fairness 360 toolkit (AIF360) is a "*comprehensive open-source toolkit of metrics to check for unwanted bias in datasets and machine learning models, and state-of-the-art algorithms to mitigate such bias*" [7]. It was developed by IBM and announced in September 2018.

It initially contained nine separate algorithms, developed in a group effort by the algorithmic fairness research community [7]. These algorithms can all be called directly in a program after the package is imported [7].

There are several ways to use the AI Fairness 360 Toolkit. You can choose from one of the now 10 algorithms used to mitigate bias in AI systems like "optimized pre-processing" or "reweighing" [4]. All of them have their own functionality and work on a different level. There are three groups of algorithms [4]. The first one works on the training data and shifts the weights to make them more equal before the model is learning from them. The second works on the creation of the model itself while the model is being trained. The third one is working on the classification process.

There is also a wide range of metrics available that can be used to create statistics and diagrams. For example, the Theil Index which can show the inequality in benefit allocation for the individuals [4].

Community collaboration is highly supported and there are many resources to choose from. On the IBM website are multiple tutorials and demos that explain how you can easily integrate the AIF360 toolkit into your program. One of these tutorials is the Jupyter notebook for the gender classification of face images that are used in the experiment in chapter 2.

2 IBM AI Fairness 360 for people with Down syndrome

In the following test series, the test impact of the AI Fairness 360 toolkit for people with Down syndrome is tested in an image gender classification model by using the AIF360 tutorial. First, it is checked, if a bias for people with Down syndrome is detected and then it is important to see if and how the AIF360 toolkit can help to mitigate this bias with the reweighing algorithm.

2.1 Test environment

We use a Python 3 Jupyter notebook to execute the tutorial, which uses a convolutional neural network (CNN) with 3 convolutional layers and 2 fully connected layers to train the network for gender-classification. It learns a baseline classifier and obtains fairness metrics to predict the gender. Originally, the tutorial was showing AI deviation for different groups defined by skin tone. Therefore, we added people with Down syndrome as a separate group and labelled them accordingly (see section 2.2). Additionally, to carrying out the desired tests, the code must be changed to store the test datasets specifically and to be able to view it independently of the training data set. This was not provided in the original version of the notebook.

As algorithm for bias mitigation the reweighing algorithm is used in this tutorial. Reweighing is a pre-processing technique that weights the examples in each (group, label) combination differently to ensure fairness before classification [5].

The entire, customized notebook with all training and test data can be found here:

https://github.com/pb055/IBMFairness360_DownSyndrom_Test

2.2 Test and training datasets

The images from people with Down syndrome used in the experiment have all been taken from free sources. Part was taken from Wikimedia and part from Flickr. Wikimedia's policy on reusing their images can be found here: [8].

IBM's tutorial is using the UTKFace dataset. The dataset contains over 20,000 face images with a wide variety of gender, age and ethnicity [9]. Disabilities are not considered explicitly in the labels, but the dataset might already contain some images of disabled people. This forced us to search freely usable pictures on our own and add them externally to the dataset. For the algorithm to recognise these pictures we had to label them accordingly in a format that contained the age, gender, race and the date and time the image was added to the dataset. For the Down syndrome to be considered we used an extra value for the "race" attribute which until that point consisted of White, Black, Asian, Indian and Others [9].

Images from the test data set:

Figure 1: Man with Down syndrome [10]

Figure 2: Little girl with Down syndrome [11]

Figure 3: Boy without Down syndrome [12]

Figure 4: Woman without Down syndrome [12]

2.3 Test scenarios and results

To determine the impact of the algorithm with and without people with Down syndrome, two different training scenarios are tested. First, we trained the model without people with Down syndrome (2.3.1) and then with a realistic number of people with Down syndrome (2.3.2). In both cases, the CNN is trained with 16.000 different images and was tested with the same test images.

Both tests are with 30 persons, 15 males and 15 females. One test is with and one without Down syndrome to see differences and a possible AI bias. Because the reweighing algorithm is based on group combinations, each test run will specify an unprivileged group that adjusts the training data for the algorithm. For people without Down syndrome we take group 0 (White) as they appear most often in both test data sets. People with Down syndrome are labelled with group 5.

The test results are shown as follows.

- Correctly predicted: Number of correctly predicted images
- Falsely predicted: Number of falsely predicted images
- Gender classification accuracy:
  Proportion of correctly recognized images in percent.
- Balanced classification accuracy:
  Accuracy which also considers the true positives (TPR) and true negatives rate (TNR).

Balanced classification accuracy = 0.5 x (TPR+TNR)

2.3.1 Training-Scenario 1: Without people with Down syndrome

In the first run of our series of experiments we trained the model only with people without Down syndrome and wanted to see what accuracies were achieved. The two provided test data sets were used and the following results were determined:

Test results for 30 people without Down syndrome

	Without reweighing:	**With reweighing:**
Correctly predicted:	25	26
Falsely predicted:	5	4
Gender classification accuracy:	83,33%	86,67%
Balanced classification accuracy:	83,33%	86,67%

Table 1: Test scenario 1 - Results without Down syndrome

Test results for 30 people with Down syndrome

	Without reweighing:	**With reweighing:**
Correctly predicted:	21	22
Falsely predicted:	9	8
Gender classification accuracy:	70,00%	73,33%
Balanced classification accuracy:	69,20	71,43%

Table 2: Test scenario 1 - Results with Down syndrome

2.3.2 Training-Scenario 2: Realistic number of people with Down syndrome

In the second run of our series of experiments we trained the model with a realistic number of people with Down syndrome and wanted to see what accuracies were achieved. According to Statista, about 50,000 people with Down syndrome currently live in Germany (0.6% of the total population) [6]. We therefore used a similar value for our training dataset and trained the model with 0.1% people with Down syndrome. From our training dataset (total 16,000) 16 people with Down syndrome were trained. The two provided test data sets were used and the following results were determined:

Test results for 30 people without Down syndrome:

	Without reweighing:	**With reweighing:**
Correctly predicted:	26	27
Falsely predicted:	4	3
Gender classification accuracy:	86,67%	90,00 %
Balanced classification accuracy:	86,67%	90,00 %

Table 3: Test scenario 2 - Results without Down syndrome

Test results for 30 people with Down syndrome:

	Without reweighing:	**With reweighing:**
Correctly predicted:	20	20
Falsely predicted:	10	10
Gender classification accuracy:	66,67%	66,67%
Balanced classification accuracy:	66,52%	66,96%

Table 4: Test scenario 2 - Results with Down syndrome

3 Assessment and Conclusion

After examining the test results, it becomes visible that there is an AI bias towards persons with Down syndrome. In the test without the AIF360 toolkit there are on average 5 persons less being classified correctly. It is shown that a person with Down syndrome is more likely to be classified as female which is part of the reason there is such a high rate of wrong classifications.

An increased quota of people with down syndrome in the training data with real-life percentages (see chapter 2.3.2) doesn't improve these results in our test but even makes accuracy worse. This might be because of the small scale of our experiment and is expected to differ with a larger number of images. The scarcity of appropriate and freely usable data however prevented a larger scale. This shows once again that there is not enough usable data of people with disabilities which impedes the research progress in this field. Datasets must include data of disabled people and they have to be explicitly labelled for it to be useful to research communities like the one surrounding AIF360.

Because of the reweighing algorithm there is a notable improvement in accuracy. Especially in the test which didn't use images of people with Down syndrome for the training process the algorithm surprisingly strongly improves accuracy. Even though in the test run with a low amount of training data there isn't such a big improvement in the percentage, the true-positive-rate is higher.

A different opportunity for the usage of AIF360 might be to use multiple models in succession for the to be predicted data. For example, could image classification models, trained to identify different disabilities label the data beforehand which could improve the classification process if the latter model uses the labelling. This approach should be subject to testing by the research community.

In conclusion we can say that the AIF360 toolkit in the form of the reweighing algorithm shows a decrease in the clearly present bias towards people with Down syndrome. It can therefore already be applied in practical appliances after making sure that an improvement and not a deterioration is also being seen in other disabilities affecting facial features. Also, solutions for the labelling must be found as the race attribute in our application is just a workaround. Other algorithms of the AIF360 toolkit should also be tested to find an optimal solution for every application.

References

1. Barocas, Solon, and Andrew D. Selbst.
"*Big data's disparate impact*" (Cal. L. Rev. 104, 2016: 671)

2. World Health Organization,
“Genes and human disease”
https://www.who.int/genomics/public/geneticdiseases/en/index1.html (last checked: 02/14/2019 at 20:00)

3. Corbett-Davies, Sam, and Sharad Goel,
"*The measure and mismeasure of fairness: A critical review of fair machine learning*" (arXiv preprint arXiv:1808.00023, 2018)

4. AI Fairness 360
"*Try a Web-Demo*"
https://aif360.mybluemix.net/ (Last checked 02/18/2019 at 10:00)

5. F. Kamiran and T. Calders
"Knowledge and Information Systems, Data Preprocessing Techniques for Classification without Discrimination", 2012.

6. Statista
"*Down-Syndrom in Deutschland*"
https://de.statista.com/infografik/15758/down-syndrom-in-deutschland/ (Last checked 02/20/2019 at 10:00)

7. Kush R. Varshney
“Introducing AI Fairness 36”
https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/ (Last checked 02/20/2019 at 10:00)

8. Wikimedia
https://commons.wikimedia.org/wiki/Commons:Reusing_content_outside_Wikimedia (Last checked 02/23/2019 at 12:00)

9. Yang Song, Zhifei Zhang
https://susanqq.github.io/UTKFace/ (Last checked 02/23/2019 at 12:00)

Images

10. Wikimedia images

https://upload.wikimedia.org/wikipedia/commons/thumb/a/ae/Pinedap.JPG/800px-Pinedap.JPG (Last checked 02/23/2019 at 12:00)

11. Flickr

http://farm9.staticflickr.com/8520/8544662380_79056a7306_b.jpg (Last checked 02/23/2019 at 12:00)

12. UTKFaces

https://susanqq.github.io/UTKFace/