

Data Quality Dimensions for Fair AI

Camilla Quaresmini and Giuseppe Primiero



UNIVERSITÀ DEGLI STUDI DI MILANO

DIPARTIMENTO DI FILOSOFIA
"PIERO MARTINETTI"

28 October 2021

Table of Contents

Background

Bias Mitigation Tools

Case Study

Approach

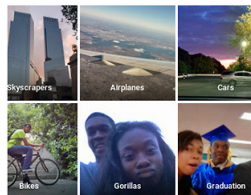
Classification

Definition

Classification is a task related to a predictive modeling problem, where a class label is predicted for a given input data, based on a classification model.

Classification: Problems

- ▶ Input data are datasets.
- ▶ Theoretically, datasets are representative of the world. But in the real life they are not.
- ▶ This results in systems full of biases.



Classification: Problems

- ▶ Boulamwini and Gebru (2) demonstrate that training and evaluation data for such algorithms are often biased, producing distorted results.
- ▶ According to a recent MIT's study by Northcutt (4), the most well-known AI datasets are full of labeling errors.
- ▶ <https://www.image-net.org>

```
drug user, substance abuser, user (15)
├─ addict (8)
│   ├── speed freak (0)
│   ├── caffeine addict, caffeine addict (0)
│   └─ drug addict, junkie, junky (5)
│       ├── opium addict, opium taker (0)
│       ├── crack addict, binger (0)
│       ├── withdrawer (0)
│       ├── cocaine addict (0)
│       └─ heroin addict (0)
├─ tripper (0)
├─ head (4)
├─ agnostic, doubter (0)
├─ greeter, saluter, welcomer (0)
├─ percher (0)
├─ gentle (0)
├─ laughter (1)
├─ baldhead, baldpate, baldy (0)
├─ advocate, advocator, proponent, exponent (109)
├─ nonreligious person (14)
│   ├── positivist, rationalist (1)
│   ├── nihilist (0)
│   ├── disbeliever, nonbeliever, unbeliever (3)
│   ├── heathen, pagan, gentle, infidel (3)
│   │   └─ paynim (0)
│   └─ idolater, idolizer, idoliser, idol worshiper (1)
├─ blasphemer (1)
├─ deist, freethinker (0)
├─ abjurer (0)
├─ Pisces, Fish (0)
├─ junior (0)
├─ celebrant, celebrator, celebrater (2)
└─ termor (0)
```

```
mutilator, maimer, mangler (0)
sniffer (0)
turner (0)
spitter, expectorator (0)
autodidact (0)
religious person (157)
nondescript (0)
capitalist (166)
second-rater, mediocrity (0)
African (74)
walk-in (0)
neighbor, neighbour (0)
miracle man, miracle worker (0)
entertainer (176)
communicator (313)
masturbator, onanist (2)
showman (0)
transvestite, cross-dresser (0)
nonperson, unperson (0)
slave (0)
weaking, doormat, wuss (3)
creditor (1)
├─ mortgagee, mortgage holder (0)
└─ picker, chooser, selector (0)
suspect (3)
├─ murder suspect (0)
├─ rape suspect (0)
└─ robbery suspect (0)
simpleton, simple (32)
survivor (0)
innocent, inexperienced person (4)
creator (175)
killer, slayer (25)
```

Bias Mitigation Tools

- ▶ Such problems are beginning to be addressed by the ML community.
- ▶ A number of tools are available to evaluate the probability that a labelling error has occurred, and repair it.

Cleanlab

- ▶ Cleanlab is an open-source Python package for finding and learning with label errors in datasets.
- ▶ It is powered by Confident Learning (CL).

Cleanlab: Description

Assumption

Data are categorical, they fit one label.

Task

Given a mapping $y^ \rightarrow \tilde{y}$ between variables, where y^* is the correct label and $\tilde{y}$ the wrong one, what is the probability*

$$p(\tilde{y} = i | y^* = j)$$

such that label i is wrong, given that label j is correct?

Definition (Confident Joint)

$$C_{\tilde{y}, y^*}[i][j] := |\hat{X}_{\tilde{y}=i, y^*=j}|$$

Case Study

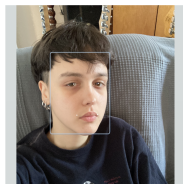
- ▶ We consider the FERET database.
- ▶ We focus on the particularly difficult (technically: noisy) attribute Gender, labels available `male;female`.
- ▶ We provide a conceptual and technical analysis with respect to available bias mitigation tools.

Case Study

- ▶ We question the underlying assumption on data dimensions:
 1. completeness on labeling;
 2. accuracy as a measure of dataset correctness.
- ▶ We aim at suggesting improvements on errors identification in the classification of two difficult examples:
 1. non-binary individuals = datapoints which have none of the available labels as the correct one;
 2. transgender individuals = datapoints which have a label that changes under different conditions.

First Case: Incomplete Label Set

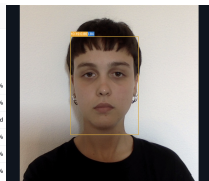
- Consider a datapoint represented by a non-binary individual, whose actual label is missing.



▼ Results

	
looks like a face	99.9 %
appears to be male	98.2 %
age range	10 - 20 years old
not smiling	99.1 %
appears to be calm	99.1 %
not wearing glasses	96.6 %

(a) *Amazon Rekognition.*



GENDER APPEARANCE 1	
	Masculine 8.04
	Feminine 8.16
AGE APPEARANCE 1	
	10-19 8.86

(b) *Clarifai.*

Figure: Example of incomplete label set.

Scheuerman's Study, (5)

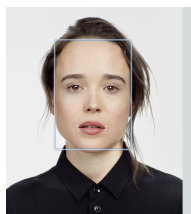
- ▶ A possible strategy is to implement a larger partition of labels adding separate categories.
- ▶ Technically, consider the completeness of the labels at a higher level of abstraction.
- ▶ 7 different genders.
- ▶ They consider four famous systems:

TPR Performance Per Gender Hashtag													
Hashtag	Amazon			Clarifai			IBM			Microsoft			All
	T	F	TPR	T	F	TPR	T	F	TPR	T	F	TPR	Avg
#woman	348	2	99.4%	333	17	95.1%	345	5	98.6%	100	0	100.0%	98.3%
#man	334	16	95.4%	344	6	98.3%	341	9	97.4%	348	2	99.4%	97.6%
#transwoman	317	33	90.6%	271	79	77.4%	330	20	94.3%	305	45	87.1%	87.3%
#transman	216	134	61.7%	266	84	76.0%	250	100	71.4%	255	95	72.8%	70.5%
#agender,													
#genderqueer,	—	—	—	—	—	—	—	—	—	—	—	—	—
#nonbinary													

- ▶ Available services offer a static classification non representative of individuals who are not cisnormative.

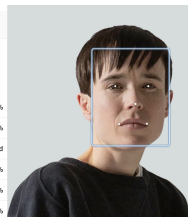
Second Case: Inconsistent Datapoint

- ▶ A datapoint representing a transgender individual.
- ▶ In this case any extension of the label set is in fact misleading.



```
"Gender": {  
  "Value": "Female",  
  "Confidence": 99.89617156982422  
},
```

(a) *Ellen Page in 2015.*



```
"Gender": {  
  "Value": "Male",  
  "Confidence": 94.46698760986328  
},
```

(b) *Elliot Page in 2021.*

Figure: Example of inconsistent datapoint in Amazon Rekognition.

Philosophical Stand

- ▶ The assumption that completeness and accuracy are the only relevant data quality dimensions for classification tasks is wrong.
- ▶ The literature on data and information quality (1; 3) provides a number of dimensions which can help improving classification and bias mitigation.
- ▶ In particular, we believe an important starting point is represented by adding the dimension of **timeliness**.

Our Strategy

Add a time frame $\mathcal{T} = t_1, \dots, t_n$ whose length depends on the dataset and the classification task over the pairing of data points to labels, to measure a probability of a label-change over time.

Proposition

Given the label set is complete at time i , it can be incomplete at time $i + n$

Proposition

Given the label i is correct for data point d at time i , it can be incorrect at time $i + n$

Implementing the strategy for Cleanlab (I)

Assumption

The probability value of a given label i being wrong, given a label j is correct (their distance) may change over time.

Task

Given a mapping $y^ \rightarrow \tilde{y}$ between variables, where y^* is the correct label and $\tilde{y}$ the wrong one, what is the probability*

$$p_{\mathcal{T}}[(\tilde{y} = i)_{t_n} | (y^* = j)_{t_{n-m}}]$$

such that label i is wrong at time t_n , given that label j was correct at time t_{n-m} ?

Definition (Confident Joint)

$$C_{\tilde{y}, y^*}[i, j, \mathcal{T}] := |\hat{X}_{\tilde{y}=i_{t_n}, y^*=j_{t_{n-m}}}|, \forall m < n$$

Implementing the strategy for Cleanlab (II)

Assumption

(Some) Data are not categorical for some labels, they fit one label at any given time, but possibly different ones at different times.

Task

Given a mapping $y^ \rightarrow \tilde{y}$ between variables, where y^* is the correct label and $\tilde{y}$ the wrong one, what is the probability*

$$p_{\mathcal{T}}[(\tilde{y} = i)_{t_n} | (y^* = i)_{t_{n-m}}]$$

such that label i is wrong at time t_n , given that label i was correct at time t_{n-m} ?

Definition (Confident Joint)

$$C_{\tilde{y}, y^*}[i, \mathcal{T}] := |\hat{X}_{\tilde{y}=i_{t_n}, y^*=i_{t_{n-m}}}|, \forall m < n$$

Conclusions

- ▶ The present study demonstrates that data quality is not just maximizing accuracy.
- ▶ A way to make current bias mitigation tools fairer is to implement more data quality metrics.
- ▶ Laying the ground for an extension of associated assessment tools.

References

- [1] Batini Carlo, Scannapieco Monica, “Data Quality: Concepts, Methodologies and Techniques”, 2006.
- [2] Buolamwini Joy, Gebru Timnit, “Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification”, 2018.
- [3] Floridi Luciano, Illari Phyllis, “The Philosophy of Information Quality”, 2006.
- [4] Northcutt Curtis, Athalye Anish, Mueller Jonas, “Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks”, 2021.
- [5] Scheuerman Morgan Klaus, Paul Jacob, Brubaker Jed, “How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services”, 2019.