Data Quality Dimensions for Fair AI

Camilla Quaresmini and Giuseppe Primiero

Department of Philosophy, University of Milan

An Important Problem

AI algorithms reflect technical errors originating with the problem of mislabeling data. As systems mistakenly identify patterns in data, making wrong classification, they are not systematically guarded against bias.

Methodology

The problem of bias in AI systems is considered from the point of view of Information Quality dimensions.
A tool for bias mitigation (Cleanlab) is examined, which defines fairness focusing on the accuracy dimension. But other dimensions of data quality are necessary.

Case Study

Potential improvements are illustrated using gender classification errors as a case study in two difficult contexts, for which dimensions of completeness and consistency are called upon:

- non-binary individuals: the label set becomes incomplete with respect to the dataset;
- transgender individuals: the dataset is inconsistent with respect to the label set.

A further extension in terms of timeliness is proposed. On this basis, a theoretical framework is provided to redefine fairness of AI systems.

Formal Analysis

To illustrate the point a toy example is considered. Compute

$$p(\tilde{y} = i \mid y* = j) \qquad (1)$$

i.e. the error rate of $y* = male$ has to be determined. Then consider the same dataset at a later time t_{n+m}. The labels might have changed.

To understand how the error rate changes, the difference between two confusion matrices has to be considered. The change rate can be computed as

$$\varepsilon = \hat{p}'(\tilde{y}; x_i; \theta) - \hat{p}(\tilde{y}; x_i; \theta) \qquad (2)$$

Equation 1 can be computed with respect to time as

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

Important Result

Timeliness is a founding dimension for developing fairer and more inclusive classification tools.

Considering a possible implementation in Cleanlab, the assumption is that the probability of assigning a label may change over time.

This can be formulated in two distinct ways.

First

$$p_{\mathcal{T}}[(\tilde{y} = i)_{t_n} \mid (y* = j)_{t_{n-m}}] \qquad (3)$$

Second

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

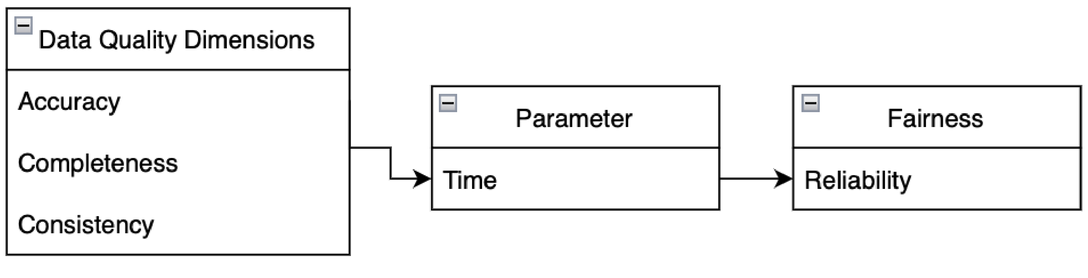

Figure 1:Useful dimensions

Results

A more general discussion on the data dimensions to be adopted in bias mitigation tools is needed.
The dimension of timeliness is crucial, hence an explicit temporal parametrization is suggested.

Completeness is considered as a relationship between a label set and an individual p belonging to a certain population P, where p is any domain item that enters P at a time t. There must exist a correct label l for each datapoint in the dataset at each time.

Completeness

A label set L for a classification algorithm in a AI system X is considered complete over a time frame $\mathcal{T} : \{t_1, \ldots, t_n\}$ denoted as $Compl_{\mathcal{T}}(L(X))$ iff given two partitions $L_{t_1} := \{l_1, \ldots, l_n\}$ and $L_{t_n} := \{l'_1, \ldots, l'_n\}$, where possibly $L_{t_1} \cap L_{t_n} \neq \emptyset$ for all $(p \in P)_{\mathcal{T}}$ s.t. $p \in d(X)_{\mathcal{T}}$ there is $l \in L_{t_1} \cup L_{t_n}$ s.t. $y^*(d) = l$.

Next, consistency of the label set with respect to datapoints possibly shifting in categorization is considered.
The method has been to reduce consistency to timeliness.

Reliability

A classification algorithm in a AI system X is considered reliable over a time frame $\mathcal{T} := \{t_1, \ldots, t_n\}$ denoted as $Rel_{\mathcal{T}}(X)$ iff $\varepsilon_{\mathcal{T}}(X) < \pi$, for some safe value π.

$\varepsilon_{\mathcal{T}}$ represents consistency expressed in terms of temporalized accuracy, denoting reliability.

Conclusion

The two previous definitions offer non-exhaustive criteria for the identification of fairness.

The problem of unfairness in AI can be expressed in terms of data quality: AI systems are limited in that they maximize accuracy, and even if systems become statistically accurate some problems remain unsolved.

Current classification methods are rooted on three assumptions on gender:

- Binarism
- Staticity
- Derivability from physical traits

These design limitations must be addressed if fairer classifications and more inclusive models of gender are to be designed.

Fairness for AI systems

$Fair_{\mathcal{T}}(X)$ only if $Rel_{\mathcal{T}}(X)$ and $Compl_{\mathcal{T}}(L(X))$.

References

[1] C. Batini and M. Scannapieco.
Data Quality: Concepts, Methodologies and Techniques.
Springer, 2006.

[2] P. Illari and L. Floridi.
The Philosophy of Information Quality.
Springer International Publishing, 2014.

[3] C. G. Northcutt, W. Tailin, and I. L. Chuang.
Learning with confident examples: Rank pruning for robust classification with noisy labels.
In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*, Sydney, Australia, 2017. AUAI Press.

[4] C. G. Northcutt, L. Jiang, and I. L. Chuang.
Confident learning: Estimating uncertainty in dataset labels.
Journal of Artificial Intelligence Research (JAIR), 70:1373–1411, 2021.