The SQuaRE Series as a Guarantee of Ethics in the Results of AI systems

Alessandro Simonetta\textsuperscript{1,2,4,\ast}, Maria Cristina Paoletti\textsuperscript{3,\ast} and Tsuyoshi Nakajima\textsuperscript{4,\dagger}

\textsuperscript{1}Department of Enterprise Engineering, University of Rome Tor Vergata, Rome, Italy
\textsuperscript{2}Italian Space Agency, via del Politecnico 001, Rome, Italy
\textsuperscript{3}Professional Association of Italian Actuaries, Rome, Italy
\textsuperscript{4}Department of Computer Science and Engineering Shibaura Institute of Technology, Tokyo, Japan

Abstract

AI is an enabling technology that can be utilized in various fields with impressive results. However, in its adoption, there are risk factors that can be mitigated through the adoption of quality standards. It’s not by chance that the new ISO/IEC 25059 includes a specific quality model for AI systems. The article describes a research approach that proposes a way to prevent the lack of quality in training data from propagating into the deductions of an AI system. This is all based on the concept of completeness from ISO/IEC 25012 and can be referred to ISO/IEC 5259-2 characteristics of diversity, representativeness, similarity for input dataset evaluation and to ISO/IEC 25059 functional correctness for output results evaluation.

Keywords

Fairness, Machine Learning, Completeness, ISO/IEC 25012, Maximum Completeness, Bias, Classification

1. Introduction

The vast availability of data and tools has allowed the construction of predictive and classification models that form the foundation of Automated Decision-Making (ADM) systems. Many business decisions rely on recommendations generated by software systems, and in some cases, these decisions are entirely automated. The notion that this promotes the concept of decision neutrality due to being algorithm-based is quite prevalent. However, since the decision-making path of an AI system is heavily influenced by the data used during the learning phase, biases present in the data can sometimes transfer into the choices proposed by the system. In the literature, it has been demonstrated that the use of AI systems trained on biased datasets can lead to situations of discrimination [1]. The risk of skewed outcomes primarily stemming from imbalanced datasets has also been studied, and it can be mitigated by the introduction of synthetic data [2]. Learning algorithms construct the model based on the training data, so such disproportion can lead to conclusions that deviate from reality [3,4]. On the other hand, in some situations, it is challenging to obtain homogeneous, proportional, and, most importantly, representative data. In these cases, the ISO standards that can help us are [1]:


Specifically, ISO 31000 includes risk management principles that allow for the assessment of both the risk of using incomplete data during the learning phase and the risk associated with unfair predictions [1]. Other kinds of risks, such as the ability of protect data from information leakage, are for further study. ISO/IEC 27002 offers two possible new approaches for proactive security, threat detection and machine learning/artificial intelligence systems. Initially, the ISO/IEC 25100 software quality model [6] did not encompass quality characteristics of AI systems. However, starting from 2023, the SQuaRE series is enriched with the quality model for AI systems: ISO/IEC 25059 standard. Table 1 presents the new sub-characteristics identified by the working group and their scope in relation to the original standard [7].
### 2. Fairness Evaluation in ML Outputs

In the context of machine learning, evaluating fairness in machine learning models is a very sensitive and important issue. The goal is to ensure that models yield results that are independent of group membership and do not perpetuate or, in some cases, even exacerbate existing societal inequalities.

There are two different approaches: measuring the intensity of output errors or measuring the overall direction of errors. The first approach focuses on assessing disparate or unfair errors among different categories, ethnicities, or groups. The second approach evaluates whether the model tends to make errors in a particular direction or towards a specific group, ethnicity, or other sensitive attribute. Bias or fairness metrics can be used to evaluate this overall direction.

In the case of classification algorithms, the confusion matrix $P$ allows for the calculation of the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN)

$$P = \begin{bmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{n1} & \cdots & p_{nn} \end{bmatrix} \quad (1)$$

$$TP(i) = p_{ii} \quad (2)$$

$$FP(i) = \sum_{k=1}^{n} p_{ik} \quad (3)$$

$$TN(i) = \sum_{k=1}^{n} p_{ik} \quad (4)$$

The concepts of precision, recall, and accuracy are well-known in the literature and are presented below for the sake of completeness in the discussion:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (5)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (6)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (7)$$

Accuracy is a measure of functional correctness according to ISO IEC 25059 and ISO IEC TS 4213.

### 3. Statistical Evaluation Methods on Output

In a classification or decision scenario, statistical criteria allow us to evaluate discrimination in terms of statistical expressions involving the random variables $A$ (sensitive attribute), $Y$ (target variable), and $R$ (the classifier or score). Therefore, it is easy to determine whether a criterion is satisfied or not by calculating the joint distribution of these random variables. Starting from the definition of independence introduced in \[8\], for there to be independence between two values of the sensitive attribute, we need to verify that the joint probability has the same values in both cases $a_i$ and $a_j$:

$$P(R = 1|A = a_i) = P(R = 1|A = a_j) \quad (8)$$

According to this hypothesis, the ideal case of perfect fairness occurs when the probabilities have the same value. As a consequence of this consideration, a measure of non-independence is obtained by calculating the distance between the two values, which is zero in the ideal case of complete independence:

$$\Phi(a_i, a_j) = |P(R = 1|A = a_i) - P(R = 1|A = a_j)| \quad (9)$$

Table 2 shows the calculation of joint probabilities in the case of the well-known Compas dataset \[9\], in which the ML system incorrectly predicted a higher degree of recidivism among African-American detainees.
Table 2

| Race         | \(P(R = 1|A = a_i)\) | Centroid |
|--------------|----------------------|----------|
| Caucasian    | 0.33                 |          |
| Hispanic     | 0.28                 | 0.26     |
| Other        | 0.20                 |          |
| Asian        | 0.23                 |          |
| African-American | 0.58             |          |
| Native-American | 0.73              | 0.65     |

In the case of the Compas dataset, the joint probabilities cluster around two centroids, which supports the reasoning that it would be more reasonable to select these two points as representative of the two treatment groups. In fact, if the probability values cluster into subsets of values, it signifies fair independence within the group and, conversely, inequity between groups. If the distribution of probability values is nearly uniform and it is not possible to identify distinct groups, or if the number of groups is greater than two, you can calculate the independence measure through the average of distances:

\[
U(a_1, \ldots, a_m) = \frac{2}{m(m-1)} \sum_{i=1}^{m} \sum_{j=i+1}^{m} U(a_i, a_j) \quad (10)
\]

4. Mutual Information

The concept of mutual information allows for the measurement of relationships between the joint probabilities mentioned in (9). Indeed, it can measure the mutual information between A and R, which is the amount of information one random variable reveals about the other. Therefore, the condition of independence between the random variables A and R, as indicated in (9), can be expressed in terms of mutual information:

\[
I(A, R) = H(A) + H(R) - H(A, R) \quad (11)
\]

where \(H(R)\) and \(H(A)\) are the entropies associated with R and A, respectively:

\[
H(R) = \sum_{i=1}^{n} P(r_i) \log(P(r_i)) \quad (12)
\]

\[
H(A) = \sum_{i=1}^{n} P(a_i) \log(P(a_i)) \quad (13)
\]

Instead, the third term in equation 12 is:

\[
H(R, A) = \sum_{i=1}^{n} \sum_{j=1}^{m} P(r_i, a_j) \log(P(r_i, a_j)) \quad (14)
\]

The other indices can also be expressed by mutual information and in particular referring to [11] and [10]. Separation is calculated by:

\[
I(R, A|Y) = H(R, Y) + H(A, Y) - H(R, Y, A) - H(Y) \quad (15)
\]

sufficiency is expressed by the following equation:

\[
I(Y, A|R) = H(Y, R) + H(A, R) - H(Y, R, A) - H(R) \quad (16)
\]

finally, the Overall Accuracy Equality (17) is computed by:

\[
H(A, R|Y = R) = H(A, Y = R) + H(R, Y = R) - H(R = Y, A|R = Y) \quad (17)
\]

5. Data Quality Measures for Input

The underlying idea of this research is to find a way to anticipate disparities in the final outcomes of an AI system by evaluating the learning training sets from the perspective of data quality (ISO IEC 25012). In particular, it has been observed how concepts of completeness, heterogeneity (Gini index), diversity (Shannon or Simpson index) or imbalance (imbalance ratio) can be used as predictive markers to highlight the risk that a data defect may propagate within the learning system.

Initially, [12] to analyze data quality issues in the learning data, Gini indices, imbalance ratios, Shannon,
Simpson indices were used. For fairness measures, independence and separation measures - consisting of the components True Positive Rate (TPR) and False Positive Rate (FPR) - were considered using the average of distances between probabilities as a criterion for synthesizing values (11).

The research revealed that the Gini index has good predictive capability for low values of the TPR component of Separation. The imbalance ratio indicator has good predictive capability for separation but not for independence. The Shannon index showed an acceptable level of prediction for the independence measure, excellent for the separation measure, but was completely ineffective for the FPR measure of separation. The Simpson index did not appear to be useful as a predictive bias measure.

The results were quite encouraging, so there was an attempt to improve the approach by acting on two fronts: the calculation method of fairness measures and the quality index of the input data to the learning system.

Regarding the calculation method for fairness measures, the use of a central tendency index could mask compensated errors, so three different approaches were attempted: using the maximum disparity between probability values (MinMax method [13]), using the distance between groups of similar probabilities (k-means and DBSCAN), and using mutual information.

As for the quality index selected in ISO IEC 25012, we chose the characteristic of completeness, particularly the concept of maximum completeness as defined in [10].

The study demonstrated that the use of maximum completeness and the MinMax measurement system provided the best predictive capability for fairness indices: independence, separation, sufficiency, and overall accuracy equality. Additionally, the use of the MinMax technique showed better sensitivity compared to mutual information and the DBSCAN clustering system, as shown in [13].

6. Conclusions

In the realm of AI systems, data governance and data quality are extremely important concepts. Since AI algorithms rely on learning datasets, the quality of input data can impact the outcomes. In this article, we have seen how completeness can serve as a good predictor of errors in the outputs of an ML system. In this context, it is clear that the definition of guidelines for the application of data governance and data quality in AI systems is crucial. Addressing bias in the data of technological systems is a significant challenge in the digital age, as the decisions made by algorithms can have substantial societal and personal implications, which can be measured according to international ISO/IEC standards.

References


