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

Alessandro Simonetta[1,2,*,†], Maria Cristina Paoletti[3,†] and Tsuyoshi Nakajima[4,†]

[1]Department of Enterprise Engineering, University of Rome Tor Vergata, Rome, Italy
[2]Italian Space Agency, via del Politecnico snc, Rome, Italy
[3]Professional Association of Italian Actuaries, Rome, Italy
[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]:

- ISO 31000:2018 Risk management — Guidelines [2]
- ISO/IEC 25000:2014 Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guide to SQuaRE [3]
- ISO/IEC 27002:2022 Information security, cybersecurity and privacy protection - Information security controls [4]
- ISO/IEC DIS 5259-2 Artificial Intelligence — Data Quality for Analysis and Machine Learning (ML) - Part 2: Data Quality Measures [5].

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 25010 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].

IWESQ 2023
*Corresponding author.
†These authors contributed equally.
alessandro.simonetta@gmail.com (A. Simonetta);
mariacristina.paoletti@gmail.com (M. C. Paoletti)
0000-0003-2002-9815 (A. Simonetta); 0000-0001-6850-1184 (M. C. Paoletti); 0000-0002-9721-4763 (T. Nakajima)
© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

CEUR-WS.org/Vol-3612/IWESQ_2023_Paper_03.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

Table 1

ISO/IEC 25010:2011	ISO/IEC 25059:2023*	
4.2 characteristics of the software product model	**AI sub-characteristics**	
Functional suitability	correctness	adaptability
Usability	controllability	transparency
Reliability	robustness	
Security	intervenability	
4.1 characteristics of the quality in use		
Satisfaction	transparency	transparency
Absence and mitigation of risks	ethical/social risk	

* in the process of being published

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)

$$P = \begin{bmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & \vdots & \vdots \\ p_{n1} & \cdots & p_{nn} \end{bmatrix} \tag{1}$$

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

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

$$TN(i) = \sum_{k=1, k\neq i}^{n} p_{kk} \tag{4}$$

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:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

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:

$$P(R = 1|A = a_i) = P(R = 1|A = a_j) \tag{8}$$

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:

$$\mathfrak{U}(a_i, a_j) = |P(R = 1|A = a_i) - P(R = 1|A = a_j)| \tag{9}$$

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
Probability for Sensitive Attribute Race

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

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:

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

In the literature, there are various clustering algorithms, with k-means and DBSCAN being used in [10]. A different approach to measuring fairness corresponds to the maximum disproportion in the values of joint probabilities (range or variability interval). Instead of measuring the distances between probabilities belonging to groups, we can calculate the difference between the maximum and minimum values (MaxMin algorithm). What has been discussed so far is applicable, without loss of generality, to other fairness measures such as separation, sufficiency, and overall accuracy equality. In all of these cases, the researcher is interested in identifying the presence of unfairness in a sensitive attribute A and assessing its magnitude based on a value within the range $\{0, 1\}$. However, if you calculate a fairness measure for each sensitive attribute A, you may discover that different treatment groups exist in relation to different indices. Since the original problem is to understand whether there are treatment differences in the values of sensitive attributes, rather than calculating a measure for each individual attribute, we can compute fairness measures for each value of the sensitive attribute. This way, we can construct a fairness vector with components being the fairness indices and examine the relationships between different vectors. In [9], a method was used to match treatment groups based on the Pearson correlation index.

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:

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

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, and

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 synthetizing 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

[1] A. Simonetta, A. Vetrò, M. C. Paoletti, M. Torchiano, Integrating square data quality model with iso 31000 risk management to measure and mitigate software bias, CEUR Workshop Proceedings 3114 (2021) pp. 17–22.

[2] International organization for standardization, "iso 31000:2018(en) risk management — guidelines", 2018. URL: https://www.iso.org/iso-31000-risk-management.html.

[3] International Organization for Standardization, "ISO/IEC 25000:2014 Systems and software engineering Systems and software Quality Requirements and Evaluation (SQuaRE) Guide to SQuaRE", 2014. URL: https://www.iso.org/standard/64764.html.

[4] International Organization for Standardization, "ISO/IEC 27002:2022 Information security, cybersecurity and privacy protection Information security controls", 2022. URL: https:/https://www.iso.org/standard/75652.html.

[5] International Organization for Standardization, "ISO/IEC DIS 5259-2 Artificial intelligence - Data quality for analytics and machine learning (ML) - Part 2: Data quality measures", Under development. URL: https://www.iso.org/standard/81860.html.

[6] International Organization for Standardization, "ISO/IEC 25010 Systems and software engineering Systems and software Quality Requirements and Evaluation (SQuaRE) - System and software quality models", 2011. URL: https://www.iso.org/standard/81860.html.

[7] International organization for standardization, "iso/iec 25059:2023 software engineering systems and software quality requirements and evaluation (square) - quality model for ai systems", 2023. URL: https://www.iso.org/standard/80655.html.

[8] S. Barocas, M. Hardt, A. Narayanan, Fairness and machine learning, 2020. URL: https://fairmlbook.org/, chapter: Classification.

[9] J. Larson, S. Mattu, L. Kirchner, J. Angwin, Compas recidivism dataset, 2016. URL: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[10] A. Simonetta, M. C. Paoletti, A. Venticinque, The use of maximum completeness to estimate bias in ai-based recommendation systems, CEUR Workshop Proceedings 3360 (2022) pp. 76–84.

[11] D. Steinberg, A. Reid, S. O'Callaghan, F. Lattimore, L. McCalman, T. S. Caetano, Fast fair regression via efficient approximations of mutual information, CoRR abs/2002.06200 (2020). URL: https://arxiv.org/abs/2002.06200.

[12] A. Vetrò, M. Torchiano, M. Mecati, A data quality

approach to the identification of discrimination risk in automated decision making systems, Government Information Quarterly 38 (2021) 101619. URL: https://www.sciencedirect.com/science/article/pii/S0740624X21000551. doi:https://doi.org/10.1016/j.giq.2021.101619.

[13] A. Simonetta, T. Nakajima, M. C. Paoletti, A. Venticinque, Fairness metrics and maximum completeness for the prediction of discrimination, CEUR Workshop Proceedings 3356 (2022) pp. 13–20.