NOVETTA

AI Performance Assessment Standardization in *ISO/IEC JTC 1/SC 42 Artificial intelligence* - Implications for Biometrics

International Face Performance Conference
27-29 October 2020

Mike Thieme
VP, Special Projects
Novetta

Introduction / CV

- Involved in developing performance testing standards since ~2004
  - Chairperson, INCITS / M1.5 Biometric Performance Testing and Reporting
  - Secretary, INCITS / Artificial Intelligence (AI)
- Editor of ISO/IEC biometric standards ...
  - **19795-2:2007** Biometric performance testing and reporting — Part 2: Testing methodologies for technology and scenario evaluation
  - **30107-3:2017** Biometric presentation attack detection — Part 3: Testing and reporting
  - **30107-4:2020** Biometric presentation attack detection — Part 4: Profile for testing of mobile devices
- Editor of ISO/IEC AI draft standard ...
  - **TS 4213** Assessment of Machine Learning Classification Performance
- Day job … lead R&D at Novetta, an advanced analytics company

About ISO/IEC JTC 1/SC 42 - Artificial intelligence

- Technical Committee formed in 2017
- Scope: Standardization in the area of Artificial Intelligence
  - Serve as the focus and proponent for JTC 1's standardization program on Artificial Intelligence
  - Provide guidance to JTC 1, IEC, and ISO committees developing Artificial Intelligence applications
- Absorbed WG 9 Big Data
- 31 participating, 16 observing members (countries)
- Full committee meets every 6 months
- ~300 meetings annually across all SC 42 working groups

SC 42 Structure (Working Groups)

- Working Group 1: Foundational Standards
  - 3 projects under development
- Working Group 2: Data
  - 5 published standards (from Big Data WG)
  - 5 projects under development
- **Working Group 3: Trustworthiness**
  - 1 published standard
  - 7 projects under development
- Working Group 4: Use cases and applications
  - 3 projects under development
- **Working Group 5: Computational approaches and computational characteristics of AI systems**
  - 3 projects under development

Program of work may have implications for biometrics

Total: 6 published standards, 21 projects under development

Projects with Biometric Implications: TS 4213

- **TS 4213: Assessment of Machine Learning Classification Performance**
- Describes best practices for comparing performance of ML models
  - Training data, GPU optimization, channel effects, test environment, cross-validation
- Specifies classification performance metrics
  - Binary classification: precision, recall, specificity, F1, ROC, PR AUC
  - Multi-class: macro-averaging, weighted-averaging, micro-averaging
  - Multi-label: hamming loss, exact match ratio, Jaccard index
  - Computational complexity and cost
- Specific statistical tests of significance
  - Paired Student's t-test, ANOVA, Kruskal-Wallis, Chi-squared test etc.
- Specifies reporting requirements
- Open issues: whether to consider performance more holistically; performance for models with dynamic retraining

Notional Roadmap: stable draft by 4/2021, publication by 10/2021

Projects with Biometric Implications: TR 24027

- **TR 24027: Bias in AI systems and AI-aided decision-making**
- Frames bias as distinct from fairness
- Describes potential sources of bias in AI systems
  - Based on human cognitive, data, model architecture, requirements
- Describes bias metrics via confusion matrix
  - Equality of odds, equality of opportunity, parity, predictive equality
- Discusses treatment of bias, including identification and mitigation
  - Data representation and labelling, transparency tools, training data, anti-bias adversarial methods, static analysis, internal and external validity testing
- Open issues: sources of bias as inputs vs. outcomes

DRAFT TEXT: For example, the designers of a facial recognition system might place importance on the face contour feature in their design and miss the fact that the contour might be (partially/completely) covered for people with particular cultural/religious backgrounds.

Notional Roadmap: stable draft by 10/2020, publication by 4/2021

Projects with Biometric Implications: TR 24029-1

- **TR 24029-1: Assessment of the robustness of neural networks - Overview**
- Discusses approaches to assessment of NN robustness, defined as ability of an AI system to maintain its level of performance under any circumstances
- Discusses statistical methods
  - Familiar to biometrics community - FPR, FNR, positive / negative predictive value
- Discusses formal methods
  - TR focus: interpolation stability, maximum stable space for perturbation resistance, uncertainty analysis, solvers, optimization techniques, abstract interpretation
- Discusses empirical methods
  - Field trials, *a posteriori* testing, NN benchmarking
  - Unfortunately, the TR asserts that empirical tests "rely on subjective observations"

DRAFT TEXT: Benchmarking measures the performance of a system on carefully designed datasets that are publicly available in most cases. Often they are used for testing different systems competitively. Prominent examples of benchmarking are the Face recognition vendor tests (FRVT) conducted by the US Department of Commerce .

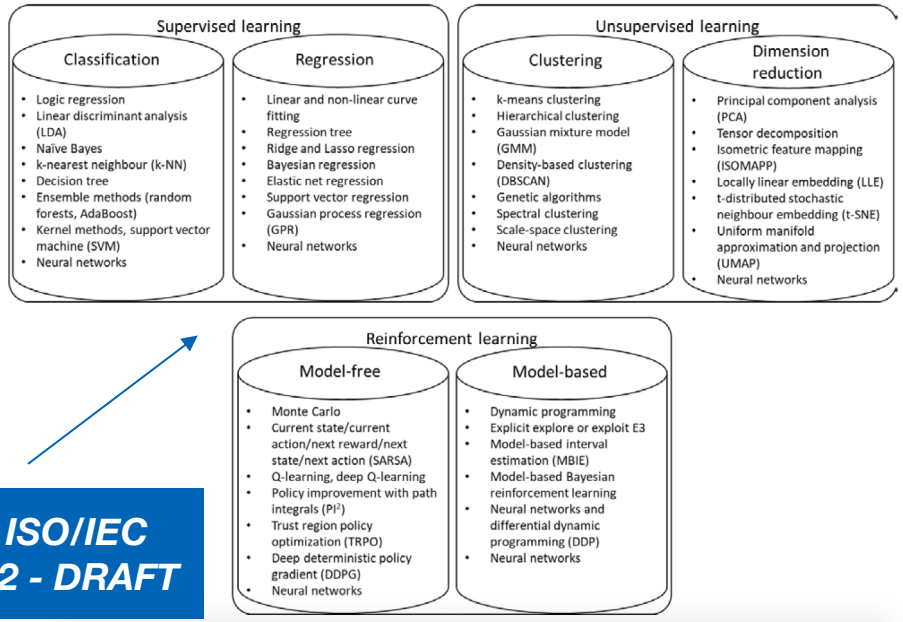
Roadmap: proceeding to publication as of 10/2020

Other SC 42 Projects To Be Aware of

- ISO/IEC TR 24028:**2020** Overview of trustworthiness in artificial intelligence
  - Discusses data poisoning, adversarial attacks, model stealing, hardware-focused threats to confidentiality and integrity, bias, unpredictability, opaqueness
- ISO/IEC CD 22989.2 Concepts and terminology **draft**
  - Many terms and definitions relevant to biometric practitioners
- ISO/IEC CD 23053.2 Framework for AI Systems Using ML **draft**
  - Conceptual material spans ML algorithms, architectures, optimization methods
- Objectives and methods for explainability of ML models and AI systems **new**
  - Broad interest in how to obtain visibility into deep neural networks

Final Considerations / Observations

- The line between artificial intelligence and machine learning is blurry
- Very little has been specified - to date, mostly considerations
- Despite pockets of highly specialized / technical activity (e.g. 24029-X), much of the focus is on governance, frameworks, process management
- Face recognition comes up often as a use case, but “biometrics” does not
- Need to distill AI or ML-specific elements (i.e. “what about this is specific to AI or ML?”)



Source: ISO/IEC 23053 CD2 - DRAFT