ISSN : 1226-7244 (Print)
ISSN : 2288-243X (Online)
논문번호 23-03-22
138
j.inst.Korean.electr.electron.eng.Vol.27,No.3,350~354,September 2023
https : //doi.org/10.7471/ikeee.2023.27.3.350

인공지능 시스템의 성능 평가 표준: ISO/IEC TR 24029-1

Evaluation Standard for Performance of Artificial Intelligence Systems: ISO/IEC TR 24029-1

이 성 수*★

Seongsoo Lee*★

Abstract

This paper describes ISO/IEC TR 24029-1, an international standard to evaluate the performance of artificial intelligence systems. ISO/IEC TR 24029-1 defines the performance measures of artificial intelligence systems in two categories, i.e. interpolation and classificiation. Performance measures in the interpolation categories mean how much the predicted values of the artificial intelligence system is close to the real values. Performance measures in the classification categories mean how much the predicted classes of the artificial intelligence system is equal to the real classes. Based on these performance measures, performance of artificial intelligence systems can be evaluated and performance of different artificial intelligence systems can be compared.

요 약

본 논문에서는 인공지능 시스템의 성능을 평가하기 위해 새로 개발된 국제 표준인 ISO/IEC TR 24029-1에 대해 다룬다. ISO/IEC TR 24029-1에서는 인공지능 시스템의 성능 지표를 Interpolation과 Classification의 두 가지 카테고리로 나누어 규정한다. Interpolation 카테고리에 해당하는 성능 지표는 인공지능 시스템이 예측한 값이 실제 값과 얼마만큼 가까운지 그 성능을 평가하는 지표이며 Classification 카테고리에 해당하는 성능 지표는 인공지능 시스템이 분류한 종류가 실제 종류와 얼마만큼 일치하는지 그 성능을 평가하는 지표이다. 이들 지표를 사용하면 인공지능 시스템의 성능을 평가하고 서로 다른 인공지능 시스템의 성능을 비교할 수 있다.

Key words : ISO/IEC TR 24029-1, Performance Evaluation, Artificial Intelligence, Interpolation, Classification

Ⅰ. 서론

최근 소프트웨어 기술과 하드웨어 기술의 급격한 발달에 힘입어 인공지능은 눈부시게 발전하고 있으며 2016년 AlphaGo, 2022년 ChatGPT의 개발은 지금까지 인간만이 할 수 있다고 여겨지던 여러 행위를 인공지능도 할 수 있음을 보여주었다.

이렇게 인공지능이 우리 사회 전반에 큰 영향을 미치면서 인공지능 시스템의 성능을 평가하기 위한 지표를 마련할 필요성이 높아지고 있으며, 특히 여러 인공지능

* School of Electronic Engineering and Department of Intelligent Semiconductor, Soongsil University (Professor)
★ Corresponding author
E-mail : sslee@ssu.ac.kr, Tel : +82-2-820-0692
※ Acknowledgment
This work was supported by the R&D Program of the Ministry of Trade, Industry, and Energy (MOTIE) and Korea Evaluation Institute of Industrial Technology (KEIT). (20023805, RS-2022-00155731, RS-2023-00232192)
Manuscript received Aug. 28, 2023; revised Sep. 7, 2023; accepted Sep. 18, 2023.

시스템을 비교하기 위해서 표준화된 성능 지표가 절실하게 요구되고 있다.

ISO(International Standard Organization), IEC (International Electrotechnical Commission), IEEE (Institute of Electrical and Electronics Engineers), DIN(German Institute for Standardization) 등 많은 국제 표준화 기관이 인공지능 시스템의 성능 평가를 위해 다양한 표준을 발표한 바 있다.

ISO/IEC AWI TR 42106[1]은 인공지능 시스템의 품질 특성에 대한 벤치마킹 표준을 제정 중이다. ISO/IEC 25059[2]는 인공지능 시스템의 품질 모델에 대한 표준을 제정하였으며 ISO/IEC TR 29119-11[3]은 인공지능 시스템의 특성 평가를 포함한 테스트 방법에 대한 표준을 제정하였다. DIN SPEC 92001-2[4]는 인공지능 시스템의 품질 평가 메타 모델과 품질 요구 조건에 대한 표준을 제정하였으며 IEEE 2937[5]은 인공지능 서버의 성능에 대한 벤치마킹 표준을 제정하였다.

그러나 이들 표준은 대부분 구체적인 성능 지표를 제시하지 않고 품질을 평가하는 모델과 절차를 제시하였거나[1]-[4] 인공지능 시스템 전체가 아닌 서버에 국한된 성능 지표를 제시하였을 뿐이기[5] 때문에 일반적인 인공지능 시스템의 성능을 평가하고 비교하기 위한 성능 지표로 사용하기에는 다소 무리가 있다.

ISO/IEC TR 24029-1[6]은 이들 표준과 다르게 인공지능 시스템의 성능 지표를 정량화하였기 때문에 다양한 인공지능 시스템의 성능을 객관적인 수치로 나타내고 비교할 수 있다. 본 논문에서는 ISO/IEC TR 24029-1에서 규정한 인공지능 시스템의 여러 가지 성능 지표를 자세히 살펴본다.

II. 인공지능 시스템의 성능 지표

ISO/IEC TR 24029-1에서는 인공지능 시스템의 출력 종류를 Interpolation과 Classification의 두 가지 카테고리로 구분한다. Interpolation은 인공지능 시스템이 수치를 예측하여 출력하는 경우이며 Classification은 인공지능 시스템이 종류를 분류하여 출력하는 경우이다. 이에 따라 ISO/IEC TR 24029-1에서는 인공지능 시스템의 출력이 Interpolation인 경우와 Classification인 경우를 구분하여 성능 평가 척도를 다음과 같이 제시하였다.

1. Interpolation

인공지능 시스템이 수치를 예측하여 출력하는 경우의 성능은 결국 인공지능 시스템이 예측한 값이 실제와 얼마나 가까운지를 의미한다. ISO/IEC TR 24029-1에서는 Root Mean Square Error, Max Error, Actual/Predicted Correlation의 3개의 성능 지표를 제시하였다.

(1) Root Mean Square Error(RMSE)

RMSE는 예측 오차(=실제값-예측값)의 표준편차를 말하며 식 (1)과 같이 주어진다. RMSE가 작다는 것은 인공지능 시스템이 예측한 값과 실제 값의 차이가 작다는 것을 의미한다.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(predicted_i - actual_i)^2}{N}} \quad (1)$$

(2) Max Error(MAXERR)

MAXERR는 예측 오차(=실제값-예측값)의 최대값을 말하며 식 (2)와 같이 주어진다. MAXERR가 작다는 것은 인공지능 시스템이 예측한 값과 실제 값의 최대 차이가 작다는 것을 의미한다.

$$MaxError = \max_{i=1}^{N} |predicted_i - actual_i| \quad (2)$$

(3) Actual/Predicted Correlation(PCC)

PCC는 실제값과 예측값의 Linear Correlation Coefficient, 흔히 Pearson Correlation Coefficient라고 불리는 값을 말하며 식 (3)과 같이 주어진다. PCC는 -1에서 1 사이의 값을 가지며, 이 값이 1에 가까울수록 예측값과 실제값의 상관관계가 높아서 예측값이 실제값과 거의 정비례하고 실제값과 근접한 주위에 모여서 분포한다는 것을 의미한다. ISO/IEC TR 24029-1에 명시되지는 않았지만 PCC에 따른 실제값과 예측값의 관계는 일반적으로 표 1과 같이 해석된다.

$$PCC = \frac{\left\{N \times \sum_{i=1}^{N}(predicted_i \times actual_i)\right\} - \left\{(\sum_{i=1}^{N} predicted_i) \times (\sum_{i=1}^{N} actual_i)\right\}}{\sqrt{N \times \sum_{i=1}^{N}(predicted_i)^2 - (\sum_{i=1}^{N} predicted)^2} \times \sqrt{N \times \sum_{i=1}^{N}(actual_i)^2 - (\sum_{i=1}^{N} actual)^2}} \quad (3)$$

Table 1. Association between actual value and predicted value according to PCC value.
표 1. PCC 값에 따른 실제값과 예측값의 관계

PCC Value	Association between actual value and predicted value
+1.0	Perfect positive association
+0.8 ~ +1.0	Very strong positive association
+0.6 ~ +0.8	Strong positive association
+0.4 ~ +0.6	Moderate positive association
+0.2 ~ +0.4	Weak positive association
0.0 ~ +0.2	Very weak positive or no association
-0.2 ~ 0.0	Very weak negative or no association
-0.4 ~ -0.2	Weak negative association
-0.6 ~ -0.4	Moderate negative association
-0.8 ~ -0.6	Strong negative association
-1.0 ~ -0.8	Very strong negative association
-1.0	Perfect negative association

Table 2. NT+, NT-, NF+, and NF- in the confusion matrix.
표 2. 혼동 매트릭스에서의 NT+, NT-, NF+, NF-

Confusion Matrix		Actual Classification	
		Positive	Negative
Predicted Classification	Positive	True Positive N_{T+}	False Positive N_{F+}
	Negative	False Negative N_{F-}	True Negative N_{T-}

2. Classification

인공지능 시스템이 종류를 분류하여 출력하는 경우의 성능은 결국 인공지능 시스템이 분류한 종류가 실제와 얼마나 일치하는지를 의미한다.

분류를 목적으로 하는 인공지능 시스템은 흔히 분류 결과를 진양성(True Positive), 진음성(True Negative), 위양성(False Positive), 위음성(False Negative)으로 나누어 이를 기반으로 각종 성능 지표를 계산하는 혼동 매트릭스[7]를 사용한다. ISO/IEC TR 24029-1에서도 혼동 매트릭스에 기반하여 식 (4)~(16)과 같이 13개의 성능 지표를 제시하였다. 식 (4)~(16)에서 사용되는 파라미터인 N_{T+}, N_{T-}, N_{F+}, N_{F-}는 표 2와 같다.

1. Sensitivity

$$Sensitivity(R_{T+}) = \frac{N_{T+}}{N_{T+} + N_{F-}} \tag{4}$$

2. Specificity

$$Specificity(R_{T-}) = \frac{N_{T-}}{N_{T-} + N_{F+}} \tag{5}$$

3. Miss Rate

$$Miss\ Rate(R_{F-}) = \frac{N_{F-}}{N_{T+} + N_{F-}} \tag{6}$$

4. Fall-Out

$$Fall-Out(R_{F+}) = \frac{N_{F+}}{N_{T-} + N_{F+}} \tag{7}$$

5. Accuracy

$$Accuracy = \frac{N_{T+} + N_{T-}}{N_{T+} + N_{T-} + N_{F+} + N_{F-}} \tag{8}$$

6. Precision

$$Precision(V_{P+}) = \frac{N_{T+}}{N_{P+}} \tag{9}$$

7. Separation Ability

$$Separation\ Ability(V_{P-}) = \frac{N_{T-}}{N_{P-}} \tag{10}$$

8. False Discovery Rate

$$False\ Discovery\ Rate = \frac{N_{F+}}{N_{P+}} \tag{11}$$

9. False Omission Rate

$$False\ Omission\ Rate = \frac{N_{F-}}{N_{P-}} \tag{12}$$

10. Positive Likelihood Relation

$$Positive\ Likelihood\ Relation(R_{L+}) = \frac{R_{T+}}{R_{F+}} \tag{13}$$

11. Negative Likelihood Relation

$$Negative\ Likelihood\ Relation(R_{L-}) = \frac{R_{F-}}{R_{T-}} \tag{14}$$

12. Diagnostic Odds Rate

$$Diagnostic\ Odds\ Rate = \frac{R_{L+}}{R_{L-}} \tag{15}$$

13. F1 Score

$$F1\ Score = \frac{2}{\frac{1}{R_{T+}} + \frac{1}{V_{P+}}} \tag{16}$$

Ⅲ. 인공지능 시스템의 성능 평가 및 최적화 절차

ISO/IEC TR 24029-1에서는 인공지능 시스템의 성능 평가 및 최적화 절차를 그림 1과 같이 규정하고 있다.

1. State Robust Goals: 인공지능 시스템의 성능 지표를 정의하고 목표 성능을 설정한다.
2. Plan Testing: 인공지능 시스템을 테스트하기 위한 방법과 절차를 결정한다.

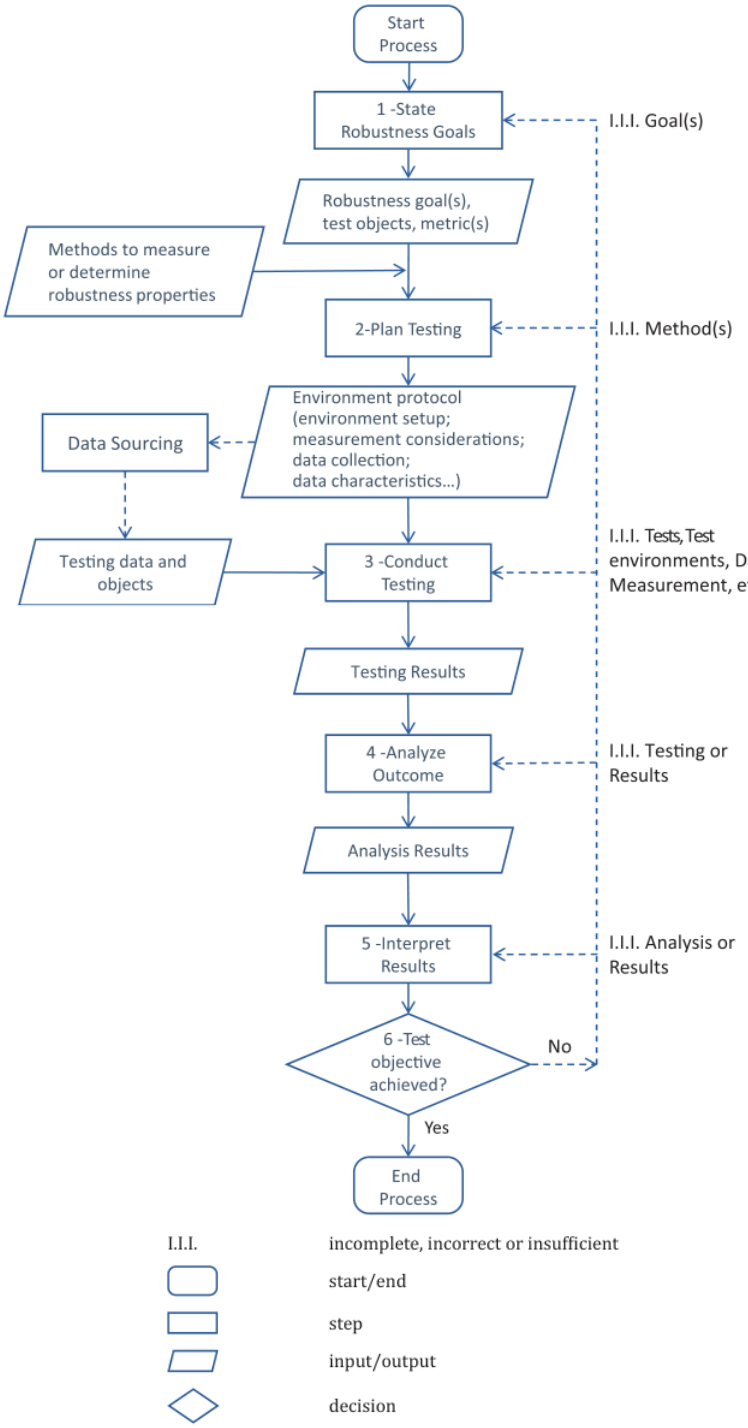

Fig. 1. Performance evaluation and optimization process of artificial intelligence system [6]
그림 1. 인공지능 시스템의 성능 평가 및 최적화 절차[6]

3. Conduct Testing: 인공지능 시스템을 테스트하는 동작을 수행한다.
4. Analyze Outcome: 인공지능 시스템의 테스트 결과를 분석한다.
5. Interpret Results: 인공지능 시스템의 테스트 결과로부터 성능 지표를 계산한다.
6. Test Object Achieved?: 인공지능 시스템의 성능 지표가 목표를 달성하였는지 확인하고 미달하는 경우 개선을 수행한 후 테스트를 재수행한다.

Ⅳ. 결론

본 논문에서는 인공지능 시스템의 성능 평가 지표를 규정한 국제 표준인 ISO/IEC TR 24029-1에 대해 살펴

보았다. 인공지능 시스템의 품질 평가에 관련된 다른 국제 표준에서는 별도로 성능 평가 지표를 규정하지 않았지만 ISO/IEC TR 24029-1는 다양한 성능 평가 지표를 구체적으로 규정하였기 때문에 다양한 인공지능 시스템의 성능을 평가하고 비교하기 위해 매우 유용할 것으로 생각된다.

References

[1] ISO/IEC AWI TR 42106, "Information Technology - Artificial Intelligence - Overview of Differentiated Benchmarking of AI System Quality Characteristics," https://www.iso.org/standard/86903.html

[2] ISO/IEC 25059:2023, "Software Engineering - Systems and Software Quality Requirements and Evaluation (SQuaRE) - Quality Model for AI Systems," https://www.iso.org/standard/80655.html

[3] ISO/IEC TR 29119-11:2020, "Software and Systems Engineering - Software Testing - Part 11: Guidelines on the Testing of AI-based Systems," https://www.iso.org/standard/79016.html

[4] DIN SPEC 92001-2, "Artificial Intelligence - Life Cycle Processes and Quality Requirements - Part 2: Robustness," https://www.din.de/en/wdc-beuth:din21:330011015

[5] IEEE 2937-2022, "IEEE Standard for Performance Benchmarking for Artificial Intelligence Server Systems," https://standards.ieee.org/ieee/2937/10376

[6] ISO/IEC TR 24029-1:2021, "Artificial Intelligence (AI) - Assessment of the Robustness of Neural Networks - Part 1: Overview," https://www.iso.org/standard/77609.html

[7] J. Ha, J. Seo, and S. Lee, "Living Lab and Confusion Matrix for Performance Improvement and Evaluation of Artificial Intelligence System in Life Environment," *j.inst.Korean. electr.electron.eng.*, vol.24, no.4, pp.1180-1183, 2020.
DOI: 10.7471/ikeee.2020.24.4.1180

BIOGRAPHY

Seongsoo Lee (Life Member)

1991 : BS degree in Electronic Engineering, Seoul National University.
1993 : MS degree in Electronic Engineering, Seoul National University.
1998 : PhD degree in Electrical Engineering, Seoul National University.
1998~2000 : Research Associate, University of Tokyo
2000~2002 : Research Professor, Ewha Womans University
2002~Now : Professor in School of Electronic Engineering and Department of Intelligent Semiconductor, Soongsil University
〈Main Interest〉 AI SoC, Automotive SoC, Security SoC, Processor SoC, Power Management SoC, Battery Management SoC, Reliability and Safety