ISSN : 1226-7244 (Print)
ISSN : 2288-243X (Online)
논문번호 20-04-41
260

j.inst.Korean.electr.electron.eng.Vol.24,No.4,1180~1183,December 2020
http://dx.doi.org/10.7471/ikeee.2020.24.4.1180

생활 환경에서의 인공지능 시스템 성능 개선 및 평가를 위한 리빙랩 및 혼동 매트릭스

Living Lab and Confusion Matrix for Performance Improvement and Evaluation of Artificial Intelligence System in Life Environment

하 지 원*, 서 지 석*, 이 성 수**★

Ji-Won Ha*, Ji-Seok Seo*, and Seongsoo Lee**★

Abstract

Recently, the daily life safety detection functionalities such as fall accident detection and burn danger detection are widely disseminated along with the development of IoT and smart home. These safety detection functionalities are mostly performed by artificial intelligence. However, simple accuracy measurement of the safety detection in laboratory environment is often far from practical performance in daily life environment. To mitigate this problem, this paper introduces two techniques, i.e. living lab and confusion matrix. Living lab is more than simple simulation of daily life environment, and it enables users to directly participate technology development and product design. Various performance measures induced from confusion matrix significantly help to evaluate the performance of artificial intelligence system for proper application purposes.

요 약

최근 들어 IoT와 스마트홈의 발전에 따라 낙상 사고 감지, 화상 위험 감지와 같이 일상 생활에서의 안전 감지 기능이 많이 보급되기 시작했다. 이러한 안전 감지 기능은 대부분 인공지능에 의해 수행된다. 그러나 실험실 환경에서 안전 감지의 정확도만 평가하는 경우에는 실제로 일상 생활 환경에서 체감하게 되는 성능과 꽤 큰 차이를 보이는 경우가 많다. 본 논문에서는 이러한 문제점을 보완하기 위해 사용하는 두 가지 기법인 리빙랩과 혼동 매트리스를 소개한다. 리빙랩은 단순히 일상 생활 환경의 모사를 넘어서 사용자가 직접 기술 개발 및 제품 설계에 참여할 수 있는 통로가 된다. 또한 혼동 매트리스에서 도출되는 다양한 성능 척도는 사용 목적에 적합하게 인공지능 시스템의 성능을 평가하는데 큰 도움을 준다.

Key words : Living Lab, Confusion Matrix, Artificial Intelligence, Life Environment, Performance Evaluation

* Korea Conformity Laboratory (Researcher, Researcher)
** Soongsil University (Professor)
★ Corresponding author
E-mail : sslee@ssu.ac.kr, Tel : +82-2-820-0692
Manuscript received Dec. 15, 2020; accepted Dec. 19, 2020.

Ⅰ. 서론

최근 들어 IoT와 스마트홈이 활발하게 보급되면서, 노인 생활자의 낙상 사고 감지, 영유아의 화상 감지 등과 같이 일상 생활에서의 안전 감지 기능이 많이 보급되기 시작했다. 이러한 안전 감지 기능은 대부분 다양한 센서의 측정값을 인공지능으로 분석하고 판단하는 형태로 수행된다.

그러나 이러한 연구 중 상당수는 대부분 실험실

에서 측정한 값으로 인공지능을 학습시키는데다가, 인식률과 같은 단순 성능 척도를 사용하기 때문에 실제 일상 생활에서 체감하는 성능과 상당한 차이를 보이는 경우가 자주 발생한다. 본 논문에서는 이러한 문제점을 해결하기 위해 인공지능 분야에서 많이 사용하는 기법인 리빙랩(Living Lab)과 혼동 매트리스(Confusion Matrix)를 간단히 소개하고, 이러한 기법들이 어떻게 인공지능의 성능을 개선시킬 수 있는지 설명한다.

No	Family	House
1	Adult 2, Infant 2	Bedroom 3, Living Room 1, Toilet 2
2	Aged 1	Bedroom 4, Living Room 1, Toilet 2
3	Aged 2	Bedroom 3, Living Room 1, Toilet 1
4	Adult 2, Infant 1	Bedroom 4, Living Room 1, Toilet 2
5	Aged 1	Bedroom 1, Toilet 1
6	Adult 2, Infant 1	Bedroom 3, Living Room 1, Toilet 1
7	Adult 2, Infant 2	Bedroom 3, Living Room 1, Toilet 2
8	Adult 2, Infant 1	Bedroom 3, Living Room 1, Toilet 2
9	Adult 2, Infant 2	Bedroom 3, Living Room 1, Toilet 2
10	Adult 2, Infant 2	Bedroom 4, Living Room 1, Toilet 2

Fig. 1. Living lab to develop detection technologies for burn danger of infant and fall accident of aged.
그림 1. 영유아의 화상 위험 및 노령자의 낙상 사고 감지 기술 개발을 위해 구성된 리빙랩

II. 리빙랩

리빙랩[1][2]은 '사용자 참여형 혁신공간'이라는 뜻으로 사용자 주도형 혁신 모델, 정부・민간・시민 간의 파트너십, 과학・사회・현장의 통합 모델을 시도하는 방법을 말한다[3]. 즉, 사용자가 실제 생활하는 공간에서 데이터 수집, 분석, 실증을 수행하고, 동시에 사용자의 참여를 통해 연구의 방향과 내용을 최적화하는 기법이다.

리빙랩의 장점으로는 (1) 사용자가 생활하는 공간에서 데이터 수집, 분석, 실증을 수행하기 때문에 실제와 유사하고 정확한 결과를 얻을 수 있고, (2)

Fig. 2. SCAPE living lab in Europe [4].
그림 2. 유럽의 iSCAPE 리빙랩[4]

사용자의 참여가 연구의 내용과 방향에 적극적으로 반영되기 때문에 보다 실용적이고 유용한 연구 결과를 얻을 수 있다는 점을 들 수 있다.

그림 1은 영유아의 화상 위험 및 노령자의 낙상 사고를 감지하기 위해 구성된 리빙랩인데, 영유아 및 노령자가 포함된 실제 가구에 측정 장비를 설치하여 데이터를 수집하고 인공지능을 학습시킴으로서 실제와 유사한 성능을 얻을 수 있었다.

많은 경우 리빙랩은 도시 전체 수준의 대규모로 구축되기도 한다. 그림 2는 유럽의 대기 오염을 줄이기 위해 6개국 6개 도시(이탈리아 볼로냐, 독일 보트롭, 아일랜드 더블린, 영국 길포드, 벨기에 하셀트, 핀란드 반타)에 구축된 iSCAPE 리빙랩으로,

동일한 하드웨어와 소프트웨어 플랫폼으로 산소, 오염가스, 소음, 습도, 기온, 조도, 기압, 자외선, 미세먼지 등을 측정한다. 수집된 데이터는 인터넷을 통해 세계 어디서든지 모니터링하고 데이터를 다운받아서 분석이 가능하다. 각 도시는 이들 데이터를 인공지능으로 분석하여 시 당국과 시민단체에 제공하고, 시 당국과 시민단체는 이를 바탕으로 도시 운영 정책을 결정한다. 시행된 정책의 효과는 곧바로 리빙랩에서 측정되며, 이를 통해 사용자 참여, 정책 결정, 연구 개발의 선순환 구조를 구축하였다.

III. 혼동 매트릭스

인공지능이 센서 시스템을 통해 어떤 사건을 감지할 때 오류가 발생할 수 있다. 일반적으로 센서 시스템의 감지 성능은 인식률같은 간단한 성능 척도를 사용하지만 인공지능과 같이 판단을 해야 하는 경우에는 미처 예상하지 못한 다양한 상황이 발생할 수 있다.

예를 들어 환자의 신체 증상을 다각적으로 모니터링한 다음에 인공지능으로 코로나-19 감염 여부를 판단하는 시스템을 개발했다고 가정한다. 이 시스템은 감염자를 비감염자로 잘못 판단하는 확률과 비감염자를 감염자로 잘못 판단하는 확률이 모두 1%라고 가정한다. 감염자로 판단된 사람은 즉시 격리되고 코로나-19 치료제를 투여받게 되지만, 격리 시의 노동 손실과 치료제의 부작용이 상당하여 건강한 사람에게는 가능한 한 격리와 치료제를 피하는 것이 바람직하다. 그래도 이 시스템의 성능은 매우 우수해 보인다.

그러나 인구 10만명인 나라에서 실제 코로나-19에 감염된 사람의 비율이 1%라고 가정했을 때 실제 감염자와 비감염자의 수는 각각 1,000명과 99,000명이며, 이 시스템은 감염자 1,000명 중에서 990명은 감염자로 제대로 판단하지만 10명은 비감염자로 잘못 판단하고, 비감염자 99,000명 중에서 98,010명은 비감염자로 제대로 판단하지만 990명은 감염자로 잘못 판단한다. 즉, 이 시스템이 감염자로 판단한 사람은 1,980명이고 이 중 절반은 실제로는 비감염자를 잘못 판단하여 필요없는 격리와 치료를 수행하게 된다. 즉, 인공지능과 같은 복잡한 판단에서 단순히 인식률과 같은 척도를 사용하면 실제로는 사용이 불가능할 정도로 성능이 나쁜데도 이를 모르고 지나칠 수 있다. 이를 보완하기 위해 인공지능에서는 혼동 매트리스[5]를 기반으로 한 다양한 성능 척도를 사용한다.

혼동 매트릭스는 측정 데이터를 표 1과 같이 진양성(TP : True Positive), 진음성(TN : True Negative), 위양성(FP : False Positive), 위음성(FN : False Negative)로 나누고, 이들을 조합하여 다양한 성능 척도를 도출한다[5][6]. 실제로 인공지능 시스템의 평가에 사용되는 성능 척도는 매우 다양하지만 가장 많이 쓰이는 것은 다음의 여섯 가지를 들 수 있다.

Table 1. Confusion matrix [5].
표 1. 혼동 매트릭스[5]

		Actual	
		Positive	Negative
Predicted	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

Table 2. Confusion matrix calculated in Corona-19 example.
표 2. 코로나-19의 예에서 계산된 혼동 매트릭스

		Actual	
		Positive	Negative
Predicted	Positive	TP = 990	FP = 990
	Negative	FN = 10	TN = 98,010

(1) TPR(True Positive Rate) 또는 Sensitivity : 실제로 사건이 발생했을 때 이를 정확히 감지하는 비율을 나타내는 척도

$$TPR = \frac{TP}{TP + FN}$$

(2) TNR(True Negative Rate) 또는 Specificity : 실제로 사건이 발생하지 않았을 때 이를 정확히 감지하는 비율을 나타내는 척도

$$TNR = \frac{TN}{TN + FP}$$

(3) PPV(Positive Predictive Value) 또는 Precision : 사건 발생을 감지했을 때 실제로 사건이 발생한 비율을 나타내는 척도

$$PPV = \frac{TP}{TP + FP}$$

(4) NPV(Negative Predictive Value) : 사건 미발생을 감지했을 때 실제로 사건이 미발생한 비율을 나타내는 척도

$$NPV = \frac{TN}{TN + FN}$$

(5) ACC(Accuracy) : 사건의 발생 여부를 시스템이 정확히 감지하는 성능을 나타내는 척도

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

(6) F1(F1 Score) : 시스템의 전체적인 성능을 나타내는 척도로 TPR과 PPV의 조화 평균으로 계산됨

$$F1 = \frac{2}{\frac{1}{TPR} + \frac{1}{PPV}} = \frac{2 \times TP}{2 \times TP + FP + FN}$$

위의 코로나-19 예에서 이들 척도를 계산해보면 TPR=0.99, TNR=0.99, PPV=0.50, NPV=0.50, ACC=0.99, F1 Score=0.66이 되며, TPR은 우수하지만 PPV가 나빠서 전체적인 성능을 나타내는 척도인 F1이 나빠지는 것을 볼 수 있다.

이들 척도를 좀 더 자세히 살펴보면 다음과 같다. TPR은 실제 양성을 양성으로 판단하는 확률, 즉 진양성을 찾아내는 성능을 나타내며, TNR은 실제 음성을 음성으로 판단하는 확률, 즉 진음성을 찾아내는 성능을 나타낸다. ACC는 실제 양성을 양성으로 판단하고 실제 음성을 음성으로 판단하는 확률, 즉 진양성과 진음성을 찾아내는 성능을 의미한다. 예를 들어 조기 검사를 통한 암 세포 발견의 경우, 실제 암 세포를 암 세포라고 판단하는 경우(= 진양성)가 가장 중요하므로 TPR이 가장 중요하며, 신원 확인을 위한 지문 인식의 경우, 실제 다른 사람인 지문을 다른 사람의 것으로 판단하는 경우(= 진음성)가 가장 중요하므로 TNR이 중요하다.

이에 비해 PPV는 양성으로 판단한 결과가 맞는 확률, 즉 위양성이 적게 발생하는 성능을 의미하며 NPV는 음성으로 판단한 결과가 맞는 확률, 즉 위음성이 적게 발생하는 성능을 의미한다. TPR과 PPV가 모두 높아야 F1이 높아지기 때문에 인공지능 시스템의 전체적인 성능을 나타내는 데는 F1이 많이 사용된다. 즉 F1이 낮으면 TPR과 PPV 중에서 하나가 낮은 것이며 이는 해당 인공지능 시스템이 진양성을 잘 찾아내지 못하거나 위양성이 많이 발생한다는 것을 의미한다.

위의 코로나-19의 예에서 진양성인 환자를 최대한 찾아내서 격리하고 치료하는 것이 중요하다면 TPR이 높아야 하며, 위양성인 환자가 불필요한 격리나 치료를 받지 않도록 하는 것이 중요하다면 PPV가 높아야 한다. 따라서 인공지능 시스템의 성능을 평가하기 위해서는 어플리케이션의 목적에 맞는 성능 척도를 잘 선택하여야 한다.

Ⅳ. 결론

본 논문에서는 인공지능 시스템이 실험실 환경에서만 학습되어 발생할 수 있는 문제점과 인공지능 시스템을 단순한 인식률만 가지고 평가했을 때 생길 수 있는 문제점을 줄이기 위해 사용되는 두 가지 기법인 리빙랩과 혼동 매트릭스를 설명하였다. 생활 환경에서 사용되는 인공지능 시스템은 우리의 삶과 밀접한 관련성을 가지고 있기 때문에 이러한 기법 등을 사용하여 더욱 정교화할 필요가 있다.

References

[1] N. Norbert, L. Sevrin, and B. Massot, “From Health Smart Homes to Living Labs for Health,” *Proceedings of IEEE International Conference on e-Health Networking, Applications, and Services*, pp.164-169, 2015.
DOI: 10.1109/HealthCom.2015.7454492

[2] J. Colomer et al, “Experience in Evaluating AAL Solutions in Living Labs,” *Sensors,* vol.14, no.4, pp.7277-7311, 2014. DOI: 10.3390/s140407277

[3] J. Seong et al, “Current Status of Korean Living Labs and Its Development Plan,” *STEPI Annual Research Topics,* 2017.

[4] https://www.iscapeproject.eu/iscape-living-labs/

[5] K. Ting, “Confusion Matrix,” *Encyclopedia of Machine Learning and Data Mining, Springer,* 2017.

[6] T. Octaviani, Z. Rustam, and T. Siswantining, “Ovarian Cancer Classification Uusing Bayesian Logistic Regression,” *IOP Conference Series: Materials Science and Engineering,* vol.546, pp. 1-7, 2019. DOI: 10.1088/1757-899X/546/5/052049