SCIENCE CHINA
Information Sciences

• LETTER •

December 2024, Vol. 67, Iss. 12, 229101:1–229101:2
https://doi.org/10.1007/s11432-023-4239-4

Multi-dimensional ability diagnosis for machine learning algorithms

Qi LIU[1], Zheng GONG[1], Zhenya HUANG[1], Chuanren LIU[2], Hengshu ZHU[3], Zhi LI[4], Enhong CHEN[1*] & Hui XIONG[5]

[1] *State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China, Hefei* 230026*, China;*
[2] *Business Analytics and Statistics, The University of Tennessee, Knoxville* 37934*, USA;*
[3] *Computer Network Information Center, Chinese Academy of Sciences, Beijing* 100083*, China;*
[4] *Shenzhen International Graduate School, Tsinghua University, Shenzhen* 518055*, China;*
[5] *Hong Kong University of Science and Technology* (*Guangzhou*)*, Guangzhou* 511453*, China*

Received 2 September 2023/Revised 27 January 2024/Accepted 10 June 2024/Published online 13 December 2024

A significant proportion of noticeable improvement in machine learning architectures actually benefits from the consistent inspiration of the way human learning [1]. For instance, curriculum learning [2] is inspired by highly organized human education systems, i.e., training the algorithms with easy samples first and gradually transforming to the hard examples can contribute to faster convergence and lower generalization error. Similarly, the evaluation of machine learning (ML) algorithms (including large language models) may also benefit from the more comprehensive and fine-grained measurement of human learning. To this end, in this study, inspired by the psychometric theories from human measurement [3], we propose a general cognitive diagnosis framework for machine learning algorithm evaluation (Camilla). Under this framework, a multi-dimensional diagnostic metric Ability is defined for collaboratively measuring the multifaceted strength of each algorithm. Specifically, given the response logs from different well-trained algorithms to data samples, we leverage cognitive diagnosis assumptions [1] and neural networks to learn the complex interactions among algorithms, samples, and the required skills (e.g., explicit category in Figure 1(b) or latent factors) for algorithms correctly responding to each sample. For simplicity, we call ML algorithms to be evaluated as learners, and the cognitive diagnosis method as diagnoser. Our goal in developing the diagnoser is to estimate the learners' multi-dimensional Ability on specific skills. The definition of Ability and skills is defined as follows.

Definition 1. Ability is a multi-dimensional metric for quantifying the proficiency of learners on specific skills. Each entry represents one skill, and the skill can be explicitly given or pre-defined as latent factors.

Each learner after training may perform well on some of the samples (or skills) while performing poorly on others. Similarly, one learner may outperform another learner on the part of the samples (or skills) while failing on the other samples. Therefore, we propose a multi-dimensional diagnostic metric Ability to quantify such internal and external performance differences. Note that, our Ability is not simply a skill-specific version of the traditional metrics like Accuracy, because we also collaboratively consider the learning difficulty and discrimination of different data samples. Here, the discrimination indicates the capability of samples to differentiate the proficiencies of learners. Finally, Ability needs to follow the monotonicity assumption.

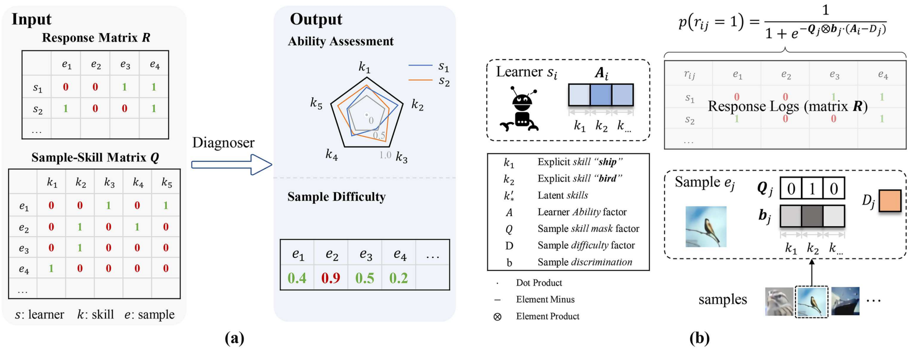
Assumption 1. The probability of correct response to the sample is monotonically increasing with the Ability of learners. Conversely, this probability for one learner is monotonically decreasing with the sample difficulty.

To design our task-agnostic diagnoser, we consider a well-trained learner set $S = \{s_1, \dots, s_N\}$ and a data sample set $E = \{e_1, \dots, e_M\}$. After running each learner s_i on E we can get response logs R, denoted as a set of triple (s_i, e_j, r_{ij}), where $s_i \in S, e_j \in E$, and r_{ij} is the response score. Specifically, in the case of classification tasks, $r_{ij} = 1$ if learner s_i answers the class label of sample e_j correctly and $r_{ij} = 0$ otherwise. Meanwhile, an explicitly or implicitly pre-defined sample-skill relevancy matrix Q should also be given. $Q = \{Q_{ij}\}_{M \times K}$, where K is the number of skills, $Q_{ij} = 1$ represents the sample e_i is 100% related to the skill k_j and $Q_{ij} = 0$ otherwise. Formally, the cognitive diagnosis problem for machine learning can be defined as follows.

Definition 2. Given the learner-sample response matrix R and the sample-skill relevancy matrix Q, the cognitive diagnosis task aims to train a diagnoser which can assess the Ability of different learners on different skills (Figure 1(a)).

With the help of a diagnoser, not only the Ability of each learner on specific skills can be quantified but also some of the sample factors (e.g., sample difficulty) are collaboratively quantified. Before introducing our diagnoser Camilla, we should note that the performance of a diagnoser is difficult to evaluate as we cannot obtain the ground-truth proficiency of learners. Therefore, we will investigate the reliability of diagnosers indirectly through their performance in predicting the response of learners on unknown samples. Let us take the classification task as an example for illustration. There are totally two cases for a successful prediction of the diagnoser: (1) the diagnoser predicts that one learner

[*] Corresponding author (email: cheneh@ustc.edu.cn)



Response Matrix $\boldsymbol{R}$

	e_1	e_2	e_3	e_4
s_1	0	0	1	1
s_2	1	0	0	1
…				

Sample-Skill Matrix $\boldsymbol{Q}$

	k_1	k_2	k_3	k_4	k_5
e_1	0	0	1	0	1
e_2	0	1	0	1	0
e_3	0	1	0	0	0
e_4	1	0	0	0	0
…					

e_1	e_2	e_3	e_4	…
0.4	0.9	0.5	0.2	

r_{ij}	e_1	e_2	e_3	e_4
s_1	0	0	1	1
s_2	1	0	0	1
…				

k_1	Explicit *skill* “***ship***”
k_2	Explicit *skill* “***bird***”
k'_*	Latent *skills*
A	Learner *Ability* factor
Q	Sample *skill mask* factor
D	Sample *difficulty* factor
b	Sample *discrimination*

·	Dot Product
–	Element Minus
⊗	Element Product

Figure 1 (Color online) (a) Formal procedure of the cognitive diagnosis task; (b) architecture of the Camilla-Base diagnoser.

can answer the class label of this sample correctly, and the learner does answer it correctly; (2) the diagnoser predicts that one learner cannot answer the class label correctly, and the learner does give a wrong answer.

Method. Based on Assumption 1 inspired by the psychometric theories from human measurement [3], we introduce the Camilla-Base, a basic version of Camilla, as in Figure 1(b).

Input. The input of Camilla-Base includes the response logs/matrix R from well-trained learners to data samples, a pre-defined sample-skill matrix Q, and the representations of learners and samples. Let one-hot vectors $s_i \in \{0,1\}^{1\times N}$ and $e_j \in \{0,1\}^{1\times M}$ denote learner s_i and sample e_j.

Mapping layer. The mapping layer consists of factors that depict the latent characteristics of learners and samples.

(1) Learner Ability. The meaning of multi-dimensional Ability factor is given in Definition 1. Each dimension of Ability corresponds to the proficiency of learners on a specific skill. We characterize the Ability factor A_i of learner s_i as $A_i = s_i \times W_A$, where $A_i \in \mathbb{R}^{1\times K}$, $W_A \in \mathbb{R}^{N\times K}$ is a trainable transformation matrix, K is the number of explicit skills pre-defined in the sample-skill matrix Q.

(2) Sample factors. The sample factors characterize the latent traits of samples consisting of three components: skill mask factor, difficulty factor, and discrimination factor.

- Sample skill mask factor. As the proficiency of learners is characterized by a multi-dimensional Ability, we consider that different samples correspond to different dimensions of the learner Ability factor, which can be evidenced by the sample-skill matrix Q. Specifically, the skill mask factor depicts the skills that are most needed for the correct response of the sample by masking the other unrelated skills. The skill mask factor Q_j of sample e_j is fixed as $Q_j = e_j \times Q$, where $Q_j \in \{0,1\}^{1\times K}$, and e_j denotes the one-hot vector of sample e_j.
- Sample difficulty factor. Intuitively, if one learner can correctly respond to a data sample while other learners cannot, then this sample with high difficulty should contribute more to this learner's corresponding Ability, and vice versa. Therefore, capturing the relationship between the Ability factor of learners and the difficulty factor of samples can help to more precisely predict how sure the learner can respond to the sample correctly. Let $D_j = e_j \times W_D$ denote the difficulty factor of sample e_j, where $D_j \in \mathbb{R}$, and W_D is a trainable transformation matrix.
- Sample discrimination factor indicates the capability of sample e_j to differentiate the mastery degree of learners. We represent the sample discrimination factor as $b_j = e_j \times W_b$, where $b_j \in \mathbb{R}^{1\times K}$, and $W_b \in \mathbb{R}^{M\times K}$ is a trainable transformation matrix.

Note that the learner Ability and sample factors are learnable parameters without any label, which are optimized through the error between the ground-truth response logs and the output of the interaction layer.

Interaction layer (output $r_{ij} \in R$). After giving the input and mapping representation, the way to define a function for modeling the complex interactions among learners, samples, and skills is one of the most important components for a diagnoser. We adopt an interaction layer which outputs the probability $r_{ij} \in R$ that the learner s_i responds to the sample e_j correctly via the comparison between the multi-dimensional learner Ability A_i and the difficulty D_j of the sample in covered skills. In this layer, we define the diagnose function as

$$p\left(r_{ij}=1|A_i,Q_j,b_j,D_j\right)=1/(1+\mathrm{e}^{-Q_j\otimes b_j\cdot\left(A_i-D_j\right)}),\quad (1)$$

where $\otimes$ is element-product and $\cdot$ is dot-product.

The description of Camilla is shown in Appendix A.

Experiments can refer to Appendix B. The discussion with related work can refer to Appendix C.

Conclusion. We have proposed the cognitive diagnostic framework Camilla and a multi-dimensional metric Ability for providing both interpretable and reliable assessment of machine learning algorithms. To the best of our knowledge, this is the first comprehensive attempt for measuring the multifaceted strength of each machine learning algorithm by exploring the connections between the research on psychometric theories and machine learning evaluation.

Acknowledgements This work was partially supported by National Natural Science Foundation of China (Grant Nos. 62337001, U20A20229), Fundamental Research Funds for the Central Universities, Guangzhou-HKUST(GZ) Joint Funding Program (Grant No. 2023A03J0008), and China Postdoctoral Science Foundation (Grant No. 2024T170497).

Supporting information Appendixes A–C. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Zhang Z, Wu L, Liu Q, et al. Understanding and improving fairness in cognitive diagnosis. Sci China Inf Sci, 2024, 67: 152106

2 Graves A, Bellemare M G, Menick J, et al. Automated curriculum learning for neural networks. In: Proceedings of the 34th International Conference on Machine Learning, 2017. 1311–1320

3 Nichols P D, Chipman S F, Brennan R L. Cognitively Diagnostic Assessment. New York: Routledge, 2012