Misclassification Detection based on Conditional VAE for Rule Evolution in Learning Classifier System

Hiroki Shiraishi
The University of Electro-Communications
Chofu, Tokyo, Japan
hirowhite@cas.lab.uec.ac.jp

Masakazu Tadokoro
The University of Electro-Communications
Chofu, Tokyo, Japan
tadokoro@cas.lab.uec.ac.jp

Yohei Hayamizu
The University of Electro-Communications
Chofu, Tokyo, Japan
hayamizu@cas.lab.uec.ac.jp

Yukiko Fukumoto
The University of Electro-Communications
Chofu, Tokyo, Japan
fukumoto@cas.lab.uec.ac.jp

Hiroyuki Sato
The University of Electro-Communications
Chofu, Tokyo, Japan
h.sato@uec.ac.jp

Keiki Takadama
The University of Electro-Communications
Chofu, Tokyo, Japan
keiki@inf.uec.ac.jp

ABSTRACT

This paper focuses on the problem of Learning Classifier System (LCS) that is hard to guarantee to generate the "correct" output (*i.e.*, the action in LCS) as the dimension size of data increases (which results in producing the "incorrect" output) and proposes the method that can detect the incorrect output of LCS. For this issue, this paper proposes the Misclassification Detection based on Conditional Variational Auto-Encoder (MD/C) which detects and rejects the incorrect output of LCS through a comparison between the original data and the restored data by CVAE (Conditional Variational Auto-Encoder) based on the output of LCS (as the condition to CVAE). The results of a ten-class classification problem using handwritten digits showed that MD/C properly rejects the incorrect output of LCS and achieves 99.0% correct rate.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Rule learning**.

KEYWORDS

Learning Classifier System, Auto-Encoder, data mining, misclassification detection

ACM Reference Format:
Hiroki Shiraishi, Masakazu Tadokoro, Yohei Hayamizu, Yukiko Fukumoto, Hiroyuki Sato, and Keiki Takadama. 2021. Misclassification Detection based on Conditional VAE for Rule Evolution in Learning Classifier System. In *2021 Genetic and Evolutionary Computation Conference Companion (GECCO '21 Companion), July 10–14, 2021, Lille, France.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3449726.3459508

GECCO '21 Companion, July 10–14, 2021, Lille, France

ACM ISBN 978-1-4503-8351-6/21/07.
https://doi.org/10.1145/3449726.3459508

1 INTRODUCTION

Learning Classifier System (LCS) [1] is an evolutionary rule-based machine learning system that can find regularities embedded in data in the form of IF-THEN rules that are easily understood by humans and are expected to be applied to tasks that require accountability. However, LCS has the following problems: (1) the learning performance for high-dimensional inputs is extremely low, and (2) the evolutionary process of the rules includes random, which cannot guarantee to obtain "completely" correct rules (*i.e.*, correct outputs). From these weaknesses, it is difficult to apply LCS to real-world problems, such as automated driving, where life-threatening decision errors are not allowed. To overcome this problem, this paper focuses on CVAE (Conditional Variational Auto Encoder) [2], which can generate data corresponding to arbitrary class labels, and proposes MD/C (Misclassification Detection based on CVAE), which detects "incorrect" outputs of LCSs based on the similarity between the original input data and the data generated by the output class of LCS. This paper investigates the effectiveness of MD/C by integrating it into XCSR (eXtended Classifier System for Real Values) [4], and calls it CVAEXCSR (XCSR based on dimensionality reduction by auto-encoder), which can cope with high-dimensional inputs.

2 PROPOSED SYSTEMS

2.1 CVAEXCSR

As shown in Figure 1, CVAEXCSR is composed of XCSR with the CVAE encoder to generate the IF-THEN rules for high-dimensional inputs. In CVAEXCSR, the CVAE encoder starts to reduce the dimension of the input data, and then XCSR generates the classifiers for the latent variables compressed from the high-dimensional inputs.

2.2 MD/C

MD/C detects whether the output of the CVAEXCSR is correct or incorrect for the input data. If MD/C judges that the output is incorrect, MD/C rejects the output. According to Eq. (1), MD/C calculates the *abnormality* $\mathcal{A}(act)$, which is an indicator of the reliability of the output *act* of CVAEXCSR. If $\mathcal{A}(act)$ exceeds the

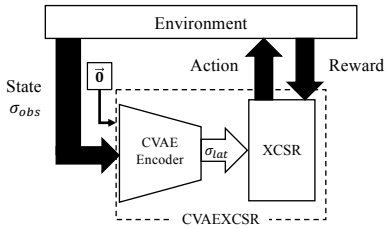

Figure 1: The Architecture of CVAEXCSR

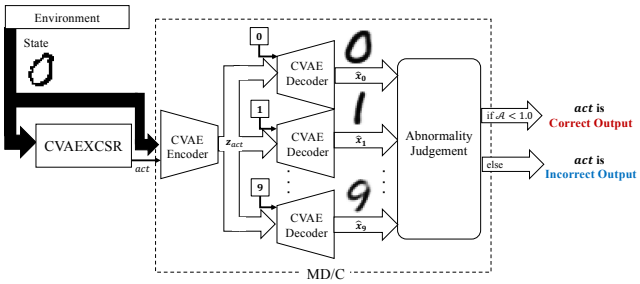

Figure 2: The Architecture of MD/C

threshold value of 1.0, it is judged as an "abnormal" and the output *act* of CVAEXCSR for it is rejected.

In Eq. (1), MSE$(\cdot)$ is the mean squared error, $\boldsymbol{x}$ is the original input data, $\widehat{\boldsymbol{x}}_{act}$ is the decoded input data according to the output *act* of XCSR assumed as the correct answer label, and $\widehat{\boldsymbol{x}}_i$ is the decoded input data according to the correct answer label of the i-th class. $\boldsymbol{\mathcal{E}} = \{\text{MSE}(\boldsymbol{x}, \widehat{\boldsymbol{x}}_i) | 0 \leq i \leq label.\text{length} - 1, i \neq act\}$ means the set of MSEs of the original input data $\boldsymbol{x}$ and the decoded input data $\widehat{\boldsymbol{x}}_i$ of the i-th class except for the output *act* of CVAEXCSR, where *label*.length indicates the number of classes to be classified.

$$\mathcal{A}(act) = \frac{\text{MSE}(\boldsymbol{x}, \widehat{\boldsymbol{x}}_{act})}{\min \boldsymbol{\mathcal{E}}} \tag{1}$$

For example, as shown in Figure 2, if the original input data $\boldsymbol{x}$ is "0" and CVAEXCSR misclassifies it as "9" (*i.e.*, $act = 9$), the decoded input data $\widehat{\boldsymbol{x}}_{act}$ will be the image "9". On the other hand, another decoded input data with the most similar shape to the original input data $\boldsymbol{x}$, *i.e.*, $\widehat{\boldsymbol{x}}_{\arg\min_i \boldsymbol{\mathcal{E}}}$, will be the image "0". Based on these relationships, it is expected that $\text{MSE}(\boldsymbol{x}, \widehat{\boldsymbol{x}}_{act}) > \min \boldsymbol{\mathcal{E}}$ in the case of misclassification. To detect this inequality, the threshold of $\mathcal{A}(act)$ is set to 1.0.

3 EXPERIMENT

3.1 Experiment Design

To investigate the effectiveness of MD/C in CVAEXCSR, this paper compares the results of three types of LCS, *i.e.,* XCSR, CVAEXCSR, and CVAEXCSR+MD/C in MNIST database classification problem [3]. In this problem, we compare the correct rate of ten handwritten numeric images from "0" to "9". The reward is set 1000 for the correct answer and 0 for the wrong answer. Note that MD/C is activated only during the exploitation (evaluation) phase of CVAEXCSR, not during the exploration (learning) phase.

Regarding CVAE, in particular, CVAE in CVAEXCSR and CVAE in MD/C are trained with 100 mini-batches and 100 training epochs. The number of nodes in CVAE in CVAEXCSR is set to {784-512-10-512-784}, meaning that the encoder compresses 784 to 10 dimensional inputs. The number of nodes in CVAE in MD/C is set to {784-784-784-784-784}, meaning that MD/C is executed without the dimensional reduction of the input data.

In the experiment, 300,000 iterations are conducted in one trial, and the 30 trials are conducted with the different random seeds.

3.2 Result

Figure 3 shows the correct rate trend, where the horizontal axis indicates the number of iterations and the vertical axis indicates the correct rate averaged over 1000 iterations. This figure suggests that XCSR has not progressed in learning at all. In contrast, CVAEXCSR gradually achieved a nearly 88% correct rate at the end of the iteration, and CVAEXCSR+MD/C achieved 99% from the first of the iteration by rejecting abnormal outputs. Note that MD/C rejected the output of CVAEXCSR 54,228.6 times out of 300,000 times.

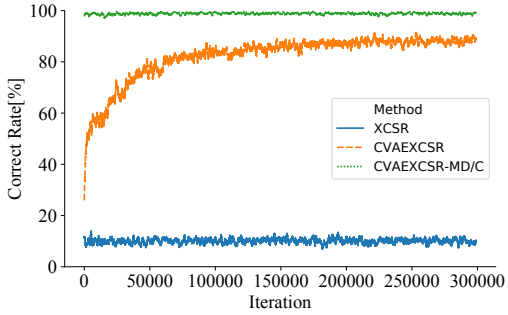

Figure 3: Correct Rate

4 CONCLUSION

This paper proposed MD/C to detect the misclassified output of an LCS and integrated it into CVAEXCSR to cope with high-dimensional inputs. The experiment on the ten-class classification problem showed that (1) MD/C can appropriately detect the incorrect output of CVAEXCSR and (2) CVAEXCSR+MD/C showed a certain level of classification performance for high-dimensional inputs that are difficult to learn by XCSR. Future work includes applying CVAEXCSR+MD/C to life-threatening problems such as medical image classification and the extension of MD/C to reinforcement learning tasks.

REFERENCES

[1] John H Holland. 1986. The possibilities of general-purpose learning algorithms applied to parallel rule-based systems. *Machine learning, an artificial intelligence approach* 2 (1986), 593–623.

[2] Durk P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. *Advances in neural information processing systems* 27 (2014), 3581–3589.

[3] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[4] Stewart W Wilson. 1999. Get real! XCS with continuous-valued inputs. In *International Workshop on Learning Classifier Systems*. Springer, 209–219.