Predictive Model for Epistasis-based Basis Evaluation on Pseudo-Boolean Function Using Deep Neural Networks

Yong-Hoon Kim
Dept. Comp. Sci., Kwangwoon Univ.
Seoul, Republic of Korea
hia5314@gmail.com

Junghwan Lee
Dept. Comp. Sci., Kwangwoon Univ.
Seoul, Republic of Korea
jazz4rabbit@gmail.com

Yong-Hyuk Kim
Dept. Comp. Sci., Kwangwoon Univ.
Seoul, Republic of Korea
yhdfly@kw.ac.kr

ABSTRACT

Complexity of a problem can be substantially reduced through basis change, however, it is not easy to find an appropriate basis in representation because of difficulty of basis evaluation. To address this issue, a method has been proposed to evaluate a basis based on the epistasis that shows the problem difficulty. However, the basis evaluation is very time-consuming. In this study, a method is proposed to evaluate a basis quickly by developing a model that estimates the epistasis from the basis by using deep neural networks. As experimental results of variant-onemax and NK-landscape problems, the epistasis has been estimated successfully by using the proposed method.

CCS CONCEPTS

• **Computing methodologies → Supervised learning by regression**; • **Mathematics of computing →** *Approximation*;

KEYWORDS

basis, deep neural networks, epistasis, pseudo-Boolean function

ACM Reference Format:
Yong-Hoon Kim, Junghwan Lee, and Yong-Hyuk Kim. 2019. Predictive Model for Epistasis-based Basis Evaluation on Pseudo-Boolean Function Using Deep Neural Networks. In *Genetic and Evolutionary Computation Conference Companion (GECCO '19 Companion), July 13–17, 2019, Prague, Czech Republic.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3319619.3326784

1 INTRODUCTION

A pseudo-Boolean function is a function $f : \{0,1\}^n \rightarrow \mathbb{R}$, and numerous problems of diverse application areas can be expressed naturally by using the pseudo-Boolean functions. Lee and Kim [3] studied the effect and importance of basis in a pseudo-Boolean optimization problem, and they demonstrated that the problem space was changed smoothly by using an appropriate basis found with a meta-GA. Lee and Kim [2] proposed a method that quickly finds the basis by performing evaluation, based on the epistasis [1] [1] from the perspective of simplifying the problem space. While we have all the solutions, we can exactly compute an epistasis. This inefficiency could be solved by estimating the epistasis of sampling data. However, the time complexity for finding the epistasis of basis is $O(l^2 s)$, where l is the chromosome size and s is the number of sampling S, and it takes a substantial amount of time for the epistasis-based basis evaluation. Therefore, in this study, we propose a method that estimates the epistasis by using a deep neural network (DNN) model to reduce the time required to calculate the epistasis.

Table 1: The fixed hyperparameters in the used DNN

Hyperparameter	Value
Learning rate	0.0001
Batch size	32
Epoch	1,000
Loss function	RMSE
Optimizer	"Adam"
Number of neurons per layer	$n^2/2$

* n : the dimension of each problem

2 EXPERIMENTS

2.1 Test Environments

The experiments were conducted in an environment that consists of Intel i7-6850K 3.60 GHz CPU and four GTX 1080Ti GDDR5X 11 GB GPUs. Furthermore, for the training data of DNNs, we used all the populations obtained from experiments of a genetic algorithm (GA), which applied the basis change from an instance of Variant-onemax and NK-landscape problems of Lee and Kim [2]. In addition, the data were modified by performing a normalization by dividing each value by a largest value among the epistases. The result of performing the deduplication of basis showed that the numbers of data in each problem were 1141 in minimum and 9816 in maximum respectively. Because we trained DNN models with small data, 10-fold cross-validation was performed in the experiments to prevent the problem of overfitting. Because the DNN models had numerous hyperparameters, the experiments were conducted by adjusting the number of hidden layers, initializer, and dropout rate, to find appropriate hyperparameters. The other hyperparameters were set up as shown in Table 1.

2.2 Results

The ratio in Figures 1 and 2 is obtained by using $100 \times \frac{|E_T - E_P|}{E_T}(\%)$, where E_T is the actual epistasis, and E_P is the estimated epistasis. Decrease in the ratio means that the estimated epistasis is getting closer to the actual epistasis. Except for three cases, it was less than 1%, showing that the epistasis was properly estimated. The ratio was at minimum when the number of hidden layers was three, the initializer was xavier, and the dropout rate was 0.5.

[1] In GA, epistasis means mutual relationship between genes, and when the value is large, it means that the problem is difficult.

GECCO '19 Companion, July 13–17, 2019, Prague, Czech Republic

ACM ISBN 978-1-4503-6748-6/19/07…$15.00
https://doi.org/10.1145/3319619.3326784

Table 2: Results of training the DNN models with the data for each problem (the number of hidden layers = 3, initializer = xavier, and dropout rate = 0.5)

Problem		Train set		Test set	Actual epistasis		Estimated Epistasis		p-value	Time (s)
		Loss at epoch 1	Loss at epoch 1000	Loss	Ave	SD	Ave	SD		
Variant-onemax	$N = 20$	2.38e-1	3.69e-3	2.81e-3	7.70e-1	1.01e-1	7.68e-1	9.00e-2	3.81e-1	718.4
	$N = 30$	2.68e-1	2.06e-3	1.75e-3	6.54e-1	8.52e-2	6.53e-1	7.29e-2	4.86e-1	926.4
	$N = 50$	1.63e-1	1.23e-3	5.38e-4	7.08e-1	6.04e-2	7.05e-1	5.63e-2	3.04e-4	6773.2
NK-landscape	$N = 20, K = 3$	2.49e-1	4.62e-3	4.07e-3	5.77e-1	1.33e-1	5.74e-1	1.22e-1	4.64e-1	540.5
	$N = 20, K = 5$	3.83e-1	9.04e-3	1.59e-3	8.41e-1	4.97e-2	8.42e-1	4.76e-2	5.54e-1	450.5
	$N = 20, K = 10$	3.23e-1	1.53e-3	1.07e-3	9.18e-1	3.78e-2	9.19e-1	3.34e-2	1.55e-1	583.1
	$N = 30, K = 3$	3.22e-1	2.79e-3	1.13e-3	7.23e-1	7.30e-2	7.25e-1	6.91e-2	3.99e-1	873.7
	$N = 30, K = 5$	2.39e-1	1.20e-3	6.02e-4	8.58e-1	4.24e-2	8.58e-1	3.91e-2	3.92e-1	1151.2
	$N = 30, K = 10$	2.41e-1	8.77e-4	5.16e-4	9.26e-1	2.81e-2	9.26e-1	2.47e-2	9.79e-1	1468.9
	$N = 30, K = 20$	2.64e-1	7.78e-4	4.23e-4	9.58e-1	1.75e-2	9.57e-1	1.57e-2	1.62e-1	1644.0
	$N = 50, K = 3$	1.78e-1	1.14e-3	4.23e-4	7.31e-1	5.72e-2	7.29e-1	5.33e-2	6.19e-3	7275.5

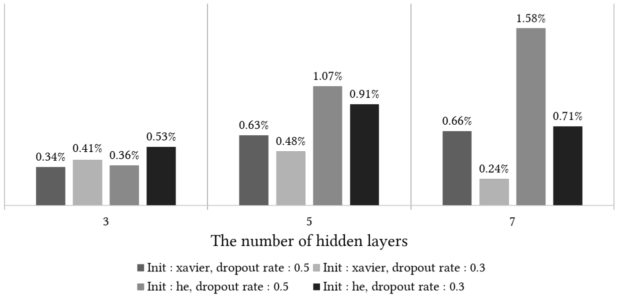

Figure 1: Closeness of actual epistasis and predicted epistasis in the variant-onemax problem

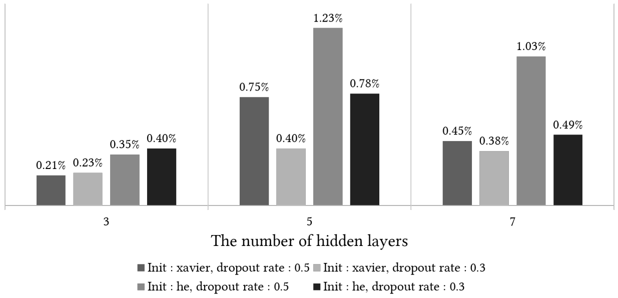

Figure 2: Closeness of actual epistasis and predicted epistasis in the NK-landscape problem

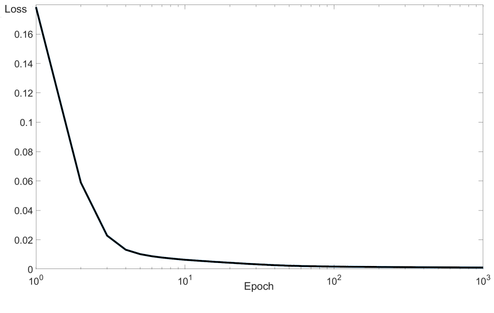

Figure 3: Changes of loss when the DNN was trained in the NK-landscape problem ($N = 50, K = 3$, the number of hidden layers = 3, initializer = xavier, and dropout rate = 0.5)

To check whether or not the training was implemented properly, the loss changes in the case of an instance of the NK-landscape problem ($N = 50, K = 3$, the number of hidden layers = 3, initializer = xavier, and dropout rate = 0.5) were shown in Figure 3. It was confirmed that as the epoch increased, the loss decreased gradually. The results of training the DNN models are summarized in Table 2. When the training was performed for the samples of each problem by repeating the epoch 1,000 times each, it was confirmed that the loss value became sufficiently small. The comparison of actual epistasis and the estimated epistasis showed that the DNN models applied in these experiments performed the estimation properly. Based on the p-value obtained from the t-test of both sides, at 90% confidence level, it was confirmed that in all the cases except two cases there is no difference statistically.

3 FUTURE WORK

We proposed predictive models using DNNs to estimate the epistasis. In the estimation results based on the training in Table 2, the proposed DNN models estimated the epistasis successfully. In this study, the hyperparameters of DNN models were limited. However, if the experiments are performed with diverse hyperparameters, it is expected that better results will be obtained compared to those of the experiments of this study.

Lee and Kim [2] used a GA that searched the basis based on the epistasis. However, as mentioned in the introduction, much time is required for the basis evaluation. To resolve this problem, a study will be conducted in future to search an optimal basis with a surrogate model-based GA [4] by estimating the epistasis using DNN models.

ACKNOWLEDGMENTS

This research was supported by a grant [KCG-01-2017-05] through the Disaster and Safety Management Institute funded by Korea Coast Guard of Korean government.

REFERENCES

[1] Yuval Davidor. 1990. Epistasis variance: suitability of a representation to genetic algorithms. *Complex Systems* 4, 4 (1990), 369–383.

[2] Junghwan Lee and Yong-Hyuk Kim. 2019. Epistasis-based Basis Estimation Method for Simplifying the Problem Space of an Evolutionary Search in Binary Representation. *CoRR* abs/1904.09103 (2019).

[3] Junghwan Lee and Yong-Hyuk Kim. 2018. Importance of finding a good basis in binary representation. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion.* ACM, 49–50.

[4] Sébastien Verel, Bilel Derbel, Arnaud Liefooghe, Hernan Aguirre, and Kiyoshi Tanaka. 2018. A surrogate model based on Walsh decomposition for pseudo-Boolean functions. In *International Conference on Parallel Problem Solving from Nature.* Springer, 181–193.