Quality Measures for Approximate Models in Evolutionary Computation

Yaochu Jin
Honda Research Institute Europe
63073 Offenbach/Main, Germany
yaochu.jin@honda-ri.de

Michael Hüsken
Institut für Neuroinformatik
Ruhr-Universität Bochum
44780 Bochum, Germany

Bernhard Sendhoff
Honda Research Institute Europe
63073 Offenbach/Main, Germany
bernhard.sendhoff@honda-ri.de

Abstract

This paper introduces different metrics for measuring the quality of meta-models in evolutionary computation. The relations between the different metrics are empirically analyzed and neural network models are trained using these different criteria.

1 Introduction

In the recent years, the interest in using approximate models (also known as meta models, or surrogates) for fitness evaluations in evolutionary computation has been increasing [1, 3]. Usually, the quality of approximate models is evaluated with the quadratic approximation error. However, the approximation task in the context of a meta-model for fitness approximation is not completely the same as in the context of optimal prediction. For a meta-model a qualitative approximation is often sufficient, whereas prediction needs a minimal quantitative difference. The examples in Fig. 1 (a) and (b) illustrate what we mean by "qualitative". The approximation accuracy of the neural networks shown might be quite unsatisfying, nevertheless, these approximate models are still able to lead an optimization algorithm to the correct minimum of the fitness function. In this sense, the quality of the meta-model for fitness approximation is sufficient, although the approximation error is high. Thus, it is worth considering other quality measures for evaluating neural networks that are used as surrogates in evolutionary computation.

In this paper, we will present some different metrics for measuring the quality of meta-models for fitness approximation in evolutionary computation. The relationship between these metrics and the quadratic approximation error is empirically studied. Neural networks are trained based on these different metrics using an evolution strategy. Discussions of the preliminary results are included at the end of the paper.

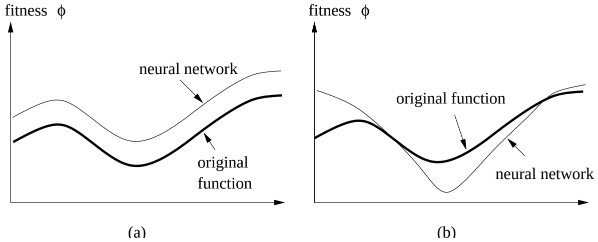


Figure 1: Although the approximation errors of the neural network models are quite large, the optimization by means of the approximate models leads to the desired minimum of the fitness.

2 Quality Measures

One of the main issues in the design and use of meta-models for evolutionary computation is their quality. However, quality of approximate models for evolutionary computation is not necessarily a close quantitative approximation of the original fitness function. Rather, the meta-model should enable the evolutionary algorithm to select the best individuals in terms of the original fitness function.

2.1 Definition of Quality Measures

The most popular measure for model quality is the mean squared difference between the individual's original fitness function $\phi^{(\text{orig.})}$ and the output of the approximate model $\phi^{(\text{model})}$

$$E^{(\text{mse})} = \frac{1}{n}\sum_{j=1}^{n}\left(\phi_j^{(\text{model})} - \phi_j^{(\text{orig.})}\right)^2 . \qquad (1)$$

Here, the mean squared difference is averaged over n different individuals taken into account for the estimation of the quality measure, e.g., the $n = \lambda$ offspring individuals in one generation.

Generally speaking, a model with good approximation quality ensures the correct evaluation and consequently the correct selection of the individuals. However, from the evolutionary computation point of view, only the correct selection is of importance. In the following, we define a number of measures that focus primarily on the correct model-based selection and not on the approximation accuracy. The exact definitions of the first two measures depend on the selection method. We give expressions only for the case of the (μ, λ)-selection with $\lambda \geq 2\mu$, nevertheless, it is in principle possible to extend the ideas and expressions to other selection schemes.

The first measure we suggest is based on the number of individuals that have been selected correctly using the approximate model:

$$\rho^{(\mathrm{sel.})} = \frac{\xi - \langle \xi \rangle}{\mu - \langle \xi \rangle} \ , \tag{2}$$

where ξ $(0 \leq \xi \leq \mu)$ is the number of correctly selected individuals, i.e., the number of individuals that would have also been selected if the original fitness function had been used for fitness evaluation. The expectation

$$\begin{aligned} \langle \xi \rangle &= \sum_{m=0}^{\mu} m \, \frac{\binom{\mu}{m}\binom{\lambda-\mu}{\mu-m}}{\binom{\lambda}{\mu}} \\ &= \frac{\mu^2}{\lambda} \ . \end{aligned} \tag{3}$$

of ξ in case of random selections is used as a normalization in (2). It can be seen that if all μ parent individuals are selected correctly, the measure reaches its maximum of $\rho^{(\mathrm{sel.})} = 1$, and that negative values indicate that the selection based on the approximate model is worse than a random selection.

The measure $\rho^{(\mathrm{sel.})}$ only evaluates the absolute number of correctly selected individuals. However, in case of $\rho^{(\mathrm{sel.})} < 1$ the measure does not indicate, whether the $(\mu + 1)$-th best or the worst offspring individual has been selected, which may have significant influence on the evolution process. Therefore, the measure $\rho^{(\mathrm{sel.})}$ is extended to include the rank of the selected individuals, calculated based on the original fitness function. The definition of the extended measure $\rho^{(\sim\mathrm{sel.})}$ is as follows. The approximate model gets a grade of $\lambda - m$, if the m-th best individual based on the original fitness function is selected. Thus, the quality of the approximate model can be indicated by summing up the grades of the selected individuals, which is denoted by π. It is obvious that π reaches its maximum, if all μ individuals are selected correctly:

$$\begin{aligned} \pi^{(\mathrm{max.})} &= \sum_{m=1}^{\mu} (\lambda - m) \\ &= \mu \left(\lambda - \frac{\mu + 1}{2} \right) \ . \end{aligned} \tag{4}$$

In analogy to (2) the measure $\rho^{(\sim\mathrm{sel.})}$ is defined by transforming π linearly, using the maximum $\pi^{(\mathrm{max.})}$ as well as the expectation $\langle \pi \rangle = \frac{\mu\lambda}{2}$ for the case of a purely random selection:

$$\rho^{(\sim\mathrm{sel.})} = \frac{\pi - \langle \pi \rangle}{\pi^{(\mathrm{max.})} - \langle \pi \rangle} \ . \tag{5}$$

Besides these two problem-dependent measures for evaluating the quality of the approximate model, two established measures — the rank correlation and the (continuous) correlation — partially fit the requirements formulated above. The rank correlation [5], given by

$$\rho^{(\mathrm{rank})} = 1 - \frac{6 \sum_{l=0}^{\lambda} d_l^2}{\lambda(\lambda^2 - 1)} \ , \tag{6}$$

is a measure for the monotonic relation between the ranks of two variables. In our case, d_l is the difference between the ranks of the l-th offspring individual based on the original fitness function and on the approximate model. The range of $\rho^{(\mathrm{rank})}$ is the interval $[-1; 1]$. The higher the value of $\rho^{(\mathrm{rank})}$, the stronger the monotonic relation with a positive slope between the ranks of the two variables. In contrast to $\rho^{(\sim\mathrm{sel.})}$, the rank correlation does not only take the ranking of the selected individuals, but also the ranks of all individuals into account.

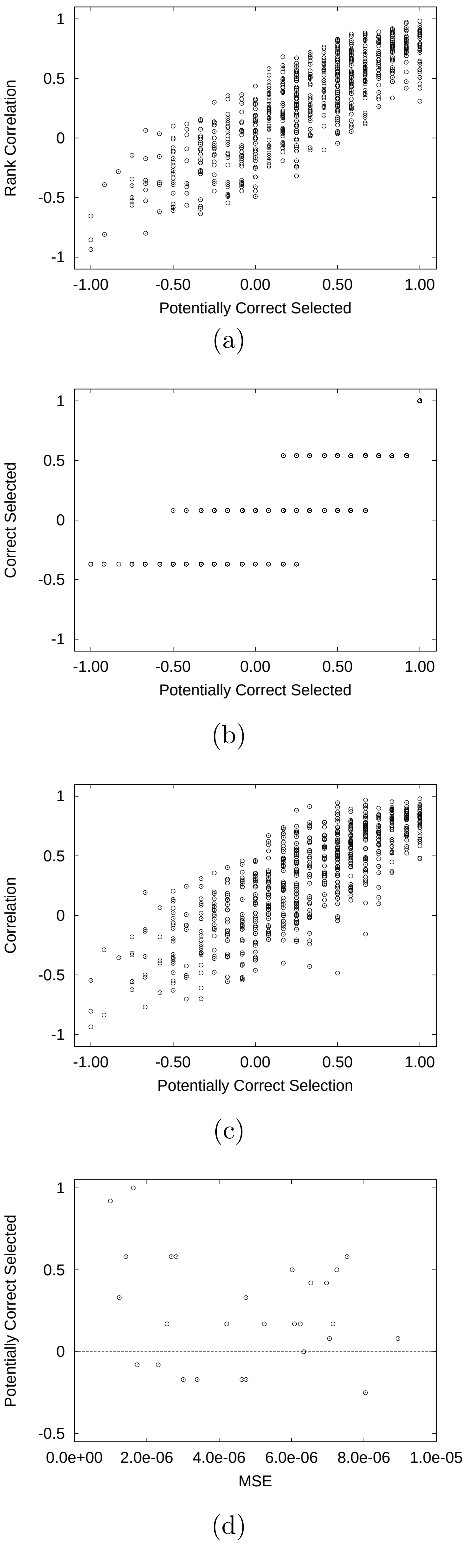
Another possibility to quantify the idea that the approximate model should ensure correct selection, but not necessarily reproduce the correct fitness values, is given by the (continuous) correlation between the approximate model and the original fitness function:

$$\rho^{(\mathrm{corr.})} = \frac{\frac{1}{n} \sum_{j=1}^{n} \left(\phi_j^{(\mathrm{m})} - \bar{\phi}^{(\mathrm{m})} \right) \left(\phi_j^{(\mathrm{o})} - \bar{\phi}^{(\mathrm{o})} \right)}{\sigma^{(\mathrm{m})} \, \sigma^{(\mathrm{o})}} \ . \tag{7}$$

Here, $\bar{\phi}^{(\mathrm{m})}$ and $\bar{\phi}^{(\mathrm{o})}$ are the mean values and $\sigma^{(\mathrm{m})}$ and $\sigma^{(\mathrm{o})}$ the standard deviations of the approximate model output and original fitness function, respectively.

3 Empirical Comparisons

The mean squared error is used as the quality criterion in training the approximate model with data from the



(a)

(b)

(c)

(d)

Figure 2: Scatter plots to illustrate the relation between the different measures. Each circle corresponds to one model, evaluated based on the data of one generation.

evolutionary blade optimization [4, 2]. In Fig. 2 we compare $\rho^{(\sim\text{sel.})}$ with the other four measures. First of all, a mainly linear relation between the measures $\rho^{(\text{corr.})}$, $\rho^{(\text{rank})}$, $\rho^{(\text{sel.})}$ and $\rho^{(\sim\text{sel.})}$ becomes obvious, Fig. 2 (a)-(c). Moreover, the relation between $\rho^{(\sim\text{sel.})}$ and $\rho^{(\text{rank})}$, Fig. 2 (a), as well as $\rho^{(\sim\text{sel.})}$ and $\rho^{(\text{corr.})}$, Fig. 2 (c) looks very similar, which is also emphasized by the high correlation between $\rho^{(\text{corr.})}$ and $\rho^{(\text{rank})}$ (not depicted). Compared with this result, the measure $\rho^{(\text{sel.})}$, Fig. 2 (b), seems to be too coarse-grained to serve as a suitable basis for evaluating the different models.

As the range of $E^{(\text{mse})}$ strongly depends on the shapes of the blades, Fig. 2 (d) is based only on the data from the same generation of the design optimization, evaluated with differently optimized models. For small values of $E^{(\text{mse})}$ the measure $\rho^{(\sim\text{sel.})}$ is decreasing with increasing $E^{(\text{mse})}$, for larger mean squared error $\rho^{(\sim\text{sel.})}$ is mainly fluctuating with zero mean. In particular these strong fluctuations indicate, that $E^{(\text{mse})}$ is only weakly related to the ability to select the correct individuals. Due to the strong linear relation between $\rho^{(\sim\text{sel.})}$, $\rho^{(\text{sel.})}$, and $\rho^{(\text{rank})}$, this result can be carried over to the other measures.

4 Neural Networks Training

In this section, we train neural networks using the root mean squared error (RMSE) criterion, the correct selection criterion (CS) defined in eqn.(2), the potentially correct selection (PCS) criterion, eqn.(5), and the correlation criterion, eqn.(7). The Lamarckian evolutionary method described in [2] is used for minimization of the approximation error or maximization of the CS and PCS ratios and the correlation.

The change of the criteria during the training using the accuracy criterion is shown in Fig. 3. It can be seen that as the approximation error decreases during the training, the correlation between the output of the model and that of the original fitness function increases and is finally close to 1, which indicates that these two variables are strongly correlated. On the other hand, the increase of CS and PCS are rather slow and are fluctuating around zero (random selection) till 100 iterations. This result agrees with the results shown in Fig. 2(d). The final value of the CS and PCS is smaller than 0.5, which means the ratio of correctly selected individuals is quite low when this model is used during evolution.

The training results using the other three criteria are somewhat surprising. It is noticed from the Figures 3(b), (c) and (d) that the approximation error always

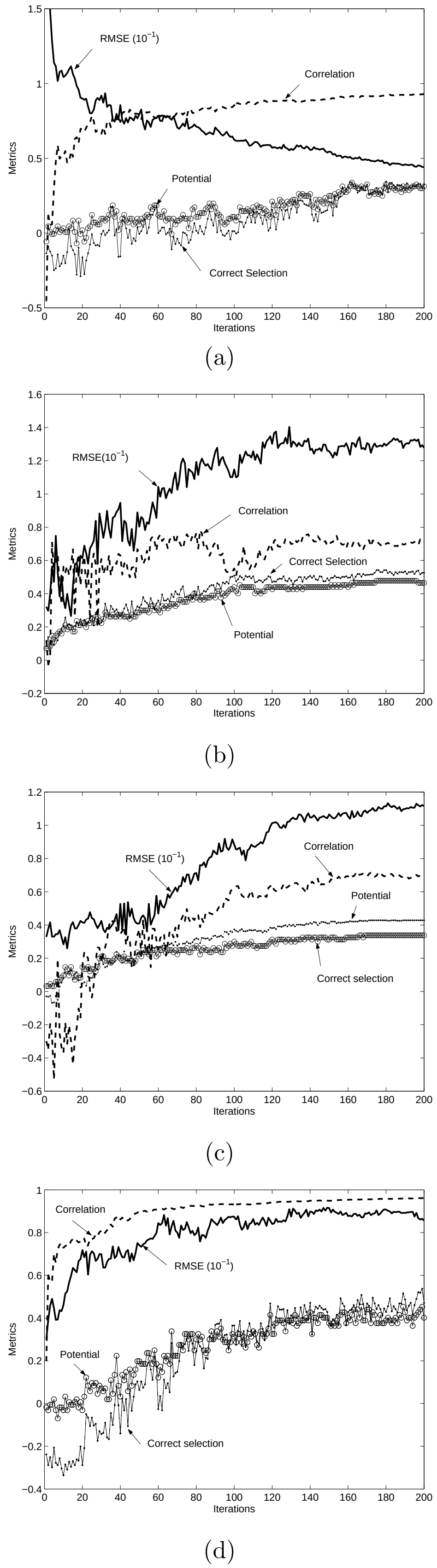


Figure 3: Training of neural networks using the (a) RMSE, (b) CS, (c) PCS and (d) correlation criterion.

increases during the training. Meanwhile, the correlation, the CS and the PCS criteria are quite well correlated. However, all these results do agree with the correlation analysis in Fig. 2, i.e., the CS, PCS and correlation criteria are almost linearly correlated with each other, while the approximation error is not strongly correlated with three criteria.

The question now is, which criterion should be used for training approximate models for evolutionary computation. It is too early to say that the CS, the PCS and the correlation criteria are not suitable for training meta-models because of the large approximation error. However, we should keep in mind that approximate models are usually used together with the original fitness function, as suggested in [4], and large approximation error could cause problems when individual-based evolution control strategies are used.

5 Conclusions

This paper suggested a number of metrics for meta-models for evolutionary computation. The relation between the metrics are empirically studied. Neural networks are trained using different criteria. Further work should be done to investigate the performance of the models trained using different metrics when they are applied to fitness evaluation in evolutionary computation.

References

[1] M.A. El-Beltagy, P.B. Nair, and A.J. Keane. Metamodeling techniques for evolutionary optimization of computationally expensive problems: promises and limitations. In *Proceedings of Genetic and Evolutionary Conference*, pages 196–203, Orlando, 1999. Morgan Kaufmann.

[2] M. Hüsken, Y. Jin, and B. Sendhoff. Structure optimization of neural networks for evolutionary design optimization. *Soft Computing Journal*, 2003. Accepted.

[3] Y. Jin. A comprehensive survey of fitness approximation in evolutionary computation. *Soft Computing Journal*, 2003. Accepted.

[4] Y. Jin, M. Olhofer, and B. Sendhoff. A framework for evolutionary optimization with approximate fitness functions. *IEEE Transactions on Evolutionary Computation*, 6(5):481–494, 2002.

[5] C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15:72–101, 1904.