Pareto Analysis of Evolutionary and Learning Systems

Yaochu Jin, Robin Gruna, Bernhard Sendhoff

2009

Preprint:

This is an accepted article published in Frontiers of Computer Science in China. The final authenticated version is available online at: https://doi.org/[DOI not available]

To appear in "Frontiers in Computer Sciences in China", 2008

Pareto Analysis of Evolutionary and Learning Systems

Yaochu Jin, Robin Gruna, and Bernhard Sendhoff

***Abstract*— This paper attempts to argue that most adaptive systems, such as evolutionary or learning systems, have inherently multiple objectives to deal with. Very often, there is no single solution that can optimize all the objectives. In this case, the concept of Pareto optimality is key to analyzing these systems.**

To support this argument, we first present an example that considers the robustness and evolvability trade-off in a redundant genetic representation for simulated evolution. It is well known that robustness is critical for biological evolution, since without a sufficient degree of mutational robustness, it is impossible for evolution to create new functionalities. On the other hand, the genetic representation should also provide the chance to find new phenotypes, i.e., the ability to innovate. This example shows quantitatively that a trade-off between robustness and innovation does exist in the studied redundant representation.

Interesting results will also be given to show that new insights into learning problems can be gained when the concept of Pareto optimality is applied to machine learning. In the first example, a Pareto-based multi-objective approach is employed to alleviate catastrophic forgetting in neural network learning. We show that learning new information and memorizing learned knowledge are two conflicting objectives, and a major part of both information can be memorized when the multi-objective learning approach is adopted. In the second example, we demonstrate that a Pareto-based approach can address neural network regularization more elegantly. By analyzing the Pareto-optimal solutions, it is possible to identifying interesting solutions on the Pareto front.

I. Introduction

In nature, species are evolving and living in constantly changing environments. It seems that evolution has found different mechanisms of adaptation [1], among which evolution and learning are two main ones. Learning, or more generally, individual level adaptation, usually includes all forms of individual phenotypic changes during an individuals lifetime. In contrast, evolution, also referred to as population level adaptation, represents the evolutionary cycle of selection and genetic variations. In nature, evolution is the main adaptation mechanism for some species such as bacteria, whereas others rely more on the individual level of adaptation.

Both evolution and learning in nature must achieve distinct goals for a better adaptation to changing environments, which is additionally constrained by limited resource of energy. These different goals may be conflicting with each other and cannot be achieved simultaneously. To achieve these goals, nature seems to have evolved systems consisting of a large number of functionally distinct yet systematically integrated subsystems, which is believed to represent a major characteristics of brain complexity [2]. In [3], multi-objectivity has been used to characterize the complexity of an evolved creature.

Yaochu Jin and Bernhard Sendhoff are with the Honda Research Institute Europe, 63073 Offenbach, Germany. Robin Gruna is with the Institut für Technische Informatik, Universität Karlsruhe (TH), 76131 Karlsruhe, Germany. Email: yaochu.jin@honda-ri.de.

This paper demonstrates, with examples, how the Pareto-based approach can be employed to analyze computational evolutionary and learning systems. Section II introduces the mathematical definition of multi-objective optimization and Pareto dominance, and discusses some general properties of Pareto fronts. In Section III, the trade-off between robustness and evolvability of a redundant Boolean representation is investigated by analyzing the Pareto front achieved with a multi-objective evolutionary algorithm. The advantage of Pareto-based approach to addressing catastrophic forgetting and neural network regularization is illustrated in Section IV. Section V briefly summarizes this paper and discusses some potentially interesting topics.

II. Pareto-based Multi-objective Optimization and Analysis

A. Multi-objective Optimization and Pareto Optimality

Consider the following multi-objective minimization problem:

$$\begin{aligned} \text{minimize} \quad & f_m(\mathbf{x}) && m = 1, 2, \cdots, M; && (1)\\ \text{subject to} \quad & g_j(\mathbf{x}) \geq 0, && j = 1, 2, \cdots, J; && (2)\\ & h_k(\mathbf{x}) = 0, && k = 1, 2, \cdots, K; && (3)\\ & x_i^L \leq x_i \leq x_i^U, && i = 1, 2, \cdots, n, && (4) \end{aligned}$$

where $f_m(\mathbf{x})$ are the M different objective functions to be minimized, $\mathbf{x} = (x_1, x_2, \cdots, x_n)^T$ is the n-dimensional decision space, $g_j(\mathbf{x})$ are the J inequality constraints, $h_k(\mathbf{x})$ are the K equality constraints, and x_i^L and x_i^U are the lower and upper bounds of the i-th decision parameter, respectively.

For the multi-objective minimization problem defined above, solution $\mathbf{x}^{(1)}$ is said to dominate solution $\mathbf{x}^{(2)}$, if $\mathbf{x}^{(1)}$ is no worse than $\mathbf{x}^{(2)}$ in all objectives, i.e.,

$$\forall m = 1, 2, \cdots, M, f_m(\mathbf{x}^{(1)}) \leq f_m(\mathbf{x}^{(2)}), \tag{5}$$

and if $\mathbf{x}^{(1)}$ is strictly better than $\mathbf{x}^{(2)}$ in at least one objective:

$$\exists m' \in \{1, 2, \cdots, M\}, \text{ such that } f_{m'}(\mathbf{x}^{(1)}) < f_{m'}(\mathbf{x}^{(2)}). \tag{6}$$

If a solution $\mathbf{x}^*$ is not dominated by any other feasible solutions, solution $\mathbf{x}^*$ is called Pareto-optimal. For most multi-objective optimization problems, there are a finite or infinite number of Pareto-optimal solutions, which are known as the Pareto set in the decision space, and the Pareto front in the objective space.

B. Analysis of Pareto Optimal Solutions

When there is no preference over a particular objective, the Pareto-optimal solutions are non-comparable, i.e., they are equally good. To pick out one solution for the final use, a user needs to provide some preference that helps the user to decide which Pareto-optimal solution best meets user's preference. Due to this reason, it has been argued that it is not necessarily to approximate the whole Pareto front, which might require more computational resources. While this kind of argument is true in some cases, it is our belief that achieving the whole Pareto front is of great value due to the following reasons:

- The shape of the Pareto front, in particular, the convexness or concaveness of the front, as well as the location of knee points and extreme points on the Pareto front, can reveal much additional domain knowledge, refer to Fig. 1. A knee point is a solution on the Pareto front that requires a large sacrifice in the other objectives to improve in one objective. Such domain knowledge may greatly help the user in decision-making for choosing the final solution.
- When aspects other than the performance of the solution in terms of the function value are considered important, for example the robustness of the Pareto-optimal solutions, it is then very helpful to achieve all the Pareto-optimal solutions. By analyzing the robustness of the solutions, the user may modify their preference and choose the best solution for the problem at hand.

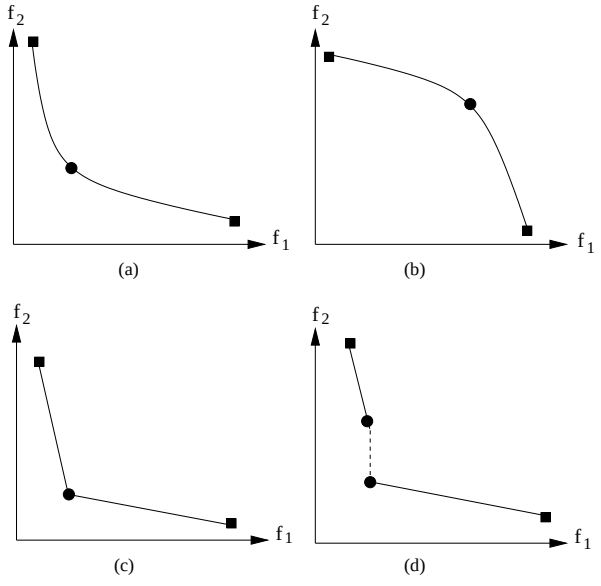


Fig. 1. Convexness and interesting points of Pareto fronts. A filled square means an extreme solution, and a filled circle denotes a knee point. (a) Convex Pareto front, (b) Concave Pareto front, (c) Convex Pareto front with two sections of piecewise linear curves, and (d) Convex Pareto front with two separate linear curves.

While it is of essential importance to analyze the Pareto-optimal solutions for multi-objective optimization problems, it is sometimes very helpful to multi-objectivize problems that originally have only one objective. For example diversity has been used as the second objective to facilitate solving multi-modal problems in reducing the number of local optimums [4], [5], [6], or solving dynamic problems [7].

One important research area in which Pareto-based multi-objective analysis has found to be of particular interest is machine learning [8]. Many typical learning problems. such as neural network regularization, ensemble generation and data clustering, can be addressed by summarizing the different objectives into a scalar objective. These problems can be solved more elegantly when the Pareto-based approach is used. A few examples will be provided in Section IV to elaborate on this point.

III. Pareto Analysis of Redundant Genetic Representations of Evolutionary System

A. Trade-offs in Biological Systems

Natural evolution can never be single-objective. Trade-offs between different targets have accompanied the history of evolution mainly due to the limited amount of energy and time. A few examples in biological systems are:

- A trade-off between energy efficiency and functionality in evolution of brain size and organization of nervous systems. It is believed that the intelligence level of organisms is roughly proportional to their relative brain size, e.g., the ratio of brain weight to body weight. However, a bigger brain is evolved at costs [9]. First, a big brain often means more energy consumption, and therefore more food consumption, which is not always available. Second, animals with a bigger brain usually have a longer life span, which means that they are likely to experience more extreme environmental changes and thus harder survival environments. In the evolution of nervous systems, it has been shown that energy efficiency has also been a main constraint for evolving their functionality. This kind of trade-off has also been shown to lead to the emergence of biologically plausible structure in the artificial evolution of a neural system [10].
- There is a trade-off between reproduction and survival (longevity) [11]. Research has been shown that male fruit flies supplied with virgin females have lower longevity than those kept without access to females [12]. This trade-off has also been shown to be important in evolving an optimal lifetime in an artificial evolutionary system [13].
- Evidence has also been found that supports a trade-off between the number of offspring and their size [14].
- Trade-offs have been found in many research topics of bioinformatics and computational biology. For example, in protein sequence alignment, maximizing the number of matching bases and minimizing the number of gaps may be two conflicting objectives. Conflicting objectives must also be taken into account in constructing phylogenetic tree and gene regulatory networks [15].

B. Pareto Analysis of Robustness-Evolvability Trade-off

It is a challenging and extremely important task to understand how natural evolution has managed to bring about the huge biological diversity and complexity from simple particles and molecules. In addition to environmental changes, it is believed that two important principles, i.e., robustness and evolvability, may have played a central role in shaping biological complexity.

Biological robustness means organisms' ability to relatively maintain their functionality under a certain degree of internal and external perturbations. An important issue directly related to robustness is evolvability, which is organisms' ability to evolve inheritable novel phenotypic functionalities that help the organism survive and reproduce. In the recently years, research on robustness and evolvability has become one of the main research topics in systems biology [16], [17].

Research on robustness and evolvability is still in its infancy [18], [19]. Not only a sophisticated quantitative definition for biological robustness and evolvability is still missing, but the biological origin, that is, how evolution has shaped the various biological mechanisms for robustness and evolvability remains to be understood.

In a broader sense, robustness contributes to evolvability in that without robustness, evolutionary tinkering will most likely lead to the lethal consequences, thus preventing evolution from creating new functionalities. For a clearer understanding of the mechanisms underlying evolvability and robustness, we investigate here evolvability in a narrow sense, that is, systems' ability to generate new phenotypes, termed innovation hereafter. Although it has been recognized qualitatively that there is a trade-off between robustness and innovation, no quantitative results have been reported.

Biological robustness can be achieved with a variety of mechanisms, such as feedback, genotypic redundancy, functional modularity, among others [16]. In the following, we investigate a redundant genotype-phenotype mapping to investigate quantitatively the robustness-evolvability trade-off.

1) A Boolean Model for Genotype-Phenotype Mapping: Consider the following genotype-phenotype mapping from n-dimensional genotype space $\mathcal{G} \in \{0,1\}^n$ to m-dimensional phenotype space $\mathcal{P} \in \{0,1\}^m$:

$$p_i = f_i^{k_i}(c_{i1}(g_1), c_{i2}(g_2), \cdots, c_{in}(g_n)), \tag{7}$$

where p_i, $i = 1, 2, \ldots, m$ is the i-th phenotype trait, g_j, $j = 1, 2, \cdots, n$ is the j-th genotype, and $\mathbf{C} = (c_{ij})_{m\times m}$, where

$c_{ij} = 1 :\Longleftrightarrow$ phenotype trait p_i is affected by gene g_j

$c_{ij} = 0 :\Longleftrightarrow$ phenotype trait p_i is independent of gene g_j.

$f_i^{k_i}(.)$ is a Boolean function with k_i inputs, where

$$k_i = \sum_{j=1}^{n} c_{ij}, \tag{8}$$

is also known as the arity of the Boolean function.

If $k_i > 1$, that is, if phenotype p_i is influenced by more than one genotype, it is called polygeny. In contrast, the number of phenotype traits affected by gene g_j is given by the sum of the elements in the column:

$$l_j = \sum_{i=1}^{m} c_{ij}. \tag{9}$$

If $l_j > 1$, then genotype g_j is said to be pleiotropy. Without loss of generality, it is assumed that $\mathbf{C}$ is chosen such that $l_j \geq 1$ for all $i = 1, \ldots, n$. This means that each gene affects at least one phenotype trait.

The mapping can be defined by the dependencies between genes and phenotype traits, and a set of Boolean functions that determine the values of the phenotype traits. Since the connection between genotype and phenotype determines the number of inputs for a certain phenotype trait and thus for the corresponding Boolean function, the dependencies between genes and phenotype traits are determined at first. Once the connection matrix $\mathbf{C}$ has been fixed, the Boolean functions $f_i^{k_i} : \{0,1\}^{k_i} \to \{0,1\}$, $i = 1 \ldots m$ can then be defined.

An example of genotype-phenotype mapping with eight genes and four phenotype traits is given in Fig. 2, where the arity of the Boolean function for the four phenotypes is three, six, six, and two, respectively. The connection matrix in this example is as follows:

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}. \tag{10}$$

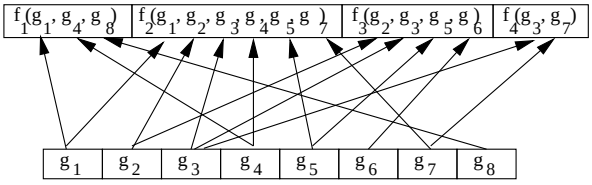


Fig. 2. An example of genotype-phenotype mapping, where the number of genes is eight, and the number of phenotype traits is four.

In the following, we are going to evolve the connection matrix $\mathbf{C}$ as well as the Boolean functions to maximize the robustness and evolvability using the NSGA-II [20], which is one of the most popular evolutionary multi-objective optimization algorithms.

2) Encoding of the Model: The first stage of the multi-objective evolutionary approach is to determine the scheme for representing the genotype-phenotype mapping model. In this work, an encoding with a fixed coding length and relatively compact has been adopted.

The encoding of the Boolean model consists of two parts: the encoding of the connection matrix C and the encoding of the Boolean functions f_i, $i = 1, \ldots, m$. The encoding of C is trivial, since each entry c_{ij} can be written in a binary vector, leading to a fixed encoding length mn, independent of the actual value of C.

Finding an encoding for f_i is more difficult. The connection matrix C determines the polygeny $k_i = \sum_{j=1}^{n} c_{ij}$ of each phenotype trait p_i, i.e., the number of inputs of the Boolean function $f_i^{k_i}$. Since $f_i^{k_i}$ is a Boolean function $f_i^{k_i}\colon \{0,1\}^{k_i} \rightarrow \{0,1\}$, it is completely defined when the corresponding outputs for each 2^{k_i} inputs are determined. It can be seen that such a canonical encoding is dependent on the actual value of the connection matrix $\mathbf{C}$. During evolutionary search the size of such an encoding is changing and a set of *ad hoc* search operators have to be defined to guarantee that newly generated solutions are feasible.

A simpler option is to define an encoding for $f_i^{k_i}()$ independent of its arity k_i. To this end, it is assumed that $k_i = n$ due to the fact that a phenotype trait cannot depend on more than n genes. Unfortunately, this approach would lead to an encoding with a length of $nm + 2^n m$. Even for a medium size of genotype and phenotype spaces, e.g., $m = 8$ and $n = 16$, the length of the chromosome will become intractably large. To address this problem, we impose more restrictions on the Boolean model so that a reasonably small encoding length can be achieved. Since the crucial part of the encoding is the encoding of the Boolean functions $f_i^{k_i}()$ and its dependency on k_i, we are going to define a class of Boolean functions that are independent of the polygeny k_i.

Let $f^{k_e} \in \mathcal{F}^{k_e}$ denote an arbitrary Boolean function with arity k_e. Then the k-ary extension of f^{k_e}, $f^k \uparrow_{f^{k_e}}\colon \{0,1\}^k \rightarrow \{0,1\}$, is recursively defined by

$$f^k\uparrow_{f^{k_e}}(g_1,\ldots,g_k) = \begin{cases} f^{k_e}(g_1,\ldots,g_{k_e}), & \text{if } k = k_e; \\ f^{k_e}(f^{k-1}\uparrow_{f^{k_e}}(g_1,\ldots,g_{k-1}),g_k), & \text{if } k > k_e; \end{cases} \tag{11}$$

for $k \geq k_e$. If $k < k_e$, then,

$$f^k \uparrow_{f^{k_e}} (g_1,\ldots,g_k) = f^{k_e}(\underbrace{1,\ldots,1}_{k_e-k},g_1,\ldots,g_k), \tag{12}$$

where the first $(k_e - k)$ positions are set to 1. f^{k_e} is termed the elementary Boolean function and k_e its elementary arity. Fig. 3 shows an example of 4-ary extension of 2-ary Boolean function f^2.

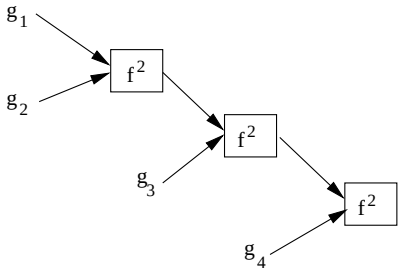
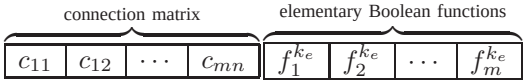


Fig. 3. An example of 4-ary extension of a 2-ary Boolean function: $f^4 \uparrow_{f^2}$.

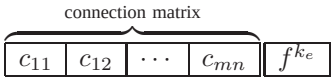
With the definition of the k-ary extensions $f^k \uparrow_{f^{k_e}}$ of Boolean functions, the following three different approaches to encoding the Boolean functions are considered, each of which results in a reasonably compact and fixed encoding length.

- *Multiple Boolean Functions* All $f^k \uparrow_{f^{k_e}}$ have the same encoding size 2^{k_e}, independent of their actual arity k, since only the elementary Boolean function f^{k_e} has to be encoded. Consequently, a restricted version of the Boolean model can be defined as follows: determine a fixed elementary arity k_e for the model and choose every Boolean function $f_i = f^k \uparrow_{f_i^{k_e}}, i = 1,\ldots m$, where k_i is the polygeny of the corresponding phenotype trait p_i and is determined by the connection matrix $\mathbf{C}$. It is important to emphasize that each phenotype function f_i is the k-ary extension of different elementary Boolean functions $f_i^{k_e}$. This model restriction leads to an encoding size of $nm+m2^{k_e}$ bit, which is reasonably small for $k_e = 2, 3, 4$. The structure of the encoding is schematically depicted below:



(13)

- *Single Boolean Function Encoding* An even simpler encoding can be achieved when the Boolean model is restricted to a single elementary Boolean function. In doing so, a fixed elementary arity k_e for the model is determined and the local Boolean functions are k_i-ary extensions of the same elementary Boolean function: $f_i := f^k \uparrow_{f^{k_e}}, i = 1,\ldots m$. This restriction leads to an encoding size of $nm + 2^{k_e}$ bit. The structure of the encoding is schematically depicted as follows:

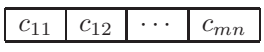


(14)

- *Majority Rule Encoding* In a further simplification of the Boolean model, we dispense with an explicit definition of the Boolean functions f_i. Instead, the phenotype traits are determined by the majority rule. There are several possibilities to break ties. In our work, we set the output of majority rule to its first input in the case of a tie. This keeps the rule balanced and deterministic:

$$f_i(g_{j_1},g_{j_2},\ldots,g_{j_{k_i}}) = \begin{cases} 1, & \text{if } \sum_{l=1}^{k_i} g_{j_l} > k_i/2; \\ 0, & \text{if } \sum_{l=1}^{k_i} g_{j_l} < k_i/2; \\ g_{j_1(i)}, & \text{otherwise.} \end{cases} \tag{15}$$

Hence, only the connection matrix has to be encoded:



(16)

These different encodings restrict the original model in a strong way and so one could say that each presents a different model on its own.

3) Objective Setup I: Maximizing Local Neutral Degree and Local Variability: There is no widely accepted quantitative definition for robustness and evolvability. In this setup, we use local neutral degree for estimating the robustness

and the maximum local innovation for approximating the evolvability of a genotype-phenotype mapping.

Given an genotype-phenotype mapping $\phi\colon \mathcal{G} \rightarrow \mathcal{P}$ and a neighborhood relation N over the set of genotypes $\mathcal{G}$, then the *local neutral degree* $\nu_\phi(g)$ of mapping ϕ is defined by [21]

$$\nu_\phi(g) := \frac{|\{g' \in \mathcal{G} : \phi(g) = \phi(g') \wedge N(g, g')\}|}{|\{g' \in \mathcal{G} : N(g, g')\}|} \tag{17}$$

Similar to the definition of local neutral degree, a definition of the local variability can be defined as follows. Given an genotype-phenotype mapping $\phi\colon \mathcal{G} \rightarrow \mathcal{P}$ and a neighborhood relation N over the set of genotypes $\mathcal{G}$, then the *local variability* $\delta_\phi(g)$ of mapping ϕ at genotype g is defined as

$$\delta_\phi(g) := \frac{|\{g' \in \mathcal{G} : \phi(g) \neq \phi(g') \wedge N(g, g')\}|}{|\{g' \in \mathcal{G} : N(g, g')\}|} \tag{18}$$

Note that the variability δ captures the fraction of unique phenotypes in the non-neutral neighborhood, which is accomplished by the set notation. Accordingly, $\forall g \in \mathcal{G}, \nu_\phi(g) + \delta_\phi(g) \leq 1$ holds by definition.

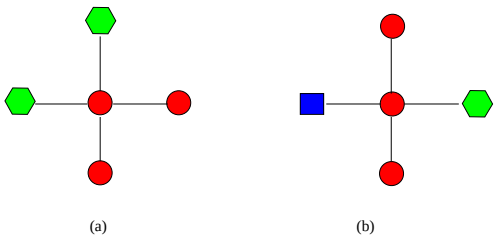


Fig. 4. Two illustrative examples on calculating local neutral degree and local variability. In the figure, filled circles, squares and pentagons denote different phenotypes, while the lines represent local neighborhood in the genotype. (a)$\nu_\phi(g) = 0.5$, $\delta_\phi(g) = 0.25$, and (b)$\nu_\phi(g) = 0.5$, $\delta_\phi(g) = 0.5$.

Two illustrative examples on how to calculate local neutral degree and local variability are provided in Fig. 4.

The definition of local neutral degree has been discussed in a number of studies. However, in most of these studies, the local neutral degree measure is investigated with respect to the single genotypes or neutral sets mapping to the same phenotype. In this work, we want to evaluate different characteristics of neutrality with respect to entire genotype spaces. We use the local neutral degree to calculate the mean neutral degree of a genotype-phenotype mapping, which is accomplished by randomly sampling genotypes and averaging their local neutral degree. A high mean neutral degree indicates that there are many genotypes with a high local neutral degree, and therefore the mean neutral degree reflects how many neutral mutations or how much neutrality is provided by the mapping. In the genotype space where the neutral degree is high, mutations are most likely neutral rather than deleterious. As a consequence, populations on neutral networks tend to drift toward regions with a high neutral degree to evolve robustness. This gives rise to the consideration that the mean neutral degree is related to mutational robustness.

In order to evaluate the robustness of a mapping, the mean of its neutral degree distribution $\bar{\nu}_\phi$ is estimated. The local neutral degree of randomly sampled genotypes is determined and averaged over all samples. Consequently, the mean neutral degree reflects the density of neutral spaces. A mapping with a high mean neutral degree has more regions in the genotype space that are surrounded by neighbors which map to the same phenotype. Therefore, phenotypes associated with dense neutral spaces have a higher robustness against mutations than those associated with less dense neutral spaces.

The mean variability of a genotype-phenotype mapping can be calculated in a similar way. It is the sensitivity of phenotypes to genotypic mutations and therefore can be considered as the opposite of robustness to genetic changes. A high mean variability can be viewed to contribute to evolvability, when a condition of evolvability is defined as to reduce the number of mutations needed to produce phenotypically novel traits [22], [23]. In [24], Fontana states that the capacity to evolve in response to selective pressures depends on phenotypic variability. This suggests that variability can be considered as prerequisite for evolvability.

It is important to note that the neutral degree and variability are related according to $\forall g \in \mathcal{G} : \nu_\phi(g) + \delta_\phi(g) \leq 1$ and thus the mean neutral degree $\bar{\nu}_\phi$ and the mean variability $\bar{\delta}_\phi$ are conflicting properties of a mapping. Therefore, the true Pareto front is given by $\bar{\nu}_\phi + \bar{\delta}_\phi = 1$. In order to study this trade-off relationship and how this is resolved by the Boolean model, we will evaluate different encodings of the model in the proposed multi-objective optimization framework.

In the experiment, different encodings of the Boolean model were optimized with respect to the mean variability and mean neutral degree. The results for a population of 50 individuals and 100 generations are summarized in Fig. 5.

It can be seen that the different encodings lead to quite different results. The random initial population with the multiple Boolean function encoding is unevenly distributed across the objective space. The mappings lie at some distance from the true Pareto front $\bar{\nu}_\phi + \bar{\delta}_\phi = 1$, albeit mappings with $\bar{\nu} > 0.6$ are almost on the true Pareto front. After 100 generations, the mappings lie evenly distributed on the theoretical trade-off surface. This is different to the case in which single Boolean functions encoded. The randomly initialized population lies almost entirely on the theoretical Pareto front, but only on a section $\bar{\nu} < 0.5$. After multi-objective optimization was performed, the mappings are more evenly distributed along this section. Only a few mappings lie on the Pareto front section with $\bar{\nu} < 0.5$. The solutions seem to have an uniform distance from each other. For $0.2 < \bar{\nu} < 0.5$, however, no Pareto optimal solutions have been found. In the case of the majority rule encoding, the randomly initialized mapping form a compact cluster and evolutionary search finds Pareto optimal mappings with $0.5 < \bar{\nu} < 0.8$. In summary, all genotype-phenotype mappings found by different encodings of the Boolean models and evolutionary search lie on the true Pareto front $\bar{\nu}_\phi + \bar{\delta}_\phi = 1$. However, the found Pareto optimal

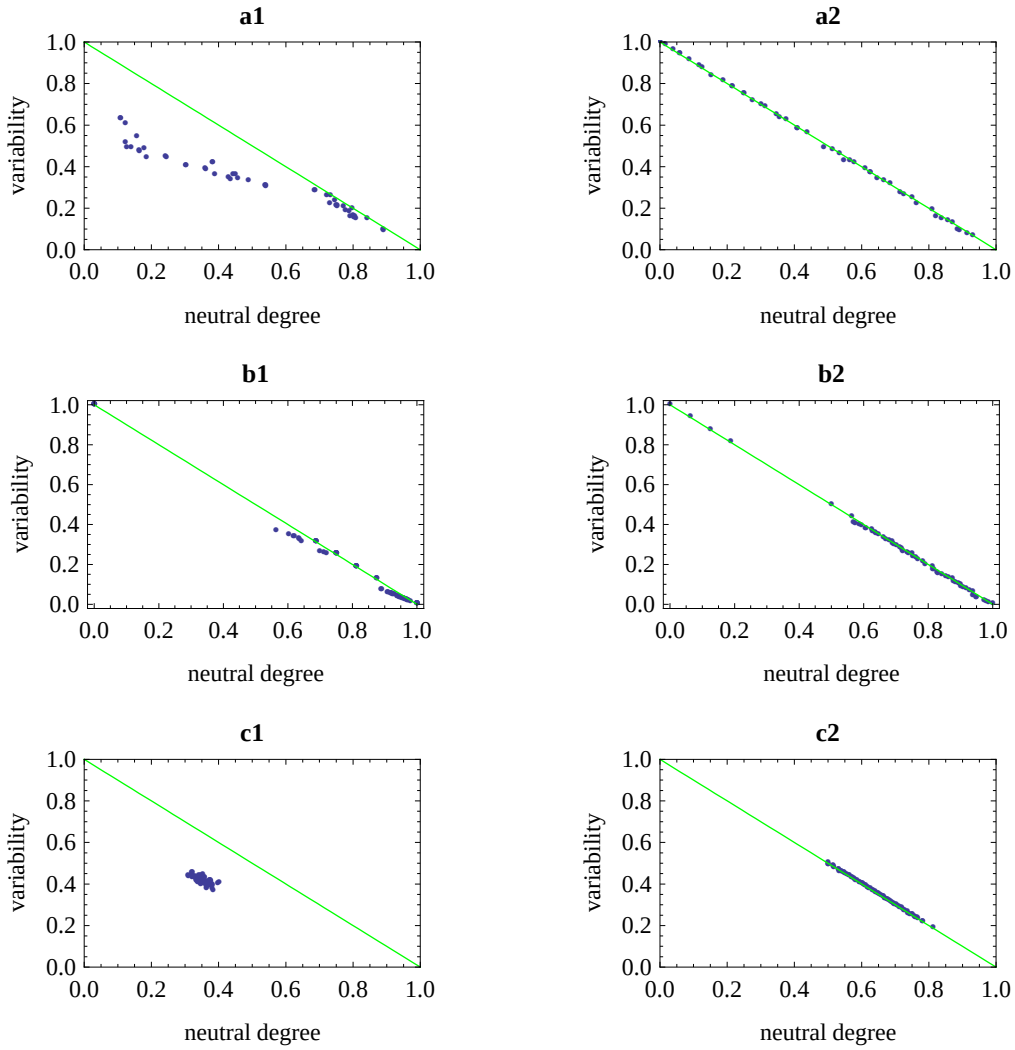


Fig. 5. Multi-objective optimization evaluation of the Boolean model ($n = 16$, $m = 8$) for maximal mean neutral degree $\bar{\nu}_\phi$ and maximal mean variability $\bar{\delta}_\phi$. The model is encoded by different techniques: **a** Multiple Boolean functions encoding with elementary arity 2. **b** Single Boolean function encoding with elementary arity 2. **c** Majority rule encoding. Plots in the left panel show the results of 50 randomly initialized genotype-phenotype mappings with $n = 16$ and $m = 8$. Those in the right panel show the approximated Pareto optimal set of genotype-phenotype mappings after 100 generations. The dotted line indicates the theoretical Pareto front, determined by the definition of neutral degree and variability, i. e. $\bar{\nu}_\phi + \bar{\delta}_\phi = 1$. The different encodings result in different approximations of a Pareto optimal set.

solutions are distributed along different regions. Only with the multiple Boolean function encoding was the evolutionary algorithm able to find the complete Pareto optimal front.

These results indicate that the different encodings of the Boolean functions differ in their capability of realizing various genotype-phenotype mappings. Since the multiple Boolean function encoding restricts the original Boolean model at the least, it is able to approximate the widest range of the theoretical Pareto front. For example, a genotype-phenotype mapping $\phi\colon \{0,1\}^{16} \to \{0,1\}^8$ with variability $\bar{\delta}_\phi = 1$ and neutral degree $\bar{\nu}_\phi = 0$ can be implemented. Despite that this mapping has a redundancy of 2^8, it possesses no neutrality. This means that the genotype space is structured in such a way that genotypes that map to the same phenotype are never neighbored and thus every single point mutations leads to a new phenotype. Therefore, such a mapping can be thought of as having low robustness and high evolvability.

Alternatively, mappings with neutral degree $\bar{\delta}_\phi \geq 0.9$ and variability $\bar{\nu}_\phi \leq 0.1$ can be implemented by the multiple and single Boolean functions encoding. In this case, almost all single point mutations are neutral and lead to the same phenotype. Therefore, such a mapping can be considered as highly robust and less evolvable.

In order to understand how the Boolean model is capable of implementing such different mappings, we analyze the underlying parameters of the model. These are given by the connection matrix $\mathbf{C}$ and the set of Boolean functions f_i, $i = 1, \ldots, m$. To obtain a first impression of the parameters' structures, we visualized them. Fig. 6 shows three example of encoding structures for different encoding schemes. Visually, no significant features or patterns in the encoding structure

can be identified. A profound analysis of the parameters is difficult because of the high degree of freedom in the model. A statistical approach is presented in [25] to interpret the parameter structure of the majority rule encoding.

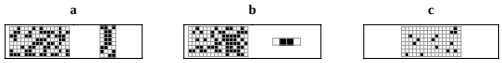


Fig. 6. Exemplary parameter visualization of the Boolean model ($n = 16$, $m = 8$) yielded by different encoding techniques. Each frame corresponds to an implementation of the connectivity matrix $\mathbf{C}$ and the set of phenotype functions (if necessary) of a genotype-phenotype mapping. **a** Multiple Boolean functions encoding with elementary arity 2. The left array depicts the connection matrix, the right array illustrates the phenotype functions f_i, $i = 1, \ldots, m$. Row i depicts the truth table of the elementary Boolean function f_i^2. **b** Single Boolean functions encoding with elementary arity 2. The left array depicts the connection matrix, the right array depicts the truth table of the elementary Boolean function f^2. **c** Majority rule encoding. The array indicates the connection matrix. In all descriptions, black array entries corresponds to 1, whereas white array entries correspond to 0. An entry $c_{ij} = 1$ in the connection matrix indicates, that phenotype trait p_i depends on gene g_j. Therefore, the pleiotropy of gene g_j is given by the sum of column j, whereas the polygeny of phenotype trait p_i is given by the sum of row i.

4) Objective Setup II: Maximizing Neutral Degree and Uniformity Entropy: In order to gain a deeper insight into the relationship between neutral degree and entropy, we carried out a multi-objective optimization with neutral degree and entropy as two objectives. The experiment was performed with different encodings and the results are shown in Fig. 7. The results vary slightly with the different encodings. They differ mainly in the way in which the random initialized population is distributed in the objective space. After 200 generations, the results resemble each other. The mappings encoded with single and multiple Boolean functions form a well-distributed Pareto front. Mappings encoded with the majority rule cover only a section of the Pareto front and do not exceed a neutral degree of 0.8. This has also been observed in the previous experiment and can be attributed to the strong restriction that reduces the original Boolean model to the majority rule.

The results suggest that a genotype-phenotype mapping implemented by the Boolean model can provide high entropy up to a certain neutral degree threshold. It remains to be clarified if this observation is a consequence of the definition of neutral degree and entropy, and so is true for arbitrary genotype-phenotype mappings or this is a property of the Boolean model. Moreover, it would be interesting to investigate how the observed neutral degree threshold depends on other model parameters. We hypothesize that this threshold is determined by the redundancy of the mapping, which is $2^{16} : 2^8$ in our experiments, but we did not perform further experiments to investigate this question.

This threshold is important when considering neutral degree as a measure for mutational robustness. Our results suggest that high entropy can only be guaranteed up to a certain mean neutrality degree, and with decreasing entropy more and more phenotypes become inaccessible to the evolutionary system through the genotype-phenotype mapping. Therefore it must be discussed if the notion of robustness can be applied to mappings that are constant or not surjective. Put differently, given two redundant genotype-phenotype mappings ϕ_1 and ϕ_2, both defined over the same genotype and phenotype spaces. Then the question must be asked whether the mapping ϕ_1 with a smaller image, that is $|\{\, \phi_1(g) \mid g \in \mathcal{G} \,\}| < |\{\, \phi_2(g) \mid g \in \mathcal{G} \,\}|$, is inevitably the mapping with higher mutational robustness. We think that non-surjective genotype-phenotype mappings must be considered when defining robustness.

The above work on analyzing robustness and evolvability trade-off of redundant genetic representations for simulated evolution demonstrates that the Pareto-approach is of great value in uncovering the relationship between redundancy and evolvability of genetic representations. In the following, we will apply the Pareto approach to studying machine learning systems.

IV. Pareto Analysis of Learning Systems

A. Trade-offs in Learning

Machine learning can be largely divided into supervised learning, unsupervised learning and reinforcement learning. Any machine learning method can be seen as an optimization problem, because the target of learning is either to minimize a cost function or maximize a reward or value function.

Machine learning has traditionally been treated as a single objective optimization problem. However, if we examine different learning problems more closely, most of them have to deal with more than one objective. Let us first take a look at model selection strategies, which are the most important issue in supervised or unsupervised learning.

A well-known criterion for model selection is the Akaike's Information Criterion (AIC):

$$AIC = -2log(\mathcal{L}(\theta|y, g) + 2K, \tag{19}$$

where g is the true function from which the training data are produced, θ is the model, $\mathcal{L}(\theta|y, g)$ is the maximized log-likelihood between the model and the true function given the training data y, and K is the effective number of parameters to be estimated in the model θ. For a number of different models trained from the same training data set, the one with the minimal AIC will be chosen. From Equation 19, we can see that AIC consists of two terms. The first term is to maximize the accuracy of the model, whereas the second term K indicates the complexity of the model. As the model complexity increases, i.e., the number of free parameters in the model increases, the first term in the AIC tends to decrease, while the second becomes larger, which in fact reflects the well-known bias-variance dilemma. From the optimization point of view, these two terms cannot be minimized simultaneously and therefore, model selection, or more generally, machine learning is actually a typical multi-objective optimization problem.

In supervised learning of neural networks and fuzzy systems, several specific trade-off criteria have been considered derived from the fundamental trade-off relationship defined by the AIC. The most commonly used criterion for model

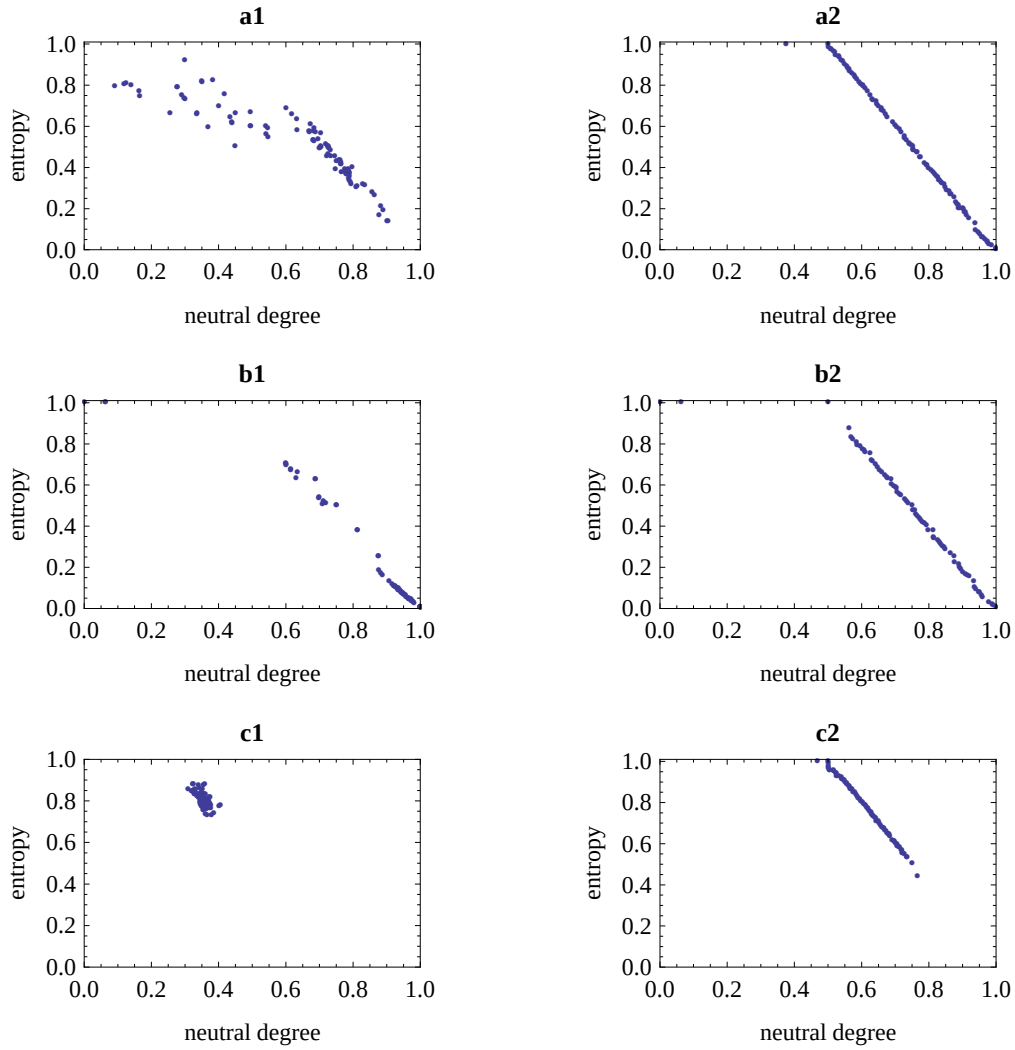


Fig. 7. Multi-objective optimization evaluation of the Boolean model with ($n = 16$, $m = 8$) for maximal mean neutral degree and uniformity entropy. The model is encoded by different techniques: **a** Multiple Boolean functions encoding with elementary arity 2. **b** Single Boolean function encoding with elementary arity 2. **c** Majority rule encoding. The figures **1** in the first columns show 75 randomly initialized genotype-phenotype mappings with $n = 16$ and $m = 8$. The figures **2** in the second column the approximated Pareto optimal set of genotype-phenotype mappings after 100 generations. The different encodings result in similar approximations of a Pareto optimal set.

fidelity is the error function, such as the mean square error or the mean absolute error. Complexity of learning models is also very often used, including the Gaussian regularizer that is the sum of the squared weights of the neural network, or the Laplacian regularizer, which is the sum of the absolute weights. If a non-gradient based algorithm is used for learning, the number of hidden nodes or the number of connections can also be used to denote the complexity of the network. In generating fuzzy systems, complexity of the fuzzy rules is directly associated with their interpretability, in other words, the easiness for human users to understand the knowledge represented by the fuzzy rules [26], [27]. Reducing the number of fuzzy subsets, the overlaps of the fuzzy subsets, the number of fuzzy rules and the rule length (in the rule premises) will improve the interpretability of fuzzy rules. When ensembles of learning models are to be generated, either functional or structural diversity of the ensemble members can be used as the objective in ensemble generation [8].

Although the bias-variance trade-off in the AIC has not been explictly taken into account in many learning algorithms, it is explicitly considered in the learning algorithms of support vector machines (SVMs). Without loss of generality, we take SVMs for classification as an example. A typical SVM for classification can be expressed as the following optimization problem:

$$\text{minimize} \quad \tfrac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\xi_i, \tag{20}$$

$$\text{subject to} \quad y_i(\mathbf{w}^T\phi(x_i) + b) \geq 1 - \xi_i, \tag{21}$$

$$\xi_i \geq 0, \tag{22}$$

$$\tag{23}$$

where $\mathbf{x}_i \in R^n$, $y_i \in \{-1, 1\}$, $i = 1, 2, \cdots, N$ are the training data set consisting of N data pairs, $\phi(x_i)$ is a kernel

function, b is the bias, ξ_i is called interior deviation or slackness, and C is a constant to be determined by the user. Obviously, minimizing $\mathbf{w}^T\mathbf{w}$ and minimizing $\sum_{i=1}^{N}\xi_i$ are two conflicting objectives, and solving the SVM is a multi-objective optimization problem [28].

One of the main principles in information processing in the brain is that the fraction of neurons that are strongly active at a time is below 1/2. This principle has been implemented in computational neuroscience, which results in a class of computational models known as sparse coding [29], where the accuracy of the model trades off with the sparseness of the active neurons in the model. So far, the coefficient determining the trade-off between the two terms has been chosen heuristically, and no in-depth discussion about the influence of the co-efficient on the final performance has been reported.

Another well-known issue in computational cognitive neuroscience is the stability-plasticity dilemma [30], which means that the learning system should be able to learn new information efficiently without completely forgetting what has been learned previously. The stability versus plasticity dilemma is often known as catastrophic forgetting in neural network based machine learning [31]. Most existing techniques try to alleviate catastrophic forgetting with models using distributed representation or a growing structure. Direct rehearsal where the previous training data is assumed to be available, or pseudo-rehearsal using pseudo-data generated from the trained model has also been suggested to address catastrophic forgetting [32].

B. Illustrative Examples

In this section, we will provide two examples of Pareto-based multi-objective learning. The first example shows how catastrophic forgetting can be approached using the Pareto-based approach. In the second example, neural network regularization is tackled with by reformulating regularization as a Pareto-based multi-objective optimization problem. In addition, we demonstrate that by analyzing the accuracy-complexity trade-off, it is able to identify Pareto optimal solutions (learning models) with a complexity that most likely matches that of the problem in question.

1) Alleviating Catastrophic Forgetting: Catastrophic forgetting means that when a trained neural network learns new patterns, the already learned information (the base patterns) will be destroyed (forgotten). Since learning the base patterns and learning the new patterns are very likely competitive, it is natural to deal with the conflicting objectives using the Pareto-based multi-objective learning [33]. In that work, pseudo-rehearsal is reformulated as a multi-objective optimization problem, as shown in Fig. 8. Before training the neural network with the new patterns, a number of pseudo-patterns are generated. The pseudo-patterns are created by generating random inputs to the trained neural network, and then recording the corresponding outputs of the network. It is hoped that the pseudo-patterns will carry the same or similar knowledge as the base patterns. Next, learning is carried out as a bi-objective optimization problem, where

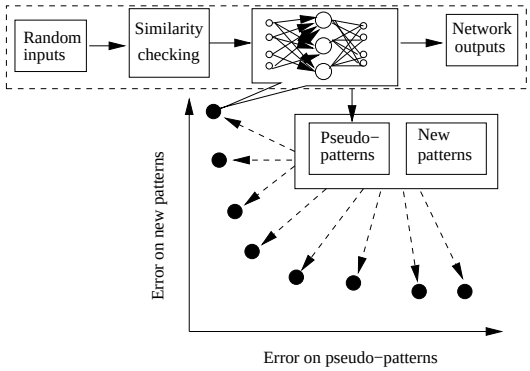


Fig. 8. Pseudo-rehearsal through multi-objective optimization.

the approximation error on the new patterns and that on the pseudo-patterns serve as the two objectives.

It has been found that it is non-trivial to properly perform life-time learning within the evolutionary cycles in multi-objective learning. This is particularly true when all the objectives are concerned with approximation error. In the case of alleviating catastrophic forgetting, three life-time learning strategies have been investigated [33]. In strategy 1, the union of pseudo-patterns and new patterns are used in life-time learning. Strategy 2 randomly chooses either the pseudo-patterns or the new patterns for lifetime learning. And in strategy 3, one of the three patterns, the pseudo-patterns, the new patterns, and the union of pseudo-patterns and new patterns, is picked out randomly for life-time learning. It is found that using the union of pseudo-patterns and new patterns in life-time only will dramatically reduce the population diversity and only a few Pareto optimal solutions can be obtained. The diversity of solutions will be improved when strategy 2 or 3 is used. Fig. 9 shows the trade-off solutions in terms of memorized based and new patterns. In this experiment, 25 pairs of base and new patterns of are generated randomly, each pair of pattern containing a 10-dimensional input and a 10-dimensional output of value either 0 or 1. Then, neural networks with a maximum of 10 hidden nodes are evolved to learn the base patterns. After the training is completed, 24 of the 25 based patterns are correctly learned. Then, 25 pseudo-patterns are generated based on the trained neural network. From Fig. 9, we can see that the Pareto-optimal solutions are able to memorize more than 15 base patterns while learning over 20 new patterns.

2) Neural Network Regularization: To improve generalization of neural networks, regularization techniques are often adopted by including an additional term in the error function:

$$J = E + \lambda\Omega, \tag{24}$$

where λ is a hyperparameter that controls the strength of the regularization, Ω is known as the regularizer, and E is usually the mean square error (MSE):

$$f = \frac{1}{N}\sum_{i=1}^{N}(y(i) - y^d(i))^2, \tag{25}$$

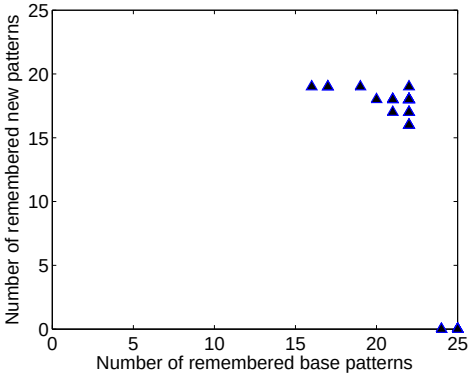


Fig. 9. Number of based and new patterns memorized by the Pareto-optimal solutions.

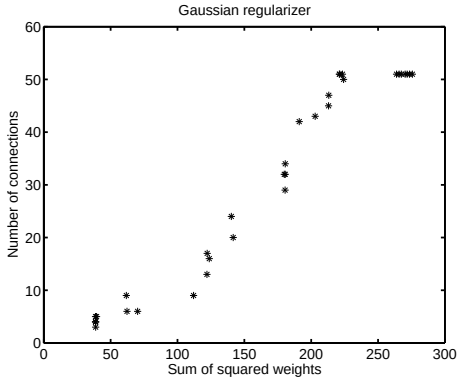


Fig. 10. Relationship between the number of connections and the sum of squared weights.

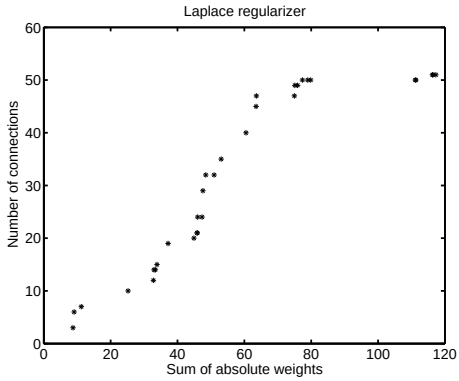


Fig. 11. Relationship between the number of connections and the sum of absolute weights.

where $y(i)$ and $y^d(i)$ are the model output and the desired output, respectively, and N is the number of data pairs in the training data.

The most popular regularization method is known as weight decay (also known as Gaussian regularizer):

$$\Omega = \frac{1}{2}\sum_k w_k^2, \tag{26}$$

where k is an index summing up all weights.

One weakness of the weight decay method is that it is not able to drive small irrelevant weights to zero, which may result in many small weights. The following regularization term has been proposed to address this problem (known as Laplacian regularizer):

$$\Omega = \sum_i |w_i|. \tag{27}$$

This regularization was used for structure learning, because it is able to drive irrelevant weights to zero.

It is quite straightforward to see that the neural network regularization in equation (24) can be reformulated as a bi-objective optimization problem:

$$\min \{f_1, f_2\} \tag{28}$$

$$f_1 = E, \tag{29}$$

$$f_2 = \Omega, \tag{30}$$

where E is usually the mean square error, and Ω is one of the regularization terms defined in equation (26) or (27).

Since evolutionary algorithms are used to implement regularized learning of neural networks, a new and more direct index for measuring complexity of neural networks can be employed, which is the number of connections in the neural network:

$$\Omega = \sum_i \sum_j c_{ij}, \tag{31}$$

where c_{ij} equals 1 if there is connection from neuron j to neuron i, and 0 if not.

When gradient-based learning algorithms are employed for regularization, the Laplace regularizer is usually believed to be better than a Gaussian regularizer in that the Laplace regularizer is able to drive irrelevant weights to zero. In this way, "structural learning" is realized with the help of the Laplace regularizer. Using the evolutionary multi-objective approach, we show that there is no substantial difference between the Gaussian and the Laplace regularizers in terms of their ability to realize structural learning, when evolutionary algorithms are used as an optimizer.

To verify this assumption, we use the Breast Cancer Data set available in the UCI Machine Learning Repository [34]. The available data are split into a training data set and a test data set, where 525 instances are used for training and 174 instances for test.

From Figs. 10 and 11, we can see that a similar relationship between the sum of squared or absolute weights and the number of connections is observed. In other words, even when the Gaussian regularizer is used, the number of connections can also be reduced to the minimum when the sum of squared weights is minimized. It does not result in many small weights as when gradient-based learning algorithms are used.

3) Identifying Models of Suitable Complexity: In this section, we show that the Pareto-approach to handling the accuracy-complexity trade-off provides an empirical, yet interesting alternative to selecting models that have good

generalization on unseen data. The basic argument is that the complexity of the model should match that of the data to be learned and the ability of the learning algorithm. When the complexity of the model is overly large, learning becomes sensitive to stochastic influences, and results on unseen data will be unpredictable, i.e., overfitting can happen. To analyze the relationship between the achieved accuracy-complexity trade-off solutions, the normalized performance gain (NPG) is defined:

$$NPG = \frac{MSE_j - MSE_i}{C_i - C_j}, \tag{32}$$

where MSE_i, MSE_j, and C_i, C_j are the MSE on training data, and the number of connections of the i-th and j-th Pareto optimal solutions. When the solutions are ranked in the order of increasing complexity, the following relationships hold:

$$\begin{aligned} C_{i+1} &> C_i, \\ MSE_{i+1} &\leq MSE_i. \end{aligned}$$

We hypothesize that if the model complexity is lower than that of the data, an increase in complexity will result in significant increase in performance (NPG). As the complexity continues to increase, the NPG reduces gradually to zero. At this point, the complexity of the model matches that of the data. Further increase in complexity will probably bring about further enhancement in performance on the training data, but with a dramatically increasing risk of overfitting the training data.

We are now going to verify empirically the suggested method for model selection also on the Breast Cancer Data set. The Pareto fronts generated from two independent runs on the three benchmark problems are presented in Fig. 12. The dots denote the results on the training data set, while the circles the results on test data. The NPG from the two independent runs for the three problems are plotted in Fig. 13. It can be seen from Fig. 12 that models with the number of connections larger than 10 to 15 start to overfit the data, which roughly corresponds to the point in Fig. 13 where the NPG drops to zero after the first peak in performance gain. This empirical result is helpful for model selection when the number of training data is too small to perform meaningful cross-validations and other statistical analyses.

V. Conclusions

This paper discusses the Pareto-based multi-objective analysis of evolutionary and learning systems. With an evolutionary multi-objective optimization algorithm, a number of Pareto-optimal solutions (forming the Pareto front) can be obtained. By analyzing the Pareto front, we are able to gain a deeper insight into in the evolutionary and learning systems. We show in the first illustrative example how the Pareto approach can be used to analyze the robustness-evolvability trade-off in a class of redundant Boolean representations for simulated evolution.

For learning systems, we show that the Pareto-based approach is able to find solutions that can learning new

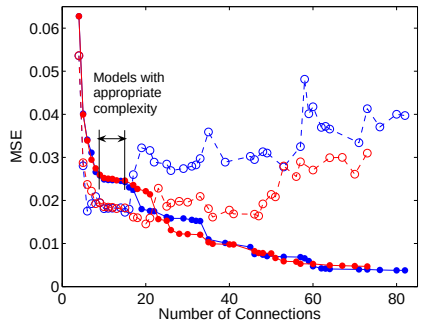
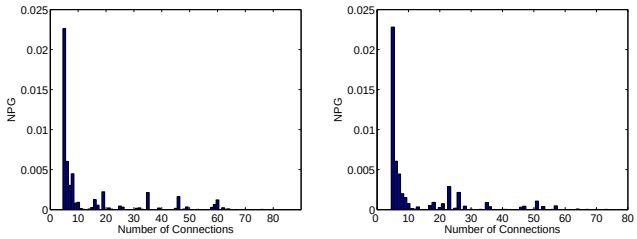


Fig. 12. Accuracy versus complexity of the Pareto-optimal solutions from two independent runs: Breast Cancer Data. Dots denote training data, and circles test data.



(a) (b)

Fig. 13. Normalized performance gain from two independent runs for the Breast Cancer Data.

knowledge without a serious interference with the already learned knowledge. In the multi-objective approach to neural network regularization, we demonstrate that the Gaussian regularizer works as efficient as the Laplacian regularizer in reducing the complexity of neural networks, when an evolutionary optimization is employed. In addition, we hypothesize that Pareto optimal solutions around the knee point are those having the appropriate complexity for the given data, which are most likely to generalize on unseen data.

Many interesting issues remain to be investigated. In this work, a simple and stationary Boolean representation is used to study robustness and evolvability of genetic representations. This should be extended to more complex, in particular dynamic genotype-phenotype mappings described e.g. by random Boolean networks or ordinary differential equations. In multi-objective learning, it is interesting to compare the convergence speed of single and multi-objective learning. Multi-objective approaches to the analysis of sparse coding still lacks. We believed the Pareto-based multi-objective approach will release much burden in tuning parameters and thus help to achieve a better understanding the the problem at hand.

Acknowledgment

The authors would like to thank Ingo Paenke for the interesting discussions on the work on multi-objective optimization of redundant representations.

REFERENCES

[1] L. Ancel and J. Bull, "Fighting change with change: Adaptive variation in an uncertain world," *Trends in Ecology and Evolution*, vol. 17, no. 12, pp. 551–557, 2002.

[2] G. Tononi, O. Sporns, and G. Edelman, "A measure for brain complexity: Relating functional segregation and integration in the nervous system," *PNAS*, vol. 91, pp. 5033–5037, 1994.

[3] J. Teo and H. Abbass, "Multiobjectivity and complexity in embodied cognition," *IEEE Transactions on Evolutionary Computation*, vol. 9, no. 4, pp. 337–360, 2005.

[4] S. Louis and G. Rawlines, "Pareto optimality, GA-easiness and deception," in *The Fifth International Conference on Genetic Algorithms*. Morgan Kaufmann, 1993, pp. 118–123.

[5] J. Knowles, R. Watson, and D. Corne, "Reducing local optima in single-objective problems by multi-objectivization," in *EMO 2001*, ser. LNCS 1993. Springer, 2001, pp. 269–283.

[6] M. Jensen, "Helper-objectives: Using multi-objective evolutionary algorithms for single-objective optimization," *Journal of Mathematical Modeling and Algorithms*, vol. 3, no. 4, pp. 323–347, 2004.

[7] L. Bui, J. Branke, and H. Abbass, "Multi-objective optimization for dynamic environments," in *Congress on Evolutionary Computation*. IEEE, 2005, pp. 2349–2356.

[8] Y. Jin and B. Sendhoff, "Pareto-based multi-objective machine learning: An overview and case studies," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 38, no. 3, pp. 397–415, 2008.

[9] J. Allman, *Evolving Brains*. Scientific American Library, 1999.

[10] B. Jones, Y. Jin, X. Yao, and B. Sendhoff, "Evolution of neural organization in the hydramat - a computational model of a radially-symmetric organism," *ACM/IEEE Transactions on Computational Biology and Bioinformations*, 2008, in revision.

[11] S. Stearns, "Trade-offs in life-history evolution," *Functional Ecology*, vol. 3, pp. 259–268, 1989.

[12] A. Mukhopadhyay and H. Tissenbaum, "Reproduction and longevity: secrets revealed by c. elegans," *Trends in Cell Biology*, vol. 17, no. 2, pp. 65–71, 2007.

[13] I. Paenke, Y. Jin, and J. Branke, "Balancing population and individual level adaptation in changing environments," *Adaptive Behavior*, 2008, submitted.

[14] E. Charnov and S. Ernest, "The offspring-size / clutch-size trade-off in mammals," *The American Naturalist*, vol. 167, no. 4, pp. 578–582, 2006.

[15] J. Handl, D. Kell, and J. Knowles, "Multi-objective optimization in computational biology and bioinformatics," *ACM/IEEE Transactions on Computational Biology and Bioinformations*, vol. 4, pp. 279–292, 2007.

[16] H. Kitano, "Biological robustness." *Nat Rev Genet*, vol. 5, no. 11, pp. 826–37, 2004.

[17] A. Wagner, *Robustness and evolvability in living systems*. Princeton University Press, 2007.

[18] R. Lenski, J. Barrick, and C. Ofria, "Balancing robustness and evolvability," *PLOS Biology*, vol. 4, no. 2, p. e428, 2006.

[19] S. Cilibert, O. Martin, and A. Wagner, "Innovation and robustness in complex gene networks," *PNAS*, vol. 104, no. 34, pp. 13 591–13 596, 2007.

[20] K. Deb, S. Agrawal, A. Pratab, and T. Meyarivan, "A Fast Elitist Non-Dominated Sorting Genetic Algorithm for Multi-Objective Optimization: NSGA-II," in *Proceedings of the Parallel Problem Solving from Nature VI Conference*, M. Schoenauer, K. Deb, G. Rudolph, X. Yao, E. Lutton, J. J. Merelo, and H.-P. Schwefel, Eds. Paris, France: Springer. Lecture Notes in Computer Science No. 1917, 2000, pp. 849–858. [Online]. Available: citeseer.ist.psu.edu/deb00fast.html

[21] P. K. Lehre and P. C. Haddow, "Phenotypic complexity and local variations in neutral degree," *Biosystems*, vol. 87, pp. 233–242, Feb. 2007. [Online]. Available: http://www.sciencedirect.com/science/article/B6T2K-4KVJCK6-8/2/099604f1791eecd1e23f7ea94729a5a9

[22] M. Kirschner and J. Gerhart, "Evolvability," *PNAS*, vol. 95, no. 15, pp. 8420–8427, 1998.

[23] P. Fernández and R. V. Solé, "Neutral fitness landscapes in signalling networks," *Journal of the Royal Society, Interface / the Royal Society*, vol. 4, pp. 41–7, 2007.

[24] B. M. R. Stadler, P. F. Stadler, G. P. Wagner, and W. Fontana, "The topology of the possible: Formal spaces underlying patterns of evolutionary change," *Journal of Theoretical Biology*, vol. 213, no. 2, pp. 241–274(34), 2000.

[25] R. Gruna, "Analysis of redundant genotype-phenotype mappings - Investigation of the effect of neutrality on evolvability and robustness," Master's thesis, AIFB, Universität Karlsruhe, 2007.

[26] Y. Jin, "Fuzzy modeling of high-dimensional systems: Ccomplexity reduction and interpretability improvement," *IEEE Transactions on Fuzzy Systems*, vol. 8, no. 2, pp. 212–221, 2000.

[27] ——, *Advanced Fuzzy Systems Design and Applications*. Heidelberg: Physica Verlag/Springer Verlag, 2003.

[28] C. Igel, "Multi-objective model selection for support vector machines," in *Evolutionary Multi-Criterion Optimization*, ser. LNCS 3410, 2005, pp. 534–546.

[29] B. Olshausen, "Relations between the statistics of narutal image and the response property of cortical cells," *Nature*, vol. 381, pp. 607–609, 1996.

[30] W. Abraham and A. Robins., "Memory retention - the synaptic stability versus plasticity dilemma," *Trends in Neuroscience*, vol. 28, no. 2, pp. 73–78, 2005.

[31] M. McCloskey and N. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," *The Psychology of Learning and Motivation*, vol. 24, pp. 109–165, 1989.

[32] R. French, "Catastrophic forgetting in connectionist networks," *Trends in Cognitive Sciences*, vol. 3, no. 4, pp. 128–135, 1999.

[33] Y. Jin and B. Sendhoff, "Alleviating catastrophic forgetting via multi-objective learning," in *International Joint Conference on Neural Networks*, 2006, pp. 3335–3342.

[34] L. Prechelt, "PROBEN1 - A set of neural network benchmark problems and benchmarking rules," Fakultät für Informatik, Universität Karlsruhe, Tech. Rep., 1994.

[35] J. Handl and J. Knowles, "Exploiting the tradeoff - The benefits of multiple objectives in data clustering," in *Evolutionary Multi-Criterion Optimization*, ser. LNCS 3410. Springer, 2005, pp. 547–560.