Contrasting the landscapes of feature selection under different machine learning models

Arnaud Liefooghe, Ryoji Tanabe, Sébastien Verel

Contrasting the Landscapes of Feature Selection under Different Machine Learning Models

Arnaud Liefooghe[1][0000−0003−3283−3122], Ryoji Tanabe[2][0000−0003−4049−0393], and Sébastien Verel[1][0000−0003−1661−4093]

[1] LISIC, Université du Littoral Côte d'Opale, France
{arnaud.liefooghe,sebastien.verel}@univ-littoral.fr
[2] Yokohama National University, Yokohama, Japan
{tanabe-ryoji-sn}@ynu.ac.jp

Abstract. Feature selection plays a crucial role in improving the performance of machine learning (ML) models for various prediction tasks and in explaining their recommendations. Feature selection can be defined as an optimization problem whose evaluation function calls on an ML algorithm — a method known as the *wrapper* approach. While a thorough understanding of the landscape of the feature selection problem might help guide the development of efficient evolutionary algorithms and algorithm selection technologies, only a couple of previous studies have explored this problem's landscape. In addition, only k-nearest neighbors classification is typically used as an ML model. This paper investigates how the choice of an ML model influences the search difficulty of the feature selection problem. Specifically, we examine the feature selection problem with 14 classification datasets and 6 ML models by means of landscape analysis and local optima networks, and we relate them to the performance of three feature selection algorithms. Our findings have important implications for feature selection problems and algorithms.

Keywords: Feature selection · Machine learning · Landscape analysis

1 Introduction

Feature selection consists in choosing the most effective features from a given set to maximize the prediction accuracy of a machine learning (ML) model [3, 6]. Discarding irrelevant or redundant features can help prevent the ML model from overfitting. Existing techniques for feature selection include filter, embedded, and wrapper approaches [6]. Unlike the first two approaches, wrapper feature selection performs iterative optimization of a subset of features. The quality of a feature subset is evaluated by actually training an ML model using those selected features as predictors. Evolutionary wrapper feature selection has demonstrated its promising performance in the literature [4, 24]. Given a dataset with n features, the feature selection problem seeks a subset of $p \leqslant n$ features that maximizes the accuracy of the considered ML model. As such, candidate solutions can be represented by a binary string of length n. Each position in the binary

string indicates whether the corresponding feature is selected or not. Then, the fitness function corresponds to a prediction score, such as classification accuracy.

We note that evaluating the quality of a feature subset is computationally intensive. Each call to the evaluation function requires training an ML model. However, despite being time-consuming due to its iterative nature, the wrapper approach typically yields a better feature subset than the filter and embedded approaches [4, 24]. This explains why computationally cheap ML algorithms, like k-nearest neighbors (`kNN`), are generally used in wrapper feature selection [4]. However, it remains unclear whether the challenges and solutions induced by other ML models are the same.

Landscape analysis plays a crucial role in understanding the structure of optimization problems through the lens of search algorithms [18]. Insights gained from landscape analysis are not only useful for understanding algorithm performance, but also for designing new efficient optimizers and selecting the most suitable approach [2, 9, 10]. In recent years, analyzing the landscape of hyperparameter optimization problems in ML has emerged as a popular topic in the evolutionary computation community; see, e.g., [16, 17, 19, 20].

Unfortunately, the landscape analysis of the feature selection problem has attracted less attention. In fact, this issue has only been addressed in couple of previous studies from Mostert et al. [11, 12]. In [11], the authors conduct a landscape analysis of the feature selection problem in terms of fitness distribution, fitness level, and neutrality. They find that the filter method generally outperforms the wrapper sequential feature selection method on instances with large neutral regions. In [12], they further analyze the landscape of the feature selection problem by means of local optima networks (LONs) [14]. They find that irrelevant features create plateaus in the landscape, which could be harmful to search algorithms that stagnate due to these equivalent solutions. However, both of their analyses focus on `kNN` only. As such, it is still unclear how the ML model actually shapes the landscape of the feature selection problem.

Inspired by the above discussions, this paper aims at improving our understanding of feature selection by exploring how the choice of the ML model influences the problem landscape. Our contributions can be summarized as follows:

(1) We contrast the landscapes and LONs resulting from 14 classification datasets with varying numbers of features, classes, and observations.
(2) We contrast the landscapes and LONs resulting from 6 established ML models that use different induction approaches.
(3) We connect our findings from landscape analysis to the actual performance of three established feature selection algorithms.

The paper is structured as follows. Section 2 provides preliminary information. Section 3 outlines the experimental setup. Section 4 presents the results of our analysis. Section 5 concludes the paper and discusses further research.

2 Background

We start by introducing feature selection and landscape analysis below.

2.1 The Feature Selection Problem

Given a set V of n features and a score function f evaluating the performance of an ML algorithm, the feature selection problem seeks a subset of features S that achieves the maximum score value:

$$S^{\diamond} = \arg\max_{S \subseteq V} f(S).$$

In the context of evolutionary feature selection, a subset S is generally represented by an n-dimensional binary string $x = (x_1, \ldots, x_n)^{\top}$. For each feature $i \in \{1, \ldots, n\}$, $x_i = 1$ when the corresponding feature is included in S, and $x_i = 0$ otherwise. The total number of feasible subsets is thus 2^n. The wrapper feature selection problem can be seen as a black-box pseudo-Boolean optimization problem. Multiple evolutionary algorithms have been proposed to solve this problem in the literature. The reader is referred to [4, 24] for a review.

2.2 Landscape Analysis

We define the fitness landscape of feature selection as a triplet $(\mathcal{X}, \mathcal{N}, f)$:

- $\mathcal{X}$ is the search space, that is the set of bitstrings of length n, with $|\mathcal{X}| = 2^n$.
- $\mathcal{N}: \mathcal{X} \mapsto 2^{\mathcal{X}}$ is a neighborhood relation. Following previous studies on this problem domain [22], we use the 1-bit-flip relation: two solutions are neighbors if their Hamming distance is one.
- $f: \mathcal{X} \mapsto \mathbb{R}$ is a fitness function, that is the ML prediction score or, more specifically, the classification accuracy. We assume f is to be maximized.

A solution $x^{\diamond} \in \mathcal{X}$ is a *global optimum* if there is no $x \in \mathcal{X}$ such that $f(x^{\diamond}) < f(x)$. A solution $x^{\star} \in \mathcal{X}$ is *local optimum* if there is no $x \in \mathcal{N}(x^{\star})$ such that $f(x^{\star}) < f(x)$. The distribution and connectivity of local optima in the landscape are crucial as they act as attraction points for search, consequently hindering the ability to reach a global optimum. In addition to traditional landscape analysis measures, we consider local optima networks for studying them, as in [12].

The *local optima network* (LON) [14] adapts the idea of representing physical energy landscapes as complex networks [5] in order to condense the information from the landscape into a weighted graph of local optima. A LON is defined as a directed, weighted graph $G = (N, E)$. The nodes N of the graph represent local optima. An edge $e \in E$ exists between two nodes if there is a non-zero probability for the search process to transition from one node to another. To be more specific, the basin of attraction of a local optimum $x^{\star} \in \mathcal{X}$ refers to the set of solutions that converge to $x^{\star}$ when applying a simple hill-climbing local search. The count of these solutions is the size of the basin and is used as the *width* of a LON node. We follow the concept of escape edges in LONs [21]: The *weight* of an edge represents the probability of transitioning from one local optimum's basin of attraction to another, upon applying a perturbation followed by a local search. As in [12], we set the perturbation strength to 2 bit-flips.

In the monotonic LON, only the edges that lead to an improved local optimum are kept. In the *compressed monotonic LON* (CM-LON), nodes with the same fitness value are collapsed, and any duplicate edges are aggregated [13].

Table 1: 14 classification datasets considered in this work.

dataset	nominal	numerical	classes	data	features	$\lvert\mathcal{X}\rvert$
⋆ diabetes	×	○	2	768	8	256
⋆ breast-cancer	○	×	2	286	9	512
⋆ breast-w	×	○	2	699	9	512
⋆ page-blocks	×	○	5	5473	10	1 024
⋆ vowel	○	○	11	990	12	4 096
⋆ heart-statlog	×	○	2	270	13	8 192
schizo	○	○	2	340	14	16 384
credit-approval	○	○	2	690	15	32 768
⋆ zoo	○	○	7	101	16	65 536
vote	○	×	2	435	16	65 536
pendigits	×	○	10	10 992	16	65 536
letter	×	○	26	20 000	16	65 536
vehicle	×	○	4	846	18	262 144
lymph	○	○	4	148	18	262 144

3 Experimental Setup

This section outlines the experimental setup of our analysis. Although various performance measures exist for ML, we simply focus on classification accuracy in this work. We consider the following 6 ML algorithms for classification: `kNN`, support vector classification (`SVC`), logistic regression (`LR`), decision tree (`DT`), random forests (`RF`), and naive Bayes (`NB`). We employ the implementation available in `scikit-learn` [15] with default parameters — see https://scikit-learn.org/stable/. While `kNN`, `SVC`, `LR`, and `NB` are deterministic algorithms, `DT` and `RF` are not. In order to minimize the effect of randomness, we conduct ten independent executions of `DT` and `RF` on each dataset. On top of that, we conduct a 5-fold cross-validation for all methods. We use the same instantiation of cross-validation folds across runs. The average score across folds and runs is then considered. All (average) score (i.e. fitness) values are rounded to 10^{-9} to prevent numerical issues.

Table 1 describes the 14 classification datasets considered in this work. The previous study on the landscape of feature selection [12] used seven of those — marked with a star ($\star$) in Table 1. They were extracted from the `UCI` repository [1]. We supplement them with seven additional datasets to aim for more generalizable results. As pointed out in [4], the `breast-cancer`, `vote`, `heart-statlog`, `zoo`, and `lymph` datasets have often been used for benchmarking evolutionary feature selectors due to their small number of observations. The number of features in all datasets is at most 18. Like Mostert et al. [12], we focus on datasets with few features. This allows us to fully enumerate all 2^n solutions. Indeed, information about all 2^n solutions is required to compute the exact LON and other landscape characteristics. Investigating datasets with more features would present several challenges that we leave open for future work. We converted nominal features into numerical features using one-hot encoding to ensure the experimental condition is the same for all ML algorithms. We further remark that the empty solution $x_0 = (0, \ldots, 0)^\top$ implies that the ML model cannot use any feature. To prevent any bias, we decided to deem x_0 as unfeasible and to

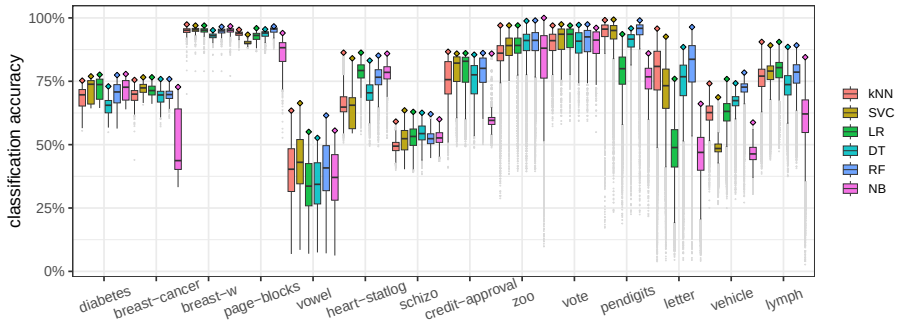


Fig. 1: Distribution of classification accuracy (i.e. fitness) values. The best value for each dataset and ML model is represented with a diamond shape.

discard it from our analysis. Considering the 14 datasets and 6 ML models, this leads to a total of 84 different subset selection problems (or landscapes).

Finally, we apply three wrapper methods to the considered problems: the forward sequential feature selector (`F-SFS`), the backward SFS (`B-SFS`), and a genetic algorithm (`GA`). Both `F-SFS` and `B-SFS` are well-established approaches [15]. The `GA` follows a simple steady-state approach with uniform crossover and standard bit-flip mutation with a rate of $1/n$. The population size is set to n, and the maximum number of calls to the evaluation function is set to n^2.

4 Empirical Results and Discussion

This section presents the dissimilarity of landscapes across problems and the results of feature selection algorithms relative to the landscape analysis.

4.1 Distribution of Fitness Values

We begin by simply reporting the distribution of fitness values, measured in terms of classification accuracy, for each model and dataset in Fig. 1. We observe significant differences in the range of values between ML models and datasets. However, this does not necessarily imply that some ML models are superior, as the quality of each subset of features is inherent to the considered ML model. Additionally, the fitness values span larger ranges in some cases. This suggests that some ML models may generate solutions that are more similar or equivalent than others, a characteristic known as *landscape neutrality*. We will delve deeper into this in subsequent discussions.

4.2 Correlation of Solution Rankings Produced by ML Models

We continue by measuring the correlation between the relative ranking of solutions produced by the different ML models. For each dataset and each pair of models, Fig. 2 gives the Spearman rank correlation coefficient between the

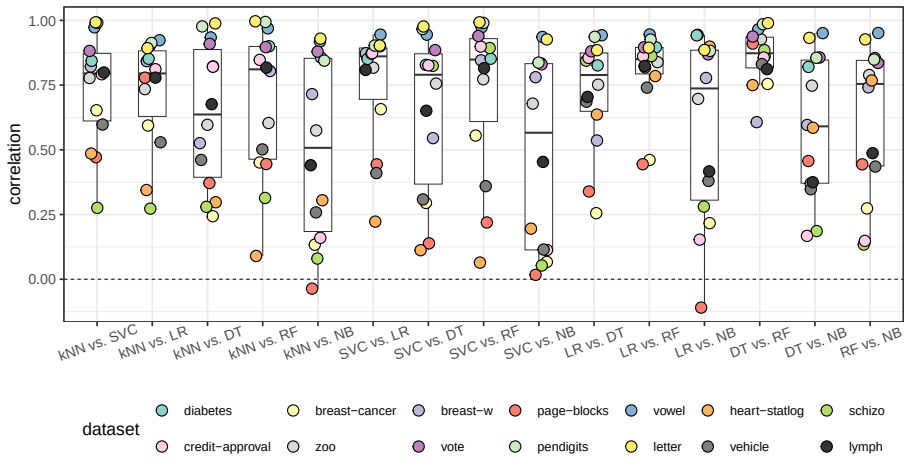


Fig. 2: Correlation of solution fitness values between each pair of ML models.

accuracy of all solutions from the landscape. This gives an indication of the level of agreement between the two models in evaluating the fitness of solutions. For each pair of models, a point represents a specific dataset while the boxplot summarizes the distribution of coefficients. We observe that some ML models agree more than others, although none of them perfectly align on all datasets.

We can see both small and large correlations for some datasets. Overall, there is a small correlation on the `heart-statlog` and `schizo` datasets in most cases when comparing `kNN` with other ML models. Likewise, the solution rankings between `NB` and others are mostly uncorrelated on the `breast-cancer`, `page-blocks`, and `credit-approval` datasets. Interestingly, a significant correlation exists between any pair of ML models on the `vowel`, `pendigits`, and `letter` datasets, all of which have a high number of classes.

4.3 Global Optima

Fig. 3 reports the count of global optima for each dataset and ML model. We follow [12] and report the number of global optima *plateaus*, such that neutral networks are collapsed. This means that any set of global optima that are connected by the neighborhood relation counts as one. For feature selection, this implies that one or more features actually have no effect on classification accuracy, regardless of whether they are included or excluded in the optimal subset.

In most problems, there is a single global optima plateau. The only exception is for the `zoo` dataset, where `kNN` reveals four, while `SVC` and `LR` have two. `LR` also has two global optima plateaus for `breast-w` and `lymph`, and three for `credit-approval` and `vote`. The only case where `NB` has two global optima plateaus is for `lymph`. These observations could imply that finding a global optimum is easier for these problems. However, we will later see that this neutrality also occurs at sub-optimal levels, which can potentially mislead the search. Delv-

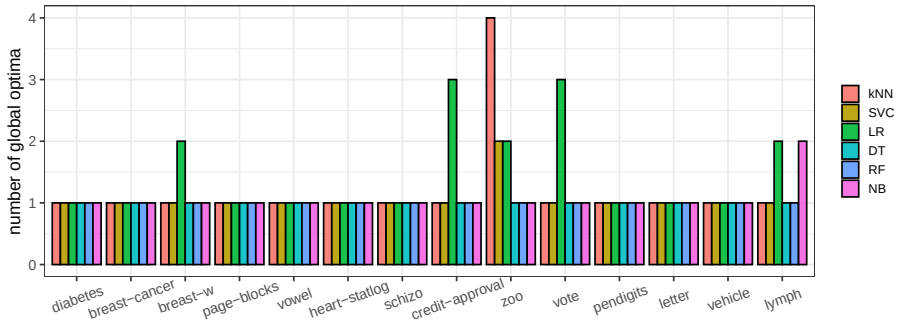


Fig. 3: Number of global optima.

ing deeper into the analysis of global optima, we find that the optimal proportion of selected features remains fairly consistent across ML models — details are not provided due to space restriction. While there are differences between datasets, there is no significant difference between the ML models.

4.4 Fitness-Distance Correlation

We continue our analysis of the global structure of the landscape with the fitness-distance correlation (`FDC`) [8]. As one of the earliest metrics from landscape analysis, `FDC` estimates how the fitness function properly guides the search towards the global optimum. Fig. 4 reports the Pearson correlation, from all solutions in the landscape, between classification accuracy (fitness) and the Hamming distance to the nearest global optimum. According to Jones and Forrest [8], for maximization, the closer the `FDC` to -1, the more *straightforward* the problem: fitness increases as we approach the global optimum. Conversely, an `FDC` close to 1 is *misleading* for search, while an `FDC` around 0 makes the problem *difficult* due to the lack of correlation between fitness and distance.

We note large variations in the `FDC` depending on the dataset. For instance, using `kNN`, the `page-blocks` and `vowel` datasets respectively achieve insignificant and relatively large negative `FDC` values. This suggests that the former has a weak global structure, while the latter has a strong one. Interestingly, the four datasets with $n = 16$ features (i.e. `zoo`, `vote`, `pendigits`, and `letter`) yield quite different FDC values. The `FDC` for `pendigits` and `letter` is close to -1 whereas that of `zoo` and `vote` is close to 0. This suggests that the landscapes induced by the first two datasets are easier than those from the last two. The main difference between them is the number of classes and observations. Both these factors appear to significantly influence the global structure of the landscape.

The type of ML model appears to affect the global structure of the feature selection problem as well. For instance, for the `breast-w` dataset, the `FDC` is slightly positive for `DT` whereas it is moderately negative for the other models. However, we do not observe any clear trend across datasets when examining a specific ML model.

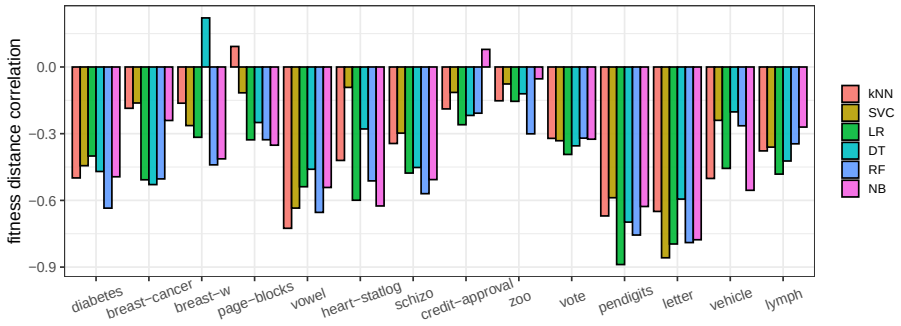


Fig. 4: Fitness-distance correlation.

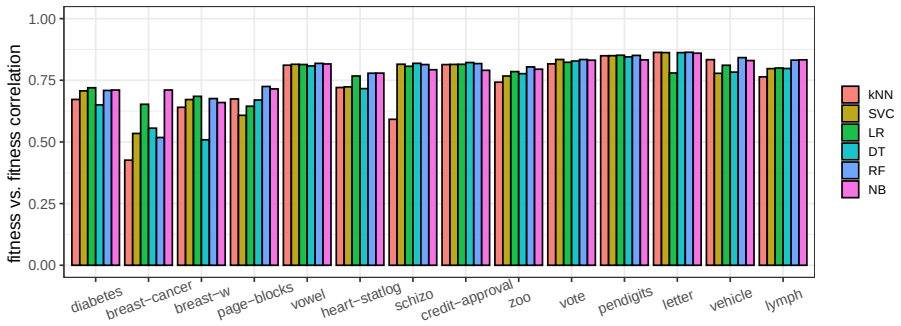


Fig. 5: Correlation between the fitness values of neighbor solutions (ruggedness).

4.5 Ruggedness

Let us now analyze the *ruggedness* [23] of the landscape depending on the considered dataset and ML model. We here measure the ruggedness as the Spearman correlation among the fitness values of neighboring solutions. For each pair of neighbors $x, x' \in \mathcal{X}$ such that $x' \in \mathcal{N}(x)$, we measure the correlation between $f(x)$ and $f(x')$. A larger correlation indicates a smoother landscape. The results are given in Fig. 5.

Overall, the correlation is high, suggesting that landscapes are all relatively smooth. Exceptions where the correlation drops below 0.75 exist for the `diabetes`, `breast-cancer`, `breast-w`, and `page-blocks` datasets. However, this could be an artifact of fewer neighbor pairs in these cases, as all these datasets have $n \leqslant 10$ features. The ruggedness also appears to be consistent across ML models for a specific dataset. As such, we find that the landscape difficulty induced by different ML algorithms is not due to the ruggedness but to other factors that we examine further below.

4.6 Neutrality

Fig. 6 gives the level of neutrality of the 84 landscapes under consideration, measured as the average proportion of equivalent solutions in the neighborhood

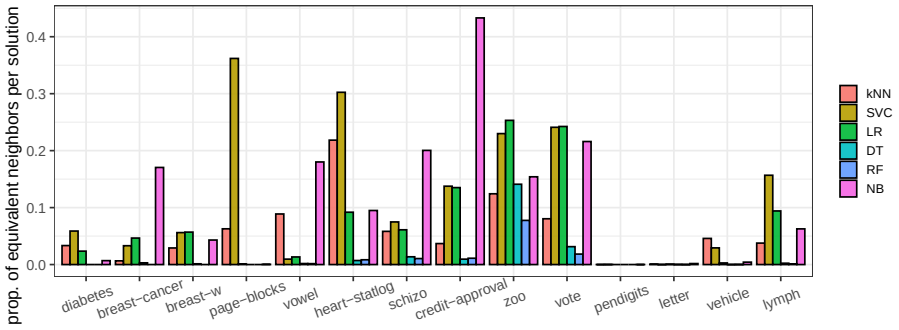


Fig. 6: Rate of neutrality between neighboring solutions.

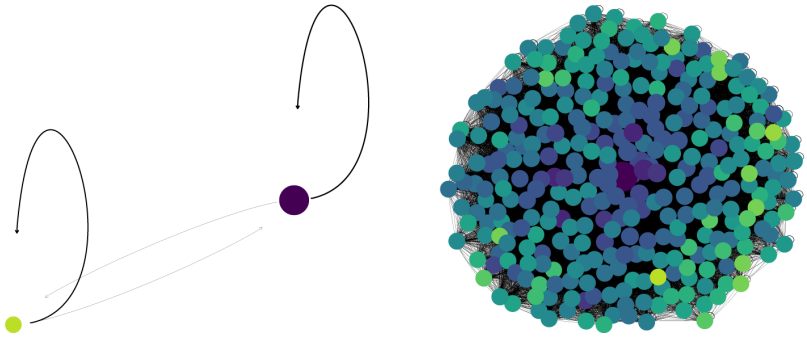
of each solution. A landscape is *neutral* when many neighbors share the same fitness value. It can be pictured as having multiple plateaus. We can see that the neutrality of the feature selection problem is highly dependent on the dataset and ML model. For instance, the `zoo`, `heart-statlog`, `credit-approval`, and `vote` datasets exhibit a high neutrality. By contrast, the `pendigits` and `letter` datasets have almost no neutral neighbors. Note that Mostert et al. [12] propose reducing neutrality by removing irrelevant features. However, they did *not* consider the datasets that produce less neutrality in their analysis.

By relating our results with existing algorithms, it is worth noting that some approaches for handling equivalent solutions have been proposed in the context of multi-objective feature selection [7]. Despite most prior studies using `kNN`, it is important to remark that our results underscore significant variations in neutrality across ML models. Notably, the level of neutrality is actually larger for `SVC` and `NB`, whereas it is quite low for `DT` and `RF`. The lower neutrality of the latter two ML models may be attributed to their stochastic nature, although we did perform multiple training for those. This suggests that existing wrapper algorithms might be worth revisiting in light of the ML model being used.

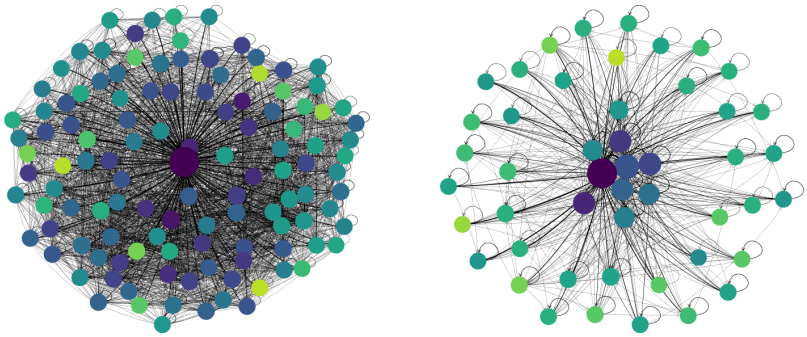
4.7 Local Optima Networks

Escaping from local optima is one of the main challenge for search algorithms. Therefore, understanding the number and distribution of local optima, along with the size of their basins of attraction, is crucial for comprehending the difficulty of the landscape. LONs have been widely used for this purpose. We built LONs and CM-LONs for all landscapes. A full analysis of these networks is out of reach in this paper due to space limitations. However, we provide some examples in Fig. 7. The size of each node corresponds to the size of its basin. The node color represents the classification accuracy, with darker shades representing better solutions. The edge width corresponds to the transition probability. The datasets considered in these examples all have $n = 16$ features.

The plots show that the structure of the landscape can be visually distinguished across datasets and ML models. For instance, the LON appears much denser for `LR` (bottom left) than for `kNN` (top left) when using the `letter` dataset.



(a) LON (`kNN`, `letter`) (b) LON (`kNN`, `vote`)



(c) LON (`LR`, `letter`) (d) CM-LON (`kNN`, `vote`)

Fig. 7: Examples of obtained LONs.

Similarly, the LON is denser for the `vote` dataset (top right) than for the `letter` dataset (top left) under `kNN`. In fact, the LON of `kNN` for `letter` (top left) aligns with the expectation of a relatively smooth landscape. The LON (top right) and CM-LON (bottom right) of `kNN` for the `vote` dataset can also be compared. We here observe a significant reduction of nodes and edges. By contrast, the CM-LON of `LR` for `letter` (not reported) has only one node less than its LON. We observed that the compression rate from LON to CM-LON, though varying among datasets, was fairly consistent across ML models. Nonetheless, `kNN` often attained a higher compression rate compared to other ML models. We delve deeper into the analysis of local optima using statistics over the LONs below.

4.8 Local Optima

In Fig. 8 (top), we report the number of local optima for all datasets and ML models. Similar to what we did with global optima, we treat plateaus of local optima as a single count. However, considering the variability of plateaus across different landscapes, Fig. 8 (bottom) quantifies the number of solutions within each local optima plateau, accounting for neutrality at the local optima level.

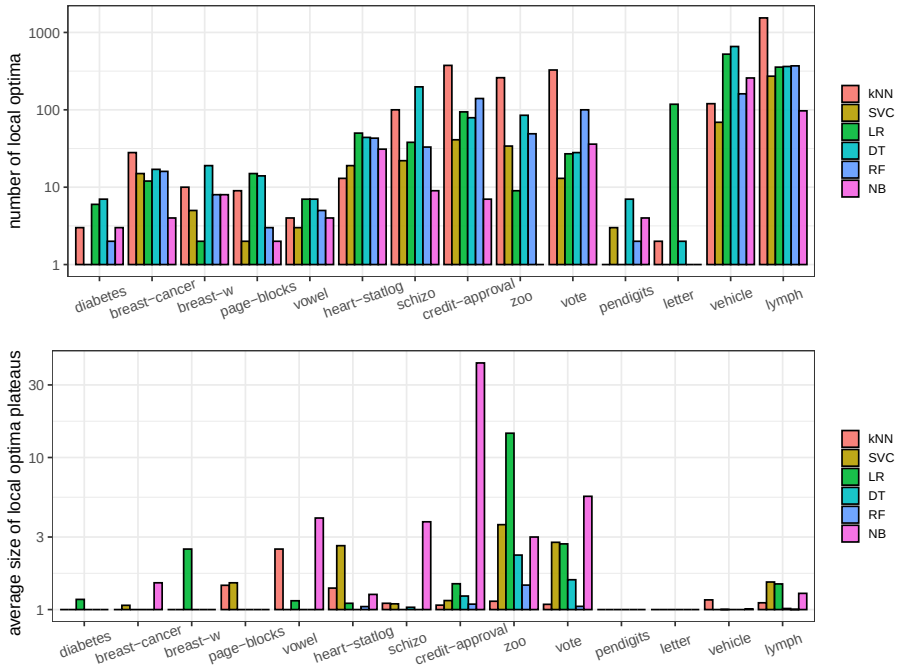


Fig. 8: Number of local optima (top) and average number of solutions in local optima plateaus (bottom) — notice the log scale.

Here as well, we observe significant differences across datasets and ML models. `kNN` and `LR` tend to produce more local optima. In fact, they are the models with the highest number of local optima in about one-third of the datasets. By contrast, `NB` often produces the fewest local optima, in about half of the datasets. This could explain the lower correlation we observed between `NB` and other ML models on the corresponding datasets, as discussed in Section 4.2. However, as anticipated in Section 4.6, `NB` also has the largest plateaus of local optima. This implies that local optima are typically clustered, making the `NB` algorithm potentially more resistant to noise caused by irrelevant features. Surprisingly, we found a single local optimum (or plateau) for `SVC` on the `diabetes` dataset, for `NB` on the `zoo` dataset, for `kNN` and `LR` on the `pendigits` dataset, and finally for `SVC`, `RF` and `NB` on the `letter` dataset. This means that local optima are all global optima, indicating that the corresponding landscapes are *uni-modal*.

Among the different datasets, `schizo`, `credit-approval`, `zoo`, `vote`, `vehicle`, and `lymph` often produce more local optima. By contrast, `pendigits` and `letter` typically yield significantly fewer local optima, and without any plateau. This might be surprising, given that these last two datasets are not the ones with the fewest features. Yet, we anticipated these landscapes were easier when examining the `FDC` in Section 4.4. It is noteworthy that `pendigits` and `letter` are the datasets with the largest number of observations — more than 10 000, signifi-

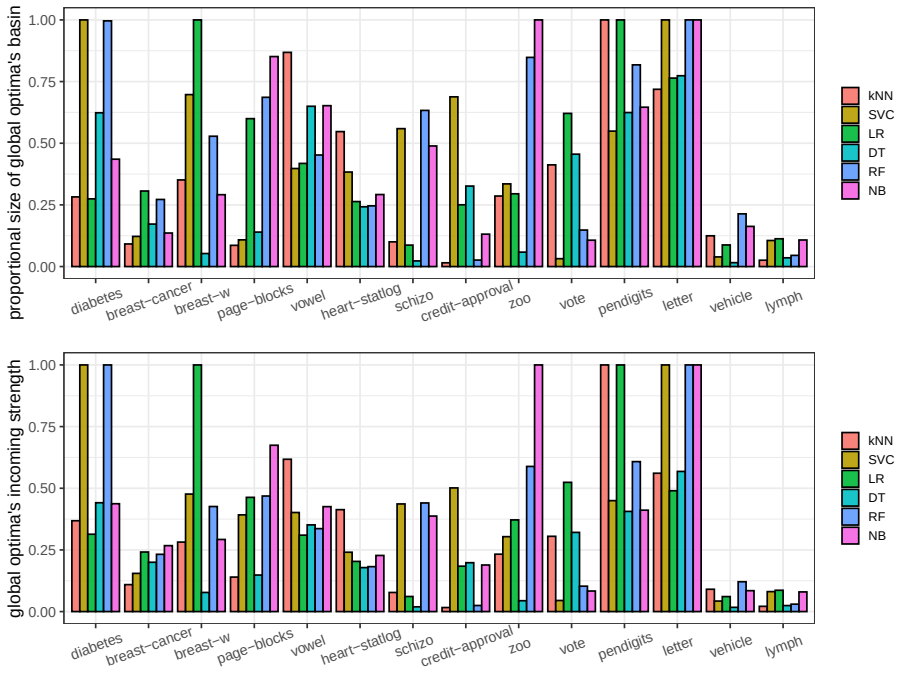


Fig. 9: Proportional sizes of global optima's basins of attraction (top), and incoming strength of escape edges towards global optima (bottom).

cantly more than other datasets. Training ML models with such large datasets is time-consuming, particularly for `kNN` which computes the pairwise distance between observations. We believe this might be the reason why large datasets are seldom used for benchmarking evolutionary feature selectors. Nevertheless, our findings underscore the importance of considering those. The question of whether having more observations consistently leads to fewer local optima requires further investigations that we leave open for future research. This could have significant implications for the explainability of feature selection.

Similar to global optima, the proportion of features selected on local optima stays fairly consistent across ML models, usually around 0.5 for all landscapes — details are omitted due to space constraint. However, the `pendigits` and `letters` datasets show a significant difference, as local optimal subsets often include more, if not all, features.

4.9 Basins of Attraction

Fig. 9 presents statistics on local optima's basins of attraction. The proportional size of global optima's basin (on top) gives the probability for a simple local search to fall into a global optimum when starting from a random solution. The incoming strength of escape edges towards global optima (at the bottom) represents the probability of transitioning from a local to a global optimum. A

Table 2: Success rate of `SFS-F`, `SFS-B` and `GA` on the considered problems. Best values are highlighted in gray.

dataset	SFS-F						SFS-B						GA					
	kNN	SVC	LR	DT	RF	NB	kNN	SVC	LR	DT	RF	NB	kNN	SVC	LR	DT	RF	NB
diabetes	.00	.00	.00	.00	1.0	1.0	.00	1.0	.00	1.0	1.0	1.0	.55	.32	.42	.55	.68	.55
breast-cancer	.00	.00	.45	1.0	1.0	.00	.00	.00	.00	.00	.00	.00	.23	.29	.42	.39	.42	.06
breast-w	.00	.00	.00	.00	.00	.00	.00	1.0	1.0	.00	1.0	.00	.35	.35	.77	.00	.39	.55
page-blocks	1.0	1.0	.00	.00	1.0	.00	.00	.00	1.0	.00	.00	1.0	.03	.13	.58	.32	.45	.45
vowel	.00	.00	.00	1.0	.00	1.0	1.0	.00	.00	1.0	1.0	1.0	.81	.48	.55	.58	.55	.81
heart-statlog	.00	.00	.00	1.0	.00	.55	.00	.00	.00	.00	.00	.00	.45	.35	.35	.16	.52	.45
schizo	.00	1.0	.00	.00	1.0	.00	.00	.00	.00	.00	.52	.00	.19	.16	.39	.16	.68	.55
credit-approval	.00	.00	.00	1.0	.00	.00	.00	.00	.00	.00	.00	.00	.00	.01	.26	.45	.00	.00
zoo	.00	.00	.06	.01	.00	.00	.00	.00	.00	.00	.00	.00	.32	.03	.26	.16	.74	.32
vote	.00	.00	.00	.00	.00	.00	.00	.00	.19	1.0	.00	.00	.35	.00	.48	.52	.16	.26
pendigits	.00	.00	1.0	1.0	1.0	.00	1.0	1.0	1.0	1.0	1.0	1.0	.81	.35	.87	.87	.97	.61
letter	1.0	1.0	.00	1.0	1.0	1.0	.00	1.0	1.0	1.0	1.0	1.0	.48	.87	.81	.74	.58	.68
vehicle	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	1.0	.06	.03	.19	.06	.06	.26
lymph	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.29	.29	.35	.06	.19	.29

higher value means that a local search is more likely to fall into a global optimum after perturbation.

We first remark that for 9 of the landscapes, both measures equal 1. This is expected as they correspond to those with a single optimum, as discussed in Section 4.8. In other landscapes, the basins of global optima remain relatively large, and complementary investigations reveal that the largest basin often ranks high in fitness. In addition, datasets with more features ($n = 18$) tend to have proportionally smaller global optima's basins. However, it is worth noting that these cases simply have more local optima, while the rank of the largest basin remains low. The global optima's incoming strength varies with the dataset and ML model. For example, a high value can be observed for the `diabetes`, `pendigits`, and `letter` datasets under any ML model. By contrast, the value is low for the `vehicle` and `lymph` datasets. These findings shall be related to the number of local optima in these datasets, as discussed in Section 4.8.

4.10 Algorithm Performance

We conclude our analysis by relating our observations on the landscape of feature selection problems with the performance of established wrapper algorithms, as described in Section 3. Table 2 gives the performance of the three considered algorithms for each dataset and ML model. We measure performance as the proportion of runs (among 31) in which the algorithm under consideration was able to identify a global optimum. Our primary focus is to compare algorithm performance across different ML models and datasets for each feature selector.

Firstly, we note that there is a substantial number of problems where `SFS-F` and `SFS-B` agree with each other. With a few exceptions, `SFS-F` and `SFS-B` typically have a success rate of either 0% or 100% due to their deterministic nature. The only element of randomness lies in how ties are broken. This explains why the success rate falls in between in some landscapes with neutrality; see

Section 4.6. In fact, the optimal subset is always identified by SFS-F and SFS-B across different datasets and ML models when the landscape has (1) a single global optima's basin of attraction, and (2) no plateaus. Overall, DT and RF tend to be easier to solve than other ML models. By contrast, SFS-F and SFS-B seldom identifies an optimal solution for kNN or LR, which were shown to produce more local optima in Section 4.8. As highlighted above, both DT and RF have landscapes with relatively strong FDC, almost no neutrality, and few plateaus. In addition, the largest basin of attraction often leads to an optimal solution for RF.

Now shifting our attention to GA, we notice a higher variation in the success rates. For a given dataset, the ML model resulting in the highest success rate is almost always identical to either SFS-S or SFS-B. Similar to those, either DT or RF is among the best for 9 out of the 14 datasets. Interestingly, a strong correlation seems to exist between the best GA setting and its FDC value. Indeed, in half of the datasets, the ML model with the lowest degree of difficulty in terms of FDC has the highest success rate. The corresponding ML model consistently shows a relatively lower neutrality rate as well. Furthermore, for almost all datasets, the largest basin of attraction for the ML model with the highest success rate falls into a global optimum. Finally, as anticipated above, for problems with $n = 16$ features, the landscapes induced by pendigits and letter appear to be easier to solve than those of zoo and vote. This holds for all three algorithms.

5 Conclusions

In this paper, we conducted a landscape analysis to study various aspects of difficulty in wrapper methods for feature selection. We examined 14 classification datasets and 6 ML models, resulting in 84 different landscapes. Our findings suggest that the difficulty and solutions are inherent to the landscape being considered. Specifically, we observed significant differences across ML models. This highlights the need to explore ML models beyond kNN, which is commonly used in the existing literature. Given the observed variations among landscapes, we do not recommend to use one ML model as a proxy for another ML model.

In addition to considering additional datasets, ML models, and feature selection algorithms, we aim to assess the impact of other classification scores. We also plan to address the challenges raised by problems with a larger number of features, thus going beyond complete enumeration. Our experimental analysis indicated that studying how the landscape difficulty varies with the number of classes and observations, particularly in terms of neutrality and multimodality, requires further consideration. In addition, the established sequential feature selection approach essentially performs a local search starting from a specific solution. We believe our methodology could help formalize its probability of successfully identifying a global optimum. At last, we expect to enhance the explainability of feature selection by analyzing and interpreting the features that wrapper methods most frequently choose.

Acknowledgments. This research was partially supported by the French Embassy in Japan under the Japan Exploration Program.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bache, K., Lichman, M.: UCI machine learning repository (2013), http://archive.ics.uci.edu/ml
2. Boese, K.D., Kahng, A.B., Muddu, S.: A new adaptive multi-start technique for combinatorial global optimizations. Oper. Res. Lett. **16**(2), 101–113 (1994). https://doi.org/10.1016/0167-6377(94)90065-5
3. Dash, M., Liu, H.: Feature selection for classification. Intell. Data Anal. **1**(1-4), 131–156 (1997). https://doi.org/10.1016/S1088-467X(97)00008-5
4. Dökeroglu, T., Deniz, A., Kiziloz, H.E.: A comprehensive survey on recent metaheuristics for feature selection. Neurocomputing **494**, 269–296 (2022). https://doi.org/10.1016/j.neucom.2022.04.083
5. Doye, J.P.K.: The network topology of a potential energy landscape: a static scale-free network. Phys. Rev. Lett. **88**, 238701 (2002). https://doi.org/10.1103/PhysRevLett.88.238701
6. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. J. Mach. Learn. Res. **3**, 1157–1182 (2003), https://dl.acm.org/doi/10.5555/944919.944968
7. Jiao, R., Nguyen, B.H., Xue, B., Zhang, M.: A survey on evolutionary multiobjective feature selection in classification: Approaches, applications, and challenges. IEEE Trans. Evol. Comput. (2024, in press). https://doi.org/10.1109/tevc.2023.3292527
8. Jones, T., Forrest, S.: Fitness distance correlation as a measure of problem difficulty for genetic algorithms. In: Eshelman, L.J. (ed.) Proceedings of the 6th International Conference on Genetic Algorithms, Pittsburgh, PA, USA, July 15-19, 1995. pp. 184–192. Morgan Kaufmann (1995)
9. Kerschke, P., Trautmann, H.: Automated algorithm selection on continuous black-box problems by combining exploratory landscape analysis and machine learning. Evol. Comput. **27**(1), 99–127 (2019). https://doi.org/10.1162/evco“a“00236
10. Liefooghe, A., Daolio, F., Verel, S., Derbel, B., Aguirre, H.E., Tanaka, K.: Landscape-aware performance prediction for evolutionary multiobjective optimization. IEEE Trans. Evol. Comput. **24**(6), 1063–1077 (2020). https://doi.org/10.1109/tevc.2019.2940828
11. Mostert, W., Malan, K.M., Engelbrecht, A.P.: Filter versus wrapper feature selection based on problem landscape features. In: Aguirre, H.E., Takadama, K. (eds.) Proceedings of the Genetic and Evolutionary Computation Conference Companion, GECCO 2018, Kyoto, Japan, July 15-19, 2018. pp. 1489–1496. ACM (2018). https://doi.org/10.1145/3205651.3208305
12. Mostert, W., Malan, K.M., Ochoa, G., Engelbrecht, A.P.: Insights into the feature selection problem using local optima networks. In: Liefooghe, A., Paquete, L. (eds.) Evolutionary Computation in Combinatorial Optimization - 19th European Conference, EvoCOP 2019, Held as Part of EvoStar 2019, Leipzig, Germany, April 24-26, 2019, Proceedings. Lecture Notes in Computer Science, vol. 11452, pp. 147–162. Springer (2019). https://doi.org/10.1007/978-3-030-16711-0“10
13. Ochoa, G., Veerapen, N., Daolio, F., Tomassini, M.: Understanding phase transitions with local optima networks: Number partitioning as a case study. In: Evolutionary Computation in Combinatorial Optimization, EvoCOP 2017. Lecture Notes in Computer Science, vol. 10197, pp. 233–248 (2017)

14. Ochoa, G., Verel, S., Daolio, F., Tomassini, M.: Local Optima Networks: A New Model of Combinatorial Fitness Landscapes. In: Recent Advances in the Theory and Application of Fitness Landscapes. Emergence, Complexity and Computation, vol. 6. Springer (2014). https://doi.org/10.1007/978-3-642-41888-4“˙9
15. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., VanderPlas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011). https://doi.org/10.5555/1953048.2078195
16. Pimenta, C.G., de Sá, A.G.C., Ochoa, G., Pappa, G.L.: Fitness landscape analysis of automated machine learning search spaces. In: Paquete, L., Zarges, C. (eds.) Evolutionary Computation in Combinatorial Optimization - 20th European Conference, EvoCOP 2020, Held as Part of EvoStar 2020, Seville, Spain, April 15-17, 2020, Proceedings. Lecture Notes in Computer Science, vol. 12102, pp. 114–130. Springer (2020). https://doi.org/10.1007/978-3-030-43680-3“˙8
17. Pushak, Y., Hoos, H.H.: Automl loss landscapes. ACM Trans. Evol. Learn. Optim. **2**(3), 10:1–10:30 (2022). https://doi.org/10.1145/3558774
18. Richter, H., Engelbrecht, A.: Recent Advances in the Theory and Application of Fitness Landscapes. Emergence, Complexity and Computation, Springer (2014)
19. Schneider, L., Schäpermeier, L., Prager, R.P., Bischl, B., Trautmann, H., Kerschke, P.: HPO $\times$ ELA: investigating hyperparameter optimization landscapes by means of exploratory landscape analysis. In: Rudolph, G., Kononova, A.V., Aguirre, H.E., Kerschke, P., Ochoa, G., Tusar, T. (eds.) Parallel Problem Solving from Nature - PPSN XVII - 17th International Conference, PPSN 2022, Dortmund, Germany, September 10-14, 2022, Proceedings, Part I. Lecture Notes in Computer Science, vol. 13398, pp. 575–589. Springer (2022). https://doi.org/10.1007/978-3-031-14714-2“˙40
20. Thomson, S.L., Ochoa, G., Veerapen, N., Michalak, K.: Channel configuration for neural architecture: Insights from the search space. In: Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2023. pp. 1267–1275. ACM, Lisbon, Portugal (2023). https://doi.org/10.1145/3583131.3590386
21. Verel, S., Daolio, F., Ochoa, G., Tomassini, M.: Local optima networks with escape edges. In: Hao, J., Legrand, P., Collet, P., Monmarché, N., Lutton, E., Schoenauer, M. (eds.) Artificial Evolution - 10th International Conference, Evolution Artificielle, EA 2011, Angers, France, October 24-26, 2011, Revised Selected Papers. Lecture Notes in Computer Science, vol. 7401, pp. 49–60. Springer (2011). https://doi.org/10.1007/978-3-642-35533-2“˙5
22. Verel, S., Ochoa, G., Tomassini, M.: Local optima networks of NK landscapes with neutrality. IEEE Trans. Evol. Comput. **15**(6), 783–797 (2011). https://doi.org/10.1109/tevc.2010.2046175
23. Weinberger, E.D.: Correlated and uncorrelatated fitness landscapes and how to tell the difference. Biological Cybernetics **63**(5), 325–336 (1990)
24. Xue, B., Zhang, M., Browne, W.N., Yao, X.: A survey on evolutionary computation approaches to feature selection. IEEE Trans. Evol. Comput. **20**(4), 606–626 (2016). https://doi.org/10.1109/tevc.2015.2504420