This is a post-peer-review, pre-copyedit version of an article published in Liefooghe A & Paquete L (eds.) *Evolutionary Computation in Combinatorial Optimization*. Lecture Notes in Computer Science, 11452. 19th European Conference on Evolutionary Computation in Combinatorial Optimization, EvoCOP 2019, held as part of EvoStar 2019, Leipzig, Germany, 24.04.2019-26.04.2019. Cham, Switzerland: Springer Verlag, pp. 147-162. The final authenticated version is available online at: https://doi.org/10.1007/978-3-030-16711-0_10

Insights into the Feature Selection Problem using Local Optima Networks

Werner Mostert[1], Katherine M. Malan[2], Gabriela Ochoa[3], and Andries P. Engelbrecht[1]

[1] Stellenbosch University, Stellenbosch Central, Stellenbosch, South Africa
werner.mostert1@gmail.com
[2] University of South Africa, Muckleneuk, Pretoria, 0002, South Africa
malankm@unisa.ac.za
[3] University of, Stirling FK9 4LA, United Kingdom
gabriela.ochoa@cs.stir.ac.uk

Abstract. The binary feature selection problem is investigated in this paper. Feature selection fitness landscape analysis is done, which allows for a better understanding of the behaviour of feature selection algorithms. Local optima networks are employed as a tool to visualise and characterise the fitness landscapes of the feature selection problem in the context of classification. An analysis of the fitness landscape global structure is provided, based on seven real-world datasets with up to 17 features. Formation of neutral global optima plateaus are shown to indicate the existence of irrelevant features in the datasets. Removal of irrelevant features resulted in a reduction of neutrality and the ratio of local optima to the size of the search space, resulting in improved performance of genetic algorithm search in finding the global optimum.

Keywords: Local Optima Networks · Feature Selection · Fitness Landscape Analysis.

1 Introduction

To further the development of a generalised theoretical framework for feature selection, and to better understand the feature selection problem, this paper analyses fitness landscapes of the feature selection problem.

The binary feature selection problem has the goal of finding the subset of all features that are the most relevant for a classification task. A full enumeration of candidate solutions (i.e subsets of features) is performed to construct a complete fitness landscape for a number of real-world classification problems. The size of the solution space grows exponentially as the number of features increases, making a full enumeration computationally infeasible for a large number of features. Therefore, only datasets with a small number of features (less than 18) are considered in this paper.

A number of different approaches have been proposed for solving the feature selection problem [4, 12–14, 23, 24]. Chandrashekar et al. [4] showed that the performance of different feature selection techniques are often problem dependent

and that the methods show vast disparity in their success ratios. This paper uses two wrapper methods and one filter method to analyse the behaviour of feature selection algorithms with respect to problem landscape characteristics. Fitness landscapes can be a valuable tool to understand problems and to analyse search algorithm behaviour [20], since fitness landscapes possess structural attributes that influence the performance of search algorithms [16].

Local optima networks [19] have previously been used to characterise the global structure of fitness landscapes for benchmark combinatorial problems. The quadratic assignment problem [6], NK-landscapes [22] and the number partitioning problem [18] are examples of combinatorial problem case studies using local optima networks as an analytical tool. For small instances of the feature selection problem (in this study with less than 18 features), a full enumeration of all candidate solutions is computationally practical. This allows for the construction of complete local optima networks and a full analysis of the global structure of fitness landscapes for instances of classification problems.

New insights to the nature of the feature selection problem are obtained using local optima networks. The local optima networks for the real world classification problems reveal interesting fitness landscape characteristics of the feature selection problem, before and after feature removal. Removal of irrelevant features shows a reduction in neutrality in the fitness landscape, a reduction in the ratio of local optima to the size of the search space, and a reduction in problem difficulty for genetic algorithm search.

The following section gives an overview of the feature selection problem and local optima networks, while Section 3 describes the experimental process and algorithm details. Finally, the results obtained are discussed in Section 4 and the paper is concluded in Section 5.

2 Background and Related Work

This section discusses the general feature selection problem and gives an overview of local optima networks.

2.1 The Feature Selection Problem

The feature selection problem is concerned with finding a set of the most relevant features from the set of all available features for a classification task. A solution to the feature selection problem can be represented as a binary string of length n, where n is the number of features in the dataset and each bit indicates whether the feature is selected or not.

Feature selection is applied as a pre-processing technique in order to reduce the dimensionality of a problem by removing redundant and irrelevant features. The issue of feature irrelevance can be misleading since two mutually exclusive features could be useless, but the union of these features could be information rich with respect to the dependant variable [10]. The utilisation of a subset of relevant features as opposed to the set of all features has been shown to increase

classifier performance, reduce computational complexity, and lead to a better understanding of the data for machine learning [4].

Feature selection algorithms can be categorised into three distinct categories, namely filter, wrapper and embedded methods. Filter methods [4] establish how important features are based on information with respect to the dependent variable, using measures such as correlation or mutual information. Wrapper methods [4] use subsets of features, for which a measure of the model accuracy is obtained per subset of features. Heuristic search is used by wrapper methods to determine the set of most relevant features with the model accuracy as objective function. Embedded methods [4] search for the most relevant features by taking advantage of the built in learning process, such as with decision trees [10].

Choosing the most suitable feature selection algorithm for a given dataset is an unsolved problem, since there is a lack of an underlying theoretical framework and a limited understanding of the nature of the feature selection problem [10]. Fitness landscapes can be used to better understand the nature of the search spaces of optimisation problems [16] and in this paper, local optima networks are employed to characterise and visualise the fitness landscapes for the feature selection problem.

2.2 Local Optima Networks

A fitness landscape is formulated as a triplet (S, N, f) [21], where S is the set of all candidate solutions, N is an operator that defines a neighbourhood structure in the solution space, and $f : S \longrightarrow \mathbb{R}$ is a fitness (objective) function that assigns to all $s \in S$ a solution quality. Binary solution spaces are have a size of $|S| = 2^N$, where N represents the dimensionality of the problem.

Local optima networks [19] are inspired by the work of Doye [7] in modelling physical energy landscapes of atomic clusters as complex networks, serving as a powerful tool to analyse the global structure of a fitness landscape. For combinatorial spaces, the idea is adapted from connected energy minima to weighted graphs of local optima in the fitness landscape [19]. Definitions of the components of the local optima network model are given below, assuming a maximisation problem.

Local optimum. A local optimum $s^* \in S^*$, where $S^* \subset S$ is the set of locally optimum solutions, is defined as a solution where $\forall s \in N(s^*), f(s) < f(s^*)$.

Local optima network. A local optima network, $G = (S^*, E)$, represents the weighted graph of all local optima solutions S^*, where an escape edge $e_{ij} \in E$ exists between nodes S_i^* and S_j^*.

Escape edge. An escape edge [22] represents the probability of moving into a neighbouring basin of attraction from a local optimum, after a defined controlled perturbation followed by a hill-climbing local search to find the connecting local optima.

Basin of attraction. A basin of attraction of a local optimum s_i^* is the set $b_i = \{s \in S \mid LocalSearch(s) = s_i^*\}$. The number of solutions, $|b_i|$, in the basin of attraction represents the size of the basin.

Monotonic local optima network (M-LON). The monotonic local optima network, *M-LON* $= (S^*, E)$, is the graph where the nodes $s_i^* \in S^*$ are the local optimum plateaus, and there is an edge $e_{ij} \in E$, with weight w_{ij}, between two nodes s_i^* and s_j^* if $w_{ij} > 0$.

M-LON plateau. The set of connected nodes in the *M-LON* with equal fitness.

Compressed LON nodes. The set of *M-LON* plateaus, C^*.

Compressed monotonic local optima network (CM-LON). The local optima network, *CM-LON*$= (C^*, E)$, is the graph where the nodes $c_i \in C^*$ are *M-LON* plateaus. Weighted edges in the *CM-LON* are aggregated for the edges of nodes in the *M-LON* plateau.

3 Experimental Setting

This section describes the classification datasets used in this study. The construction of the fitness landscape using the full enumeration of feature subsets is then described, followed by the different feature selection algorithms and the approach used to construct and visualise the local optima networks.

3.1 Datasets

The University of California, Irvine (UCI), Machine Learning Repository [15] contains a wide range of datasets that can be used for various machine learning objectives. Seven of the UCI repository classification datasets, containing a variety of nominal and numerical features, were used in this study and are summarised in Table 1.

Table 1. Datasets

Dataset	*Nominal*	*Numerical*	*# Classes*	*# Data Elements*	*# Features*
breast-cancer	Yes	No	2	286	9
zoo	Yes	Yes	7	101	17
page-blocks	No	Yes	5	5473	10
vowel	Yes	Yes	11	990	13
breast-w	No	Yes	2	699	9
heart-statlog	No	Yes	2	270	13
diabetes	No	Yes	2	768	8

The number of features for the datasets considered were kept small in order to be able to compute a full enumeration of the search space, taking into account that the search space expands exponentially with every feature that is considered.

3.2 Fitness Function

A measure of classification accuracy based on a test dataset is used as the fitness value for a solution, s, where s is a subset of features. A bit string representation

is used for solutions to indicate inclusion or exclusion of a specific feature. The test dataset is obtained by doing a 50/50 split of the original dataset, where the training dataset is used to perform the feature selection.

The classic k-nearest-neighbour [1] using Euclidean distance, a simple non-stochastic classifier, is used as implemented in the Weka Machine Learning software development kit [11]. Stochasticity involved with calculating fitness is avoided since it would introduce noise into the fitness landscape. The best value for k is problem dependant. For the purpose of this paper, it is not necessary to use optimal values for k since the focus of this paper is to understand the effect of the inclusion and exclusion of features on the fitness landscape. Therefore, k is arbitrarily chosen as $k = 3$.

The measure of classification accuracy is defined as Cohen's Kappa-statistic, a measure non-biased by class imbalance. Cohen's Kappa is defined as [3]:

$$\kappa = \frac{P_0 - P_c}{1 - P_c} \tag{1}$$

where P_c is the probability of agreement by chance and P_0 is the total agreement (i.e. the number of correctly classified instances). The Kappa statistic allows for the level of agreement with respect to each class label to be measured. The Kappa statistic is a robust measure of accuracy, normalised to the range $[-1, 1]$, where total disagreement is represented as $\kappa = -1$, completely random classification as $\kappa = 0$, and total agreement as $\kappa = 1$. Using non-biased measures of classifier accuracy is important since a raw percentage of correct classification may be statistically biased due to a skewed distribution of classes.

A full enumeration of candidate solutions is done. For datasets with a large number of features, it becomes computationally infeasible to do a full enumeration (an NP-hard problem [2]). A full enumeration for small binary spaces allows the construction of a complete local optima network to analyse and visualise the feature selection problem.

3.3 Feature Selection Algorithms

This section describes the filter method and the two wrapper methods that are used to conduct feature selection in this study. The three feature selection methods below are chosen since each method addresses the problem in a very different way.

Filter Method. The information gain [5] feature evaluation measurement is selected for use by the ranker filter method, as implemented in Weka [11]. Using the training dataset, the features are ranked based on relevance with respect to the class, from high to low. Given a list of ranked features, each linear combination of features from highest relevance to lowest relevance is considered and the fitness of the solution is evaluated.

For example, for the set of all features sorted by decreasing information gain, $F = \{1, 5, 2, 4, 3\}$, the following five feature sets are considered in order: $\{1\}$, $\{1, 5\}$, $\{1, 5, 2\}$, $\{1, 5, 2, 4\}$, and $\{1, 5, 2, 4, 3\}$. The fitness of each subset of

features is computed by applying k-NN on the test dataset with the subset of features and using the Kappa statistic as an accuracy measure. The feature set with the highest fitness value is selected as the output of the filter method. The data instances used to determine information gain are restricted to the training set only. This is done to fairly evaluate the performance against the wrapper methods, which use the same set for training.

Sequential Forward Selection Wrapper Method. The sequential forward selection (SFS) algorithm is used as implemented by GreedyStepwise in Weka [11]. The SFS algorithm starts with an initial empty set of features and sequentially adds a feature that results in better fitness up until there are no features that can be added that will result in better fitness than the current solution.

Genetic Algorithm Wrapper Method. The genetic algorithm wrapper method uses the classic genetic algorithm as described by Goldberg [9]. The following parameters were set for the genetic algorithm:

- Population size : 20
- Number of generations : 20
- Crossover probability : 0.6
- Mutation probability : 0.033

The Weka [11] default parameters were used for all datasets. Since the genetic algorithm has a stochastic element, fitness is reported as the mean over 30 independent runs of the algorithm.

3.4 Local Optima Network Generation & Visualisation

The open source Java library by Fieldsend [8] was used to generate the local optima network (LON) graph. Since a full enumeration of the candidate solutions in the search space is done, an exhaustive LON is generated which indicates all of the local and global optima that exist for each problem. In the LON graph, therefore, each node is either a local or a global optimum.

An edge between two nodes in the LON indicates that the two basins containing the optima are regarded as neighbours in the search space. Two common concepts used to define neighbourhood between basins in a LON are basin transitions [17] and escape edges [22]. In this study, the escape edges definition of neighbourhood is used since the basin transition definition, for a full enumeration, produces vastly dense networks. Using escape edges, two local optima are defined as neighbours in the LON if it is possible for any candidate solution in the basin of one local optimum to reach a candidate solution in the basin of the other optimum after a controlled perturbation. In this study, the 'controlled perturbation' for defining neighbourhood between basins is set to 2 bit flips. A larger value for the number of bit flips results in a less connected LON, whereas a single bit flip results in a more densely connected LON.

The generated local optima network is visualised, with the size of the nodes proportional to the basin sizes for the local optima. The edge weights, represented as the width of the edges in the visualisation, between nodes indicates the probability of moving to the connected node.

Table 2. Local Optima Network Node & Edge Colours

Node		*Edge*	
Colour	*Node Description*	*Colour*	*Edge Description*
Red	Global optima	Green	Deteriorating fitness
Blue	Best local optima in a non-global funnel	Blue	Neutral fitness
Gray	All other local optima	Gray	Improving Fitness

The nodes and edges are coloured as described in Table 2.

Compressed monotonic local optima networks [18], as defined in Section 2.2, is an adaptation of the local optima networks for connected local optima with equal fitness. Compressed monotonic local optima networks are used for visualisation of neutrality for the feature selection fitness landscapes. The compressed local optima networks show the size of the local optima plateaus as a box, where the length of box is proportional to the size of the plateau.

4 Results

The feature selection algorithm performance for each respective dataset is summarised in Table 3. The performance ratio, P_{ratio}, is calculated as,

$$P_{ratio} = 1 - \frac{f(s_{global} - f(s_{best}))}{s_{global}}$$

where f is the fitness function using Cohen's Kapa statistic for k-KNN, s_{global} is the solution with the best fitness of all candidate solutions (obtained by a full enumeration) and s_{best} is the solution that is returned by the feature selection algorithm. A performance ratio of 1.0 means that the algorithm found a global optimum. For the genetic algorithm (GA) wrapper method, the mean of the fitness values for each run of the algorithm is used to calculate the performance ratio. The GA success ratio is presented as the number of times a single run of the genetic algorithm found a global optimum as a proportion of the number of independent runs (30) of the algorithm. The values in bold show the best performance per feature selection algorithm.

The sequential forwards selection wrapper (SFS) method found a global optimum for two of the datasets. The GA wrapper method found a global optimum for at least one of the runs for all datasets, evident since the success ratio is always greater than 0. The filter method comes close to finding the global optimum, as can be seen in the breast-w dataset with a performance ratio of 0.9920, but failed to find a global optimum for any of the datasets.

The GA wrapper method comes close to finding a global optimum in terms of fitness, but shows a largely disparate success ratio with 0.1333 for the vowel dataset and 0.8666 for the breast-w dataset.

Table 4 summarises metrics calculated for the local optima networks for each of the datasets. The metrics give an analytical indication to the structure of the

Table 3. Feature Selection Algorithm Performance

Dataset	$P_{ratio}(Filter)$	$P_{ratio}(SFS)$	$P_{ratio}(GA)$	*GA Success Ratio*
breast-cancer	0.8315	0.9827	0.9758	0.3666
zoo	0.9711	0.8281	0.9872	0.1333
page-blocks	0.9409	**1.0**	0.9896	0.0666
vowel	0.8228	**1.0**	0.9826	0.7
breast-w	**0.9920**	0.9845	**0.9992**	**0.8666**
heart-statlog	0.9443	0.9443	0.9626	0.2
diabetes	0.9284	0.8899	0.9682	0.2333

local optima networks. The values in bold represent the highest value for each metric for the different data sets that are analysed.

Table 4. Local Optima Network Metrics

Dataset	*N_optima*	*N_global*	*Bsize_global*	*Str_global*	*N_coptima*	*Prop_neutral*
breast-cancer	30	1	0.0605	0.2252	24	0.2
zoo	**4938**	2	0.0012	0.0094	**1342**	**0.7282**
page-blocks	11	1	0.0958	0.152	11	0
vowel	42	**4**	0.1472	**0.4988**	15	0.6428
breast-w	40	3	0.1192	0.2423	14	0.65
heart-statlog	135	1	0.0135	0.062	109	0.1925
diabetes	8	1	**0.1607**	0.354	8	0

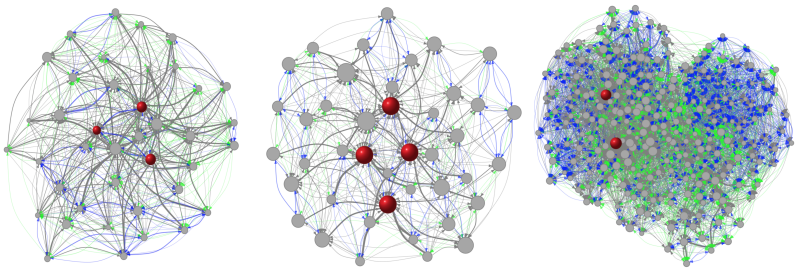
Table 4 contains the following metrics, with definitions as follows:

- *N_optima*: Number of nodes (total number of optima, including global and local)
- *N_global*: Number of different global optima
- *Bsize_global*: Normalised size of the global optimum. Notice that there can be more than one global optimum
- *Str_global*: Incoming strength of global optimum normalised by the total incoming strength
- *N_coptima*: Number of compressed local optima
- *Prop_neutral*: Proportion of local optima that have a neutral connection ($\frac{N\_optima - N\_coptima}{N\_optima}$)

Three of the datasets exhibited multiple global optima, where the remainder of the datasets exhibited a single global optimum. The dataset that contained the most global optima is the vowel dataset with four global optima. The three datasets that exhibited more than one global optimum also show the largest proportion of neutrality between nodes in the local optima network. The zoo local optima network shows a 0.7282 local optima neutrality proportion, which

indicates that 72.82% of all optima in the local optima network have at least one local optimum to which it is connected via a neutral edge (i.e. equal fitness between connected nodes).

The number of local optima are evidently correlated with the size of the search space, but does not necessarily tend to the same exponential scale since the ratio of global optima with regards to the number of candidate solutions differ.



(a) breast-w LON (b) vowel LON (c) zoo LON

Fig. 1. Local optima networks

The constructed local optima networks are shown in Figure 1. Only the datasets with multiple global optima are presented. Due to the size of the zoo local optima network, it has been pruned to 258 out of the 4938 with a fitness value threshold for visualisation purposes.

It is observed from the local optima network visualisations in Figure 2 that there are multiple global optima that are closely co-located. The nodes that are global optima form a fully connected sub-graph with neutral edges between them. Figure 2 shows a three dimensional visualisation of the local optima network that clearly illustrates the neutral plateau formed by the global optima for the breast-w dataset. The global optima neutral plateau can be seen in Figure 2 by the three red nodes (global optima) that are fully connected with blue (neutral) edges.

To investigate the presence of other (local) optima plateaus in the LON, a compressed monotonic LON (*CM-LON*) representation is used. This representation shows interconnected local optima via neutral edges as a box in the local optima network, where the length of the box is representative of the number of local optima that are connected. Nodes are still represented as circles where there are no local optima connected via a neutral edge. The *CM-LON* for each of the three datasets with multiple global optima are presented in Figure 3. The colouring scheme remains the same as for the full local optima networks.

Note that the individual nodes of the global optima are now collapsed into a single node, confirming the observation that the global optima are connected via neutral edges. The breast-w dataset exhibits a proportionally large local optima plateau and several smaller local optima plateaus, where the zoo dataset

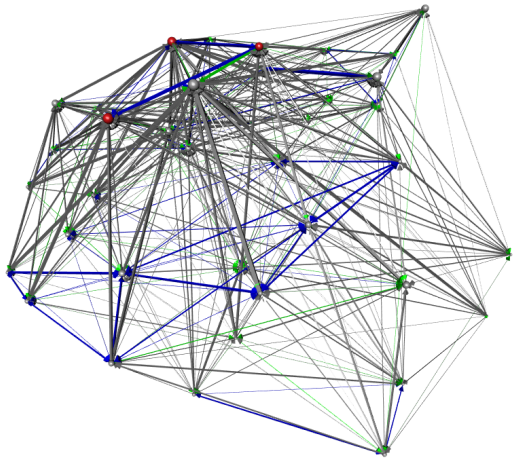
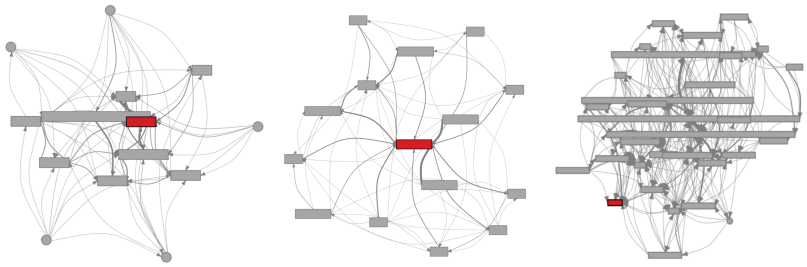


Fig. 2. 3D Local Optima Network for breast-w dataset.



(a) breast-w *CM-LON* (b) vowel *CM-LON* (c) zoo *CM-LON*

Fig. 3. Compressed Monotonic Local Optima Networks.

exhibits multiple proportionally larger local optima plateaus. Neutrality in a landscape represents a lack of information for search and metaheuristic search algorithms, such as genetic algorithms. Genetic algorithms require strategies for coping with high levels of neutrality. The vowel *CM-LON* shows that there is no local optimum in the network that does not have at least one other connected local optimum with equal fitness.

The formation of global optima plateaus shows that there are features in the global optima solutions that have no effect on fitness. Therefore, the symmetric intersection of the feature subsets will result in the set of irrelevant features. Table 5 shows the bit string representation of 10 local optima with the highest fitness represented as s_i^* where $i = 1, ..., 10$ is decreasing in fitness. The bits in bold represent the irrelevant features.

The irrelevant features for the three datasets with multiple global optima are removed and the local optima networks reconstructed. The expectation is that

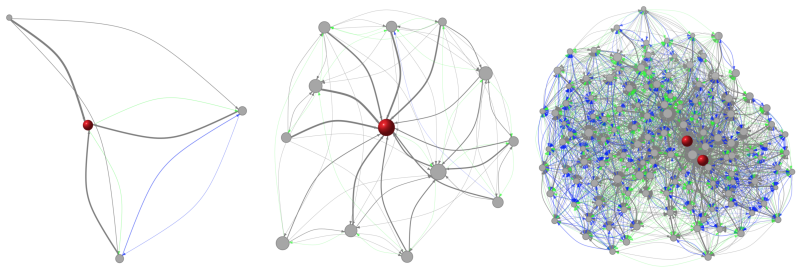
zoo	
Solution	*Features*
s_1^*	00111101100011001
s_2^*	00111101100011011
s_3^*	01111101010011001
s_4^*	01111101010011011
s_5^*	00011101110011011
s_6^*	00011101110111011
s_7^*	00011101100011101
s_8^*	01011101100011101
s_9^*	00011101100011111
s_{10}^*	01011101100011111

breast-w	
Solution	*Features*
s_1^*	110**0**01**1**1**0**
s_2^*	110**1**01**0**1**1**
s_3^*	110101111
s_4^*	010001111
s_5^*	110011010
s_6^*	101001100
s_7^*	101001101
s_8^*	111101000
s_9^*	010001010
s_{10}^*	111001010

vowel	
Solution	*Features*
s_1^*	**00**11111100100
s_2^*	**01**11111100100
s_3^*	1011111100100
s_4^*	1111111100100
s_5^*	0001111110100
s_6^*	1001111110100
s_7^*	0001101110110
s_8^*	0101101110110
s_9^*	1001101110110
s_{10}^*	1101101110110

Table 5. Feature subsets for top 10 highest fitness local optima.

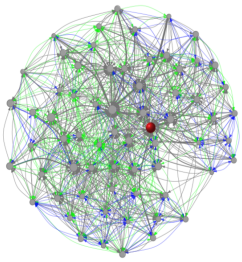
the number of global optima should decrease. Figure 4 shows the newly constructed local optima networks for the three respective datasets when removing the irrelevant features as determined in Table 5.



(a) breast-w LON (b) vowel LON (c) zoo LON

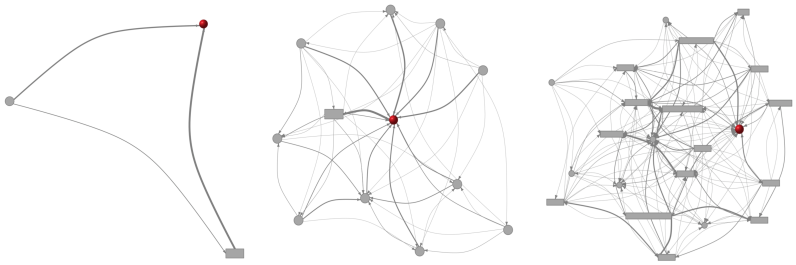
Fig. 4. Local optima networks after feature removal.

Visually, the local optima networks for the breast-w and vowel datasets indicate the existence of a single global optimum after removal of the irrelevant features. The zoo dataset, however, still has two global optima. Another feature has now been highlighted as being irrelevant, after removal of the first irrelevant feature that was detected in Table 5. This occurrence highlights the complex relationship between features and how the existence of one feature may affect the informational value of other features. The irrelevant feature is determined by calculating the symmetric intersection between the sets of features for the two global optimum solutions. After removal of the irrelevant feature, the local optima network for the zoo dataset is shown in Figure 5 (now with two features removed). The fitness landscape of the zoo problem now has a single global optimum.



(a) zoo LON

Fig. 5. Local optima networks after feature removal, round 2.



(a) breast-w *CM-LON* (b) vowel *CM-LON* (c) zoo *CM-LON*

Fig. 6. Compressed Monotonic Local Optima Networks after feature removal

The number of neutral edges in the zoo local optima network after removal of two irrelevant features still stands out among the non-global local optima. The number of neutral edges for the vowel and breast-w datasets after removing the irrelevant features has visually decreased. The compressed monotonic local optima networks are given in Figure 6, where all irrelevant features have been removed.

The contrast between the compressed monotonic local optima networks for Figure 6 (after irrelevant feature removal) and Figure 3 (before irrelevant feature removal) shows the reduction in the number and size of the local optima neutral plateaus. The local optima network metrics for the now single global optima datasets are given in Table 6.

Table 6. Local Optima Network Metrics after Feature Removal.

Dataset	*N_optima*	*N_global*	*Bsize_global*	*Str_global*	*N_coptima*	*Prop_neutral*
zoo	656	1	0.0039	0.0452	193	0.7057
vowel	13	1	0.1252	0.5948	12	0.0769
breast-w	3	1	0.3513	0.6522	3	0.25

The feature selection problem dimensionality reduction obtained by removing the irrelevant features is apparent in the now reduced number of local optima as indicated in Table 6.

The visual observation of the reduction of neutrality in local optima networks in Figure 4 is confirmed by the metrics in Table 6. The proportion of connected nodes with neutral edges for the vowel and breast-w datasets decreased drastically from 0.6428 to 0.0769 and 0.65 to 0.25, respectively. The zoo dataset showed a smaller reduction less, from 0.7282 to 0.7057.

The removal of a feature changes the search space of the feature selection problem, which will affect the behaviour of the feature selection search process. The distribution of local optima and the shape of the fitness landscape changes, therefore search algorithms have different information to guide search. Running the feature selection algorithms on the now reduced datasets yield the performance as summarised in Table 7. The values in bold represent an improvement in performance.

Table 7. Feature Selection Algorithm Performance After Feature Removal.

Dataset	$P_{ratio}(Filter)$	$P_{ratio}(SFS)$	$P_{ratio}(GA)$	*GA Success Ratio*
zoo	0.9711	0.8281	**0.9890**	0.1333
vowel	0.8228	1.0	**0.9888**	**0.8**
breast-w	**0.9923**	**0.9923**	**0.9997**	**0.9666**

The genetic algorithm wrapper method performed better for all of the datasets after removal of the irrelevant features, with significant improvement in the success ratio for the vowel and breast-w datasets improving by 14.3%. As a result of the increased success ratio, the performance ratio for the problems also increased, albeit slightly. A Mann-Whitney U test on the samples of performance ratio values for each run of the genetic algorithm, was conducted at a $p = 0.05$ level of significance. The Mann-Whitney U test indicated that the performance improvement is not statistically significant with regards to the performance ratio. The zoo dataset did not show an improvement in success ratio and showed a negligible increase in performance ratio at the third decimal.

The performance of the filter and sequential forward selection algorithms was not affected by the removal of the irrelevant features for the zoo and vowel datasets. A small to negligible improvement in performance ratio is observed on the breast-w dataset for the filter and wrapper methods. The GA wrapper method did not improve in its success ratio, which could be due to other factors affecting the difficulty of the problem. The LON for this problem is observed to contain local optima neutral plateaus that persist after the removal of the neutrality directly attributed to irrelevant features, as seen in Figure 6.

As previously mentioned, the number of local optima are correlated with the size of the search space but are also affected by the presence of irrelevant features. The zoo dataset, before irrelevant feature removal, exhibits a ratio of

0.0376 ($\frac{4938}{2^{17}}$) local optima with respect to the total number of solutions. After irrelevant feature removal the ratio changes to 0.02 ($\frac{656}{2^{15}}$). The vowel and breast-w ratio of local optima to number of solutions changed from 0.0051 and 0.0781 to 0.0063 and 0.0469, respectively. The removal of the irrelevant features therefore resulted in a reduction in the ratio of local optima to number of solutions by 46.8% for the zoo dataset and 39.9% for the breast-w dataset. The vowel dataset shows an increase in the number of optima at the third decimal which is seen as a negligible change in local optima ratio to number of solutions.

The normalised incoming strength of the global optimum consistently increases for all the datasets. The increased strength to the global optimum from connected local optima indicates that there is a better chance for search to move from a connected local optimum to the global optimum. Removing irrelevant features therefore resulted in a higher probability of reaching a global optimum and can be regarded as a reduction in search difficulty for algorithms utilising neighbourhood in the search space.

5 Conclusion

The purpose of this paper was to conduct an analysis on the fitness landscapes of the feature selection problem.

The construction of local optima networks for the feature selection problem exhibited interesting problem properties that were previously unknown. These insights allowed a better understanding of the performance differences between feature selection algorithms for search spaces with and without irrelevant features.

The construction of the local optima networks for the real world datasets that were used, showed that the feature selection problem had either a single global optimum or several global optima. For problems with multiple global optima, the global optima were located within a single bit-flip neighbourhood of other global optima forming a neutral plateau. Since the inclusion or exclusion of a particular feature had no effect on fitness (classification accuracy), it could be deduced that these features were irrelevant to the classification task.

The incoming strength of escape edges to the global optima represents the probability of moving towards the global optima from a connected local optima. The consistent increase of the normalised incoming strength, after irrelevant feature removal, of the global optimum indicates a reduction in problem difficulty. The proportion of local optima connected via neutral edges in the local optima network consistently reduced when irrelevant features were removed. The proportion of local optima connected via neutral edges fluctuated from a low amount to a very high amount on a per problem basis.

The performance ratio of the filter and wrapper feature selection algorithms did not significantly improve after removing irrelevant features. The success ratio of the genetic algorithm did improve, and remained the same for one dataset, after removing irrelevant features. There was no statistically significant improvement in the performance ratio of the genetic algorithm wrapper method. The

improvement in success ratio serves as an indication that the genetic algorithm found the global optima more often and is therefore an indication that the problem difficulty decreased.

Irrelevant features introduce misinformation in the search space, as evident in the reduction of the proportion of local optima that exists in relation to the size of the search space. Removal of irrelevant features generally reduces the number of local optima, both as a result of the dimensionality reduction and removal of misinformation.

The insights gained on the effect of irrelevant features on the fitness landscape and the global structure of the feature selection problem informs on the behaviour of the feature selection algorithms and contributes to the greater goal of meta-learning. Possible future work includes the study of the effect on the fitness landscape using different classifiers and other approaches to feature selection. The effect of increased dimensionality on the fitness landscape is another area of interest for future research.

References

1. Aha, D., Kibler, D.: Instance-based learning algorithms. Machine Learning **6**, 37–66 (1991)
2. Amaldi, E., Kann, V.: On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. Theoretical Computer Science **209**(1), 237–260 (1998)
3. Ben-David, A.: Comparison of classification accuracy using Cohen's Weighted Kappa. Expert Systems with Applications **34**(2), 825–832 (2008)
4. Chandrashekar, G., Sahin, F.: A survey on feature selection methods. Computers & Electrical Engineering **40**(1), 16–28 (2014)
5. Cover, T.M., Thomas, J.A.: Elements of information theory. John Wiley & Sons (2012)
6. Daolio, F., Verel, S., Ochoa, G., Tomassini, M.: Local optima networks of the quadratic assignment problem. In: Evolutionary Computation (CEC), 2010 IEEE Congress on. pp. 1–8. IEEE (2010)
7. Doye, J.P., Massen, C.P.: Characterizing the network topology of the energy landscapes of atomic clusters. The Journal of Chemical Physics **122**(8) (2005)
8. Fieldsend, J.E.: Computationally efficient local optima network construction. In: Proceedings of the Genetic and Evolutionary Computation Conference Companion. pp. 1481–1488. ACM (2018)
9. Goldberg, D.E.: Genetic algorithms in search, optimization and machine learning. Addison-Wesley (1989)
10. Guyon, I., Elisseeff, A.: An introduction to variable and feature selection. Journal of Machine Learning Research **3**(Mar), 1157–1182 (2003)
11. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: An update. SIGKDD Explor. Newsl. **11**(1), 10–18 (Nov 2009). https://doi.org/10.1145/1656274.1656278
12. Hancer, E., Xue, B., Zhang, M.: Differential evolution for filter feature selection based on information theory and feature ranking. Knowledge-Based Systems **140**, 103–119 (2018)

13. Kira, K., Rendell, L.A.: The feature selection problem: Traditional methods and a new algorithm. In: AAAI. vol. 2, pp. 129–134 (1992)
14. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial intelligence **97**(1), 273–324 (1997)
15. Lichman, M.: UCI machine learning repository (2013), http://archive.ics.uci.edu/ml
16. Malan, K.M., Engelbrecht, A.P.: A survey of techniques for characterising fitness landscapes and some possible ways forward. Information Sciences **241**, 148–163 (2013)
17. Ochoa, G., Tomassini, M., Vérel, S., Darabos, C.: A study of NK landscapes' basins and local optima networks. In: Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation. pp. 555–562. ACM (2008)
18. Ochoa, G., Veerapen, N., Daolio, F., Tomassini, M.: Understanding phase transitions with local optima networks: number partitioning as a case study. In: European Conference on Evolutionary Computation in Combinatorial Optimization. pp. 233–248. Springer (2017)
19. 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, pp. 233–262. Springer (2014)
20. Pitzer, E., Affenzeller, M.: A comprehensive survey on fitness landscape analysis. In: Recent Advances in Intelligent Engineering Systems, pp. 161–191. Springer (2012)
21. Reidys, C.M., Stadler, P.F.: Combinatorial landscapes. SIAM review **44**(1), 3–54 (2002)
22. Verel, S., Daolio, F., Ochoa, G., Tomassini, M.: Local optima networks with escape edges. In: International Conference on Artificial Evolution (Evolution Artificielle). pp. 49–60. Springer (2011)
23. Xue, B., Zhang, M., Browne, W.N.: Particle swarm optimization for feature selection in classification: A multi-objective approach. IEEE transactions on cybernetics **43**(6), 1656–1671 (2013)
24. Yu, L., Liu, H.: Feature selection for high-dimensional data: A fast correlation-based filter solution. In: ICML. vol. 3, pp. 856–863 (2003)