Filter versus Wrapper Feature Selection based on Problem Landscape Features

Werner Mostert
Department of Computer Science,
University of Pretoria
Pretoria, South Africa
u13019695@tuks.co.za

Katherine Malan
Department of Decision Sciences,
University of South Africa
Pretoria, South Africa
malankm@unisa.ac.za

Andries Engelbrecht
Institute for Big Data and Data
Science, University of Pretoria
Pretoria, South Africa
engel@cs.up.ac.za

ABSTRACT

Feature selection is a complex problem used across many fields, such as computer vision and data mining. Feature selection algorithms extract a subset of features from a greater feature set which can improve algorithm accuracy by discarding features that are less significant in achieving the goal function. Current approaches are often computationally expensive, provide insignificant increases in predictor performance, and can lead to overfitting. This paper investigates the binary feature selection problem and the applicability of using filter and wrapper techniques guided by fitness landscape characteristics. It is shown that using filter methods are more appropriate for problems where the fitness does not provide sufficient information to guide search as needed by wrapper techniques.

CCS CONCEPTS

• **Computing methodologies** → **Feature selection**; *Discrete space search*;

KEYWORDS

Feature selection problem, fitness landscapes, Hamming distance in a level, neutrality

ACM Reference Format:
Werner Mostert, Katherine Malan, and Andries Engelbrecht. 2018. Filter versus Wrapper Feature Selection based on Problem Landscape Features. In *GECCO '18 Companion: Genetic and Evolutionary Computation Conference Companion, July 15–19, 2018, Kyoto, Japan,* Jennifer B. Sartor, Theo D'Hondt, and Wolfgang De Meuter (Eds.). ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3205651.3208305

1 INTRODUCTION

Humans are good at establishing patterns and categorizing arbitrary objects by association. This human trait allows for the perception and cognitive detection of arbitrary objects in everyday life, such as recognizing friends and family. Within the field of machine learning, this basic human trait is artificially simulated to accomplish the same goal. Fields such as computer vision rely heavily on classification based on features of real world problems [24].

GECCO '18 Companion, July 15–19, 2018, Kyoto, Japan

ACM ISBN 978-1-4503-5764-7/18/07…$15.00
https://doi.org/10.1145/3205651.3208305

Many different classification algorithms, otherwise referred to as classifiers, have been developed [19] to recognize patterns in often large amounts of data. Classifiers generally work towards a generalized goal: some conclusion is made as to the membership of a category or class for a specific instance of data. The membership is based on patterns and associations made with regards to a set of parameters in observed data. The parameters in the observed data are often also called features or attributes of the data. Chandrashekar and Sahin [6] define a feature as "an individual measurable property of the process being observed".

Intuitively, when a human is given the task of recognizing a friend or family member, what is it that determines how the individual is identified? The observer may perhaps take into account physical features such as the subject's physical features, voice or smell. Given the fact that a classifier needs some set of features to perform classification, it becomes a complex problem to decide on which features are the most information rich to provide for performant classification. Heuristics have been used to great success for classification in combinatorial problems, with little explanation as to why they perform well or not [4]. The features that are used to perform classification could greatly influence the performance of heuristic techniques.

In order to further the development of a generalized theoretical framework for feature selection and to better understand the feature selection problem, this paper analyzes the fitness landscape of the feature selection problem with respect to a simple and non-stochastic classifier. Since feature selection is a binary problem, the landscape that is analyzed is discrete. A variety of data sets are considered with varying numbers of features, types of features and data set sizes.

There are various feature selection algorithms which have been developed to date, and there is still current research on the topic [8, 12, 14, 15, 25, 26]. These generally fall into three categories namely filter methods, wrapper methods and embedded methods [6]. There is evidently no shortage of algorithms for conducting feature selection. The algorithms are diverse in how the problem is approached, however, there seems to be no theoretical framework in place to guide researchers in making decisions on which algorithms to use [10].

Filter methods establish a ranking of how important features are based on information regarding the characteristics of the features and the relationships between said features, using measures such as correlation or mutual information. Wrapper methods, on the other hand, use classifier fitness with feature subsets to guide the search in determining feature relevance for performant classification.

It is hypothesized that it would be more appropriate to use basic filter techniques in the case where the fitness landscape does not provide sufficient information to guide the search for optimal feature subsets.

Section 2 gives an overview of the feature selection problem and fitness landscape analysis. Section 3 gives an overview on the feature selection algorithms that were utilized. Landscape characteristics that are considered in this paper are discussed in Section 4. Section 5 describes the experimental process of conducting the fitness landscape analysis and calculation of landscape characteristics with regards to the respective feature selection algorithms. Finally, the experimental results are given in Section 6.

2 FEATURE SELECTION

The feature selection problem is that of selecting a good or relevant subset of all available features to accomplish the task at hand. Feature selection is beneficial in order to better understand and visualize data, as well as to improve classification performance by effectively reducing the dimensionality of the problem [10]. Modern problems such as image classification suffers from extremely high dimensionality, thus affecting classifier performance. There exist many approaches for feature selection, of which many have proven to be problem dependent and have had varying measures of success [6].

Since choosing a subset of features from the complete feature set F is a combinatorial problem, the subset $F_s \subseteq F$ can be represented as a binary string S_i. An 'on' bit in the string represents the inclusion of the feature, whereas an 'off' bit represents its exclusion. In this format it is possible to model a full permutation of all possible inclusions and exclusions of features of the complete feature set F. In order to analyze a fitness landscape, the precondition is that it is computationally possible to calculate a fitness value at a given point in the landscape, $f(S_i)$, where f is the fitness function. A unique bit string is referred to as a solution within the context of this paper.

The issue of feature irrelevance comes to light since two features that are considered within mutual exclusion, could be useless, but the union of these features could be information rich [10]. A primitive approach to solving this problem would be to do an exhaustive search of the combination of features which results in optimal performance. Given a small number of features this is conceptually possible, however, as stated by Amaldi *et al.* [2] the full permutation of feature sets for a highly dimensional problem is a non-polynomial (NP)-hard problem.

Although feature selection methods may be used to improve classifier performance, Guyon and Elliseeff [10] found that for problems of high dimensionality, the performance increase is not always significant. Since the feature selection problem is of a complex nature, it is proposed that landscape analysis be conducted to obtain a more detailed understanding of the problem.

3 FEATURE SELECTION ALGORITHMS

Two categories of feature selection algorithms are considered in this study, namely filter methods and wrapper methods. Within the category of wrapper techniques, the sequential selection wrapper and heuristic search wrapper using a genetic algorithm, are considered.

3.1 Filter Methods

Filter methods work on the premise of ranking features in terms of their relevance. The definition of what makes a feature relevant is an indirect notion, and is measured by calculating metrics such as the correlation or mutual information between features. As previously mentioned, a feature regarded in isolation may be of little value to a classifier, but regarded in conjunction with another feature it may prove to be information rich in determining the class of a problem instance. The criteria used by filter methods takes this into account and disregards the fitness of the classifier when ranking features.

3.2 Wrapper Methods

Wrapper methods make use of the classifier fitness when selecting features. Various search algorithms may be employed to determine the optimal set of features to be used. Since using a brute force approach is not computationally feasible, simple approaches such as a sequential selection search or heuristic search can be used. The sequential feature selection (SFS) algorithm [21] starts with an empty set of features, including one additional feature at a time and evaluating each feature set to determine the classification accuracy. Given that one feature is selected, the immediate neighborhood is considered and sequentially added and re-evaluated in search of a better solution. The stopping condition for the algorithm is the case where the current solution is better than any of the solutions within the immediate neighborhood. Some similarity in the behavior of the algorithm can be observed to that of a greedy hill climbing algorithm.

The SFS algorithm may be seen as naive; thus heuristic search algorithms may be employed to take a more intelligent approach, at the cost of increased computation time. Genetic algorithms [18] are population based algorithms used for search and optimisation problems [9], which translates well to use for feature selection.

4 FITNESS LANDSCAPE CHARACTERISTICS

An underlying fitness landscape can be a valuable tool for analysis of heuristic search algorithms [20], since some fitness landscapes possess structural attributes that can lead search algorithms to good or bad solutions [17]. There are various characteristics of a fitness landscape that can be investigated and also many different techniques that may be applied in order to analyze these characteristics [17]. Fitness landscape characteristics such as fitness distributions, epistasis, neutrality, amongst others, can prove vital in understanding why certain algorithms perform well on sets of problems and why others do not.

This section describes the fitness landscape characteristics that are considered in this paper. An overview is given on fitness frequency distributions, Hamming distance in a level, and landscape neutrality in order to describe the fitness landscape of the problem.

4.1 Fitness Frequency Distribution

The fitness frequency distribution characteristic can be calculated by the fitness function alone – dependent on the grouping strategy

used. A simple grouping strategy that can be used is to bin the fitness function values into sub-ranges. This is only possible if the upper and lower bounds of the fitness function are known. Given a finite fitness range, a B number of bins can be used to divide the range into bins of equal size. The bins are initialized to a zero count and for each solution in the sample, the count of the bin corresponding to the fitness of the solution is incremented during the sampling. This results in a histogram of the number of solutions in the sample in a number of fitness value ranges. By constructing the fitness frequency distribution histogram one may profile the problem to answer the question of how many configurations of the combinatorial problem result in a specific range of fitness values. The fitness distribution as calculated here has some similarities with the density of states technique [3].

4.2 Hamming Distance in a Level

The Hamming distance in a level as proposed by Belaidouni and Hao [4] is a measure of the similarity, or the lack thereof, of problem instances within a range of fitness values. The authors describe a concept called an iso-cost level which is essentially a set of problem instances that correlate to the same fitness value. The distance D in a set A is the average distance between the elements of A and is defined as [4]:

$$D(A) = \frac{1}{|A^2|} \sum_{(s,s') \in L^2} d(s, s')$$

where L is the iso-cost level and s and s' are problem instances within the level. By using the iso-cost levels one could visualize this technique as calculating the disparity in the other relevant dimensions with respect to a specific fitness. Alternatively, it can be viewed as a measure to indicate the width of the landscape [4] with respect to each iso-cost level. If the distance of an iso-cost level is large, it indicates that solutions of that fitness value are widely distributed in the search space. On the other hand, if the distance of an iso-cost level is small, it indicates that solutions of that fitness are clustered around a specific point in the search space.

Hamming distance in a level (HDIL) as originally proposed defines a fitness level per unique cost value. A slight adaptation is made in this paper where instead of considering problem instances of a specific cost (fitness value), a grouping strategy is used for problem instances in the range of the bin sizes as used for the fitness distribution calculation.

4.3 Neutrality

The neutral theory argues that nonadaptive neutral mutation takes place for extended periods of times [13], where the landscape will remain at a constant height. This essentially translates to the fact that, in a fitness landscape, one may have various neighboring solutions with the same fitness.

Two solutions in a discrete landscape may be considered neutral if their respective fitness values are equal and they fall within the same neighborhood [22]. Within binary spaces, a simple definition of neighborhood may be solutions where the Hamming distance between solutions is equal to one. A landscape that is regarded as neutral does not necessarily imply that the landscape is flat, but does suggest successive neutrality in fitness [17] on neighbourhood paths.

5 IMPLEMENTATION

This section discusses the experimental process that was followed. Subsection 5.1 discusses the decision on which fitness function to use, as well as the interpretation of the measure used as fitness. The process followed in order to decide on the appropriate data sets to use is discussed in subsection 5.2. Implementation parameters and setup of the feature selection algorithms are discussed in section 5.3. Section 5.4 and 5.5 detail the calculation of the HDIL, fitness frequency distributions and neutrality measures.

5.1 Fitness Function

In this study, a measure of classification accuracy based on a test data set is used as the fitness value for a solution, s. The choice of fitness function is of paramount importance when conducting fitness landscape analysis since different functions result in different landscapes and even the same fitness function used with different notions of neighbourhood will result in different landscapes [17].

Classifiers with stochastic elements, such as artificial neural networks using stochastic gradient descent or random decision forests, could very likely provide for good classification accuracy measurements but would introduce noise into the fitness landscape. This is due to the fact that since a stochastic classifier is generally executed multiple times and the resultant mean performance measurement is used. For the produced fitness landscape the difference between resultant error curves within the hyper-dimensional problem space would become fuzzy. It is desirable to be able to reliably reproduce the same landscape for a non-dynamic landscape since this allows for landscape characteristics that are independent of change in time, or some other external variable that may affect the fitness calculation. Therefore the classic k-nearest-neighbor [1] with $k = 3$, a simple non-stochastic classifier was used. The value of k is decided as 3 since it produces diverse, a wide range of low to high, classification accuracies for a number of problems. It is possible that other values of k would prove similarly diverse or possibly perform better and therefore change the fitness landscape. The focus within the context of this paper is on the fitness landscape analysis of a static fitness function and the effect of using subsets of features on the landscape, and not on optimisation of the fitness function itself.

In order to obtain a measure of classification accuracy, Cohen's Kappa-statistic, a non-biased measure, is used. Cohen's Kappa is defined as [5]:

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

where P_c is the agreement probability as a result of randomness and P_0 is the total agreement probability. The concept of agreement probability is quite simple. Given a confusion matrix as in Figure 1, the Kappa statistic would be calculated as follows:

$$\begin{aligned} P_c &= (\frac{86}{100})(\frac{84}{100}) + (\frac{14}{100})(\frac{16}{100}) = 0.0162 \\ P_0 &= \frac{75}{100} + \frac{5}{100} = 0.8 \\ K &= \frac{0.8 - 0.0162}{1 - 0.0162} = 0.7967 \end{aligned} \tag{2}$$

Table 1: Data Sets

Data Set Number & Name	Nominal Features	Continuous Features	No. Classes	No. Data Elements	No. Features
Data Set 1 - Anneal.ORIG	Yes	Yes	6	898	39
Data Set 2 - Audiology	Yes	No	24	226	70
Data Set 3 - Colic.ORIG	Yes	Yes	2	368	28
Data Set 4 - Cylinder Bands	Yes	Yes	2	512	40
Data Set 5 - Hepatitis	Yes	No	2	155	19
Data Set 6 - Vowels	Yes	No	11	990	14
Data Set 7 - Page Blocks	No	Yes	5	5473	10

The rate of agreement in this case is therefore 0.7967. The Kappa statistic ranges from total disagreement at -1 through completely random classification at 0, to 1 which indicates total agreement. The Kappa statistic allows for the level of agreement for each class label to be measured. This is important since for a raw count of correct classification instances, the results may be statistically biased due to an overwhelming presence or absence of a specific class in an observed data set. The Kappa statistic removes the bias element in correct classification count [5] and is a normalized classification measure in the range $[-1, 1]$.

	A	B	Total
A	75	11	86
B	9	5	14
Total	84	16	100

Figure 1: Confusion Matrix for 100 samples

5.2 Frameworks and Data Sets

The UC Irvine Machine Learning Repository [16], which contains a wide variety of data sets that can be used for various machine learning objectives, was used to source data. A total of 81 data sets were originally considered. The data sets that were considered contained a variety of nominal and numerical data elements or features, where some data sets exclusively contained one of the types of data elements. Data sets with a large range of number of features were considered with varying numbers of data instances. No specific preprocessing of the data sets took place.

A subset of data sets were selected to supply meaningful information for contrasting fitness landscape characteristics with feature selection algorithm performance. Using the Weka Machine Learning software development kit [11] as an implementation framework, the classification accuracy was measured for each data set after applying the respective feature selection algorithms. The feature selection algorithms were applied to 50% of the instances for the respective data sets, after which the classification accuracy was determined on the other 50% of data instances, in order to avoid statistical bias.

Of the original 81 data sets, seven were selected to be considered in this paper since they exhibited a notable difference in performance between the fitness obtained when using a filter technique in comparison to using a sequential selection or genetic search wrapper technique. Table 1 summarises the seven data sets that were used in the study.

5.3 Feature Selection Algorithms

The classic information gain [7] feature evaluation measurement was selected for use by the basic filter method. The features were then ranked in accordance to the relevance of the features from high to low. Given a list of features, sorted according to relevance determined by the information gain of each feature, each linear combination of features from most relevant to lowest relevance was considered and the fitness of the solutions were evaluated. For example, given a list of features sorted by decreasing information gain, $F = 1, 5, 2, 4, 3$, the following five feature sets were considered: $\{1\}$ (feature 1 alone), $\{1, 5\}$, $\{1, 5, 2\}$, $\{1, 5, 2, 4\}$, and $\{1, 5, 2, 4, 3\}$. The feature set with the highest fitness value was then selected as the output of the filter method.

Both wrapper techniques used the k-nearest-neighbor classifier as the fitness function, as previously described. The sequential selection wrapper technique used the SFS algorithm as defined by Pudil *et al.* [21]. The heuristic wrapper method made use of the simple genetic algorithm described by Goldberg [9]. The following parameters were set for the genetic algorithm:

- Population size : 20
- Number of generations : 20
- Crossover probability : 0.6 (60%)
- Mutation probability : 0.033 (3.3%)

The genetic algorithm parameters above are the Weka [11] defaults for a simple genetic algorithm. The genetic algorithm parameters are kept the same for all executions on the various data sets in order to limit the statistical bias introduced by correct parameter optimisation for the genetic algorithm, which may be different per data set.

5.4 Sampling Methods

Taking a sample solution from a uniform distribution within the context of binary landscapes entails initializing a random bit string which maps to a configuration within the space of all possible feature selection combinations.

Problems that are of high dimensionality have an excessively large problem space, and small sample sizes with respect to the size of the search space (i.e. dimensionality of the problem in this case) may heed misleading results. Small sample sizes in a large search

space means that only a very small proportion of the problem space will be considered. Sampling of a uniform distribution may also present an issue, since landscape characteristics that are reliant on the concept of neighborhood will be suppressed.

When calculating the fitness frequency distribution and HDIL using samples obtained from a random, uniform sampling technique, the loss of neighborhood information does not present an issue, since these measures are not concerned with the topological features within the neighborhood of solutions. When approximating the neutrality of the landscape, a random walk was used as sampling method since neighborhood is important.

Since a variety of different data sets were considered, ranging from 10 to 70 features, it would not be fair to use the same sample size for the smaller and larger problem spaces. A constant of $C = 50$ was used to scale the sample size with the dimensionality (size) of the problem space. The number of random samples was then calculated as:

$$S = CI$$

where I is the number of features in the data set considered.

A minimum number of features constraint was imposed to solutions that were sampled. A minimum number of 10% of all features was required in order for a solution to be considered in the sample. This was done in an attempt to cut down on the classifier performing poorly simply because of the fact that it is considering a very small number of features.

5.5 Neutrality

A number of techniques have been proposed for characterising the level of neutrality in fitness landscapes [17]. Van Aardt *et al.* [23] proposed two normalized measures of neutrality within the range $[0, 1]$ as an approximation of the degree of neutrality in a landscape, M_1, and the relative size of the largest neutral region within the landscape, M_2. These measures are based on a sample of solutions generated by a random walk through the search space.

The random walk was implemented in the binary search space as follows: starting at a initial random solution, the problem space was navigated by flipping a random bit in the bit string, thereby exploring the problem space whilst preserving topological information in the neighborhood. The result of a random walk is a sequence of bit string solutions to the feature selection problem. Given this sequence, every three consecutive solutions were considered as a 3-point structure and determined to be neutral or not, depending on whether the fitness values remained constant between the three neighbouring solutions on not.

Van Aardt *et al.* defined the measures of neutrality as [23]:

$$M_1 = \frac{s_{neutral}}{|W|}$$

where $s_{neutral}$ is the number of neutral 3-point structures in the random walk W, and $|W|$ is the total number of 3-point structures in W.

M_2 is defined as:

$$M_2 = \frac{w_{max}}{|W|}$$

where w_{max} is the largest sequence of consecutive three point structures that were neutral in the walk W.

It would not be fair to conduct walks of the same length in the discrete landscape for problems that differ in size of dimensionality. If the same number of steps were taken for a smaller landscape, it would be explored less than the larger landscape and may therefore produce misleading results. Therefore, the number of steps to take, w, was calculated as,

$$w = k_w I$$

where k_w scales the walking distance. In this case, k_w was set to 10.

Due to the fact that the random walk is by its very nature a stochastic process, the mean result for the two metrics were recorded for a total of 30 independent random walks.

6 RESEARCH RESULTS

The following section presents results of the feature selection algorithms on the seven datasets. Results are reported as classification accuracy (as measured by the Kappa statistic) of the k-nearest neighbour classifier using the features selected. Section 6.2 presents the results from the fitness landscape analysis on the seven data sets and the link between the landscape characteristics and algorithm performance are discussed.

6.1 Feature Selection Algorithm Performance

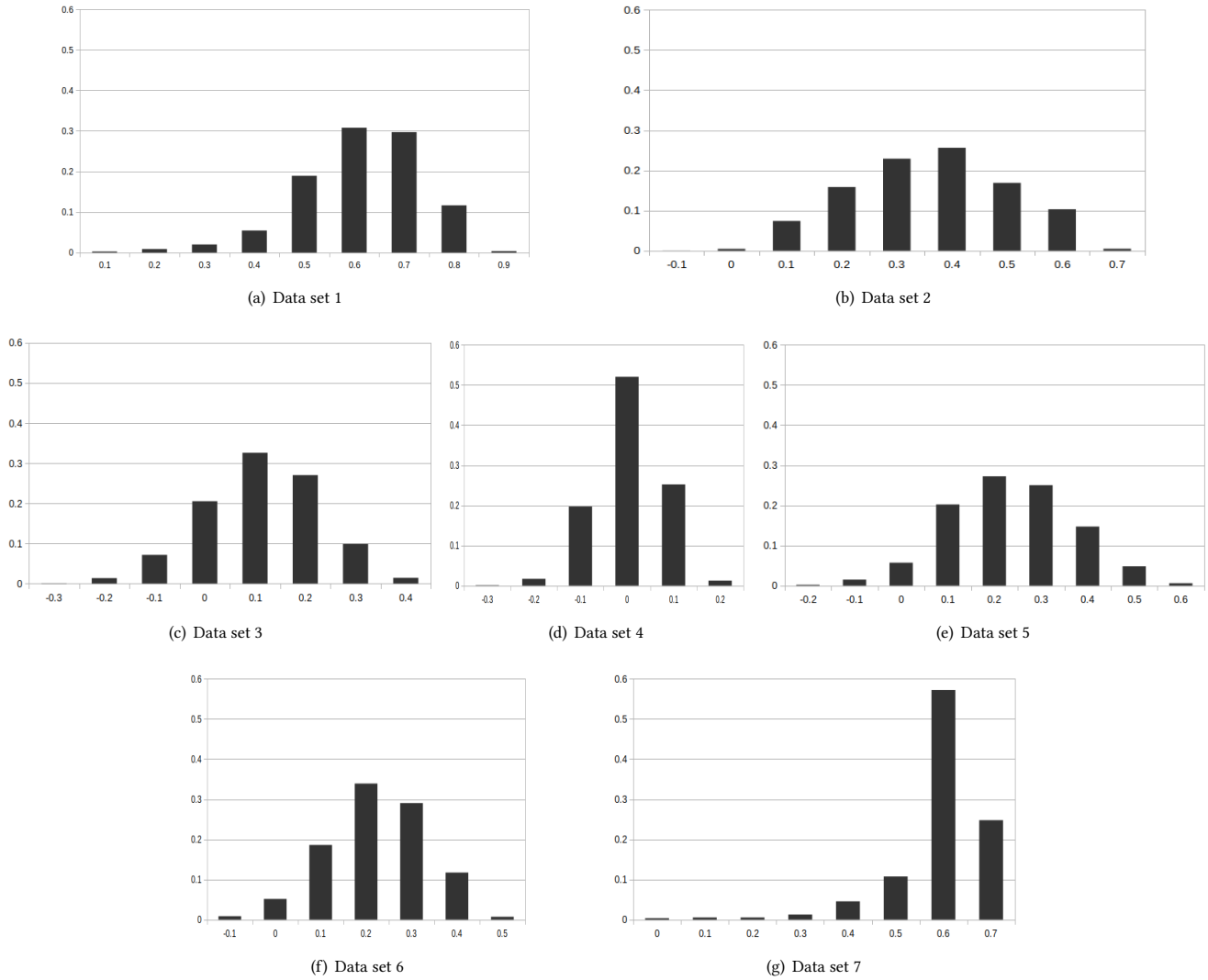
The datasets were split in half for feature selection and classification. After performing the feature selection based on the first half of each of the datasets, the result of each algorithm was then tested for accuracy of classification using the remaining half of each dataset. Table 2 indicates the classifier accuracy using the features selected by the basic filter method, the sequential selection wrapper method and the heuristic search with a genetic algorithm method. For each data set, the best result (highest Kappa statistic) is highlighted in bold. It can be seen that the filter method outperformed the wrapper techniques on four of the data set, but all techniques performed very poorly on data set 4.

Table 2: Feature Selection Techniques - Fitness

Data Set	Filter	Seq. Selection Wrapper	Heuristic Wrapper
1	0.7274	0.6928	**0.8212**
2	**0.6618**	0.4685	0.3393
3	0.1375	0.4529	**0.4713**
4	**0.1814**	-0.0879	-0.1806
5	0.5195	0.6335	**0.7921**
6	**0.4430**	0.2386	0.2748
7	**0.4721**	0.3762	0.3575

6.2 Fitness Landscape Analysis

Uniform random samples of size $50 \times I$ (where I is the number of features in the dataset) were generated for each of the seven data sets. For each random solution (a bit string of selected features), the fitness was determined by classifying the data using k-nearest neighbour. Figure 2 shows the frequency of solutions in fitness ranges binned by 0.1 interval of the Kappa statistic. Bins with a



(a) Data set 1

(b) Data set 2

(c) Data set 3

(d) Data set 4

(e) Data set 5

(f) Data set 6

(g) Data set 7

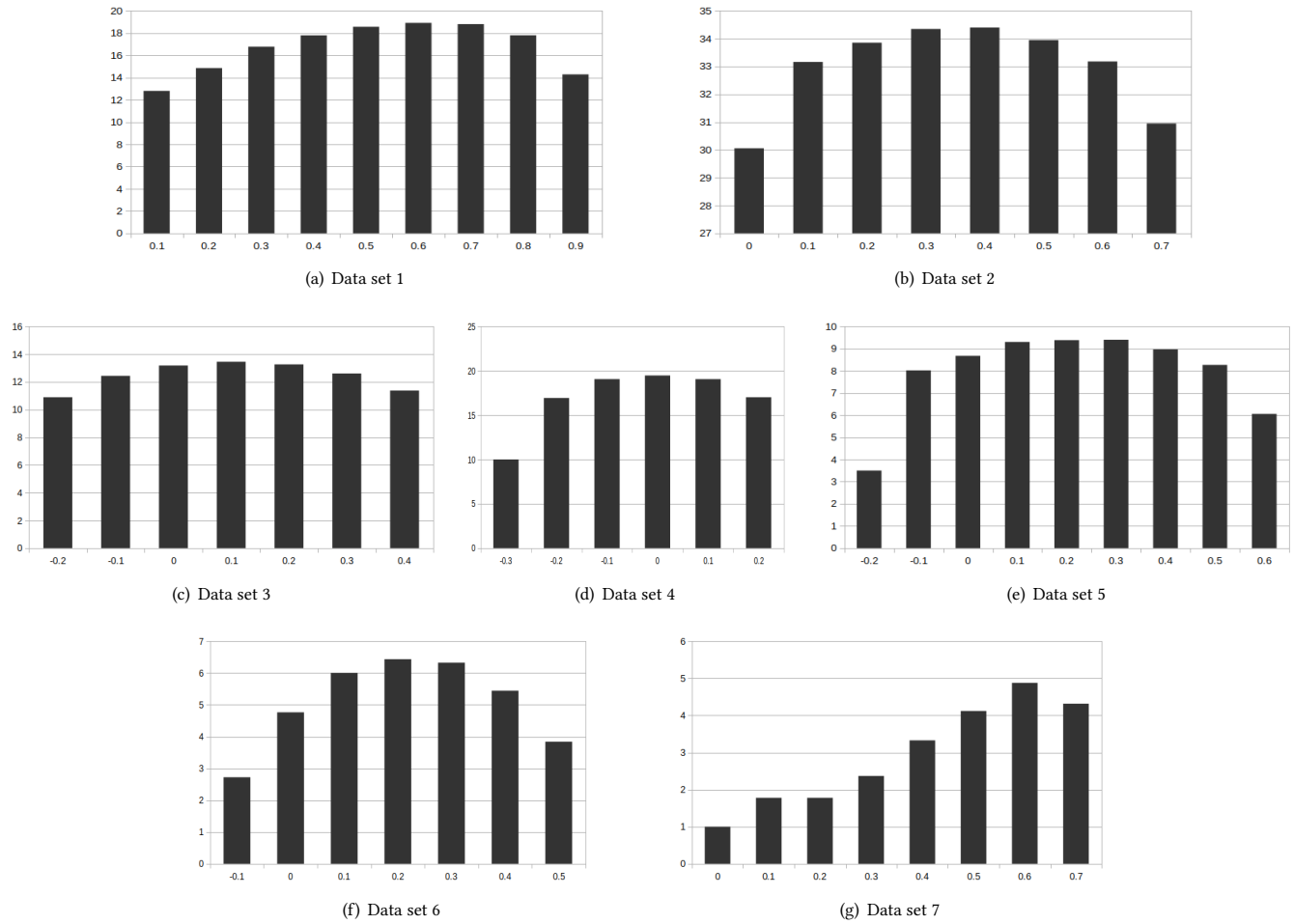
Figure 2: Fitness frequency distributions based on samples

count of 0 are omitted. The horizontal axes in Figure 2 corresponds to the Kappa statistic bins in increasing levels of accuracy. The vertical axes give the proportion of solutions in the sample that fall into each fitness bin. For comparison, all graphs in Figure 2 are plotted with the same range from 0 to 0.6.

Two of the data sets are characterized by very narrow fitness distributions with over 50% of the data points in a single fitness band: data set 4 with 1040/2000 points (52%) in the $[0, 0.1)$ band, and data set 7 with 314/550 points (57%) in the $[0.6, 0.7)$ band. These problems are therefore characterized as having a larger proportion of similar fitness values, providing less variation in fitness information to guide search. Since filter methods do not use fitness information for selecting features, these methods should not be negatively affected by these narrow fitness distributions in the way that wrapper methods could be. It is observed that, for these two data sets, the filter method outperforms the wrapper methods. Note, however, that the inverse is not true, since there are other data sets (data sets 2 and 6) without such narrow distributions on which the filter method performs better than the wrapper methods.

The Hamming distance in a level of the seven data sets are illustrated in Figure 3. The bins with 0 solutions are also omitted. The Hamming distance in a level is indicative of the width of the landscape for the different fitness levels. In other words, it indicates to which degree the solutions within the specific fitness level are clustered together or spread apart. Since Hamming Distance measures the number of bits that differ between two bit strings, the maximum possible Hamming distance between two bit strings is the length of the bit string, which for the binary feature selection landscape is I, the number of features in the data set.

It is informative to consider the HDIL graph with the corresponding fitness frequency distribution of the same data set: where the fitness distribution graph gives an indication of the number of



(a) Data set 1 (b) Data set 2

(c) Data set 3 (d) Data set 4 (e) Data set 5

(f) Data set 6 (g) Data set 7

Figure 3: Hamming distance in a level based on samples

solutions in the sample in a particular bin, the HDIL gives an indication of the distance between these solutions. For example, Figure 2 shows that data set 1 contained very few sampled solutions in the highest fitness band (bin 0.9). Figure 3 shows that these solutions were on average a Hamming distance of over 14 apart in the search space.

The HDIL profile could provide information relevant in choosing between a sequential wrapper and a heuristic wrapper technique. Sequential selection algorithms can be viewed as a form of local search, because they start with an empty attribute set and then add attributes one by one using the fitness as a guide. This means that they only consider neighbours one Hamming distance away each time. If the landscape has the good solutions spread far away from each other (HDIL is large for the highest fitness bins), then sequential wrapper methods should not do well, because the search path cannot hop across basins. A heuristic wrapper technique should not suffer from this limitation, due to the wider exploration of the search operators. It is observed that for data set 1, which had high HDIL values in the high fitness bins, the heuristic wrapper technique outperformed the sequential wrapper technique. Further work is needed in extracting normalised numerical indicators from the HDIL profiles for comparison between different data sets.

Table 3 shows the degree of neutrality in a landscape M_1, and the relative size of the largest neutral region within the landscape, M_2.

Table 3: Neutrality

Data Set	M_1	M_2
1	0.63308	0.04482
2	0.79552	0.04378
3	0.12174	0.01350
4	0.35347	0.01838
5	0.25059	0.03183
6	0.00612	0.00612
7	0.12206	0.02590

Two data sets exhibit high levels of neutrality, data set 1 at 0.6331 and data set 2 at 0.7955. It can be noted that the filter methods for

both of these data sets perform well. This occurrence can also be attributed to the fact that filter methods would not be affected by the neutrality of the landscape. Data set 4 exhibits some degree of neutrality at 0.3535, where the filter method performed better than the wrapper methods. One may observe at this point that, for all three of the landscapes that exhibit neutrality, the filter method performed better than the sequential selection wrapper technique. In one case, for data set 1, the genetic search wrapper technique outperformed the filter method. This may be due to the exploratory potential of the genetic search technique being able to traverse the neutral plateaus of the landscape, whereas the sequential selection wrapper technique may not be able to reach better solutions due to its neighborhood being too restrictive and its greedy nature of only considering fitter solutions in the immediate neighborhood.

Neutrality measure $M2$ indicates the relative size of the largest neutral region within the landscape. All of the data sets exhibited low values indicating that there are various neutral regions within the landscape instead of a small number of connected neutral networks.

7 CONCLUSION

This paper performed an initial investigation into the landscape characteristics of the binary feature selection problem. It was observed that for data sets exhibiting narrow fitness distributions, the filter method outperformed the sequential feature selection and genetic search wrapper methods. This seems to indicate that filter methods are more appropriate for problems where the fitness does not provide sufficient information to guide search as with wrapper techniques.

The Hamming distance in a level shows potential as an indicator of problem difficulty for local search techniques such as the sequential selection wrapper technique, but requires more work in developing normalised metrics for comparison between problems.

The filter method performed better than the sequential feature selection on data sets exhibiting larger amounts of neutrality. In some cases the genetic search wrapper technique performed better, possibly due to its potential to explore the landscape even for solutions surrounded by neutral regions.

Further work is needed in investigating the use of other existing landscape analysis techniques suited for binary landscapes. In addition, the effect of parameters on the characteristics of the landscape, such as k in the k-nearest-neighbour algorithm, is a possible area of further research. By using a fuller set of numerical landscape characteristics, the use of data mining to predict feature selection algorithm performance can also be investigated.

REFERENCES

[1] D. Aha and D. Kibler. 1991. Instance-based learning algorithms. *Machine Learning* 6 (1991), 37–66.

[2] Edoardo Amaldi and Viggo Kann. 1998. On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems. *Theoretical Computer Science* 209, 1 (1998), 237–260.

[3] Torsten Asselmeyer, Werner Ebeling, and Helge Rose. 1996. Smoothing representation of fitness landscapes, the genotypephenotype map of evolution. *BioSystems* 39, 1 (1996), 63–76.

[4] Meriema Belaidouni and Jin-Kao Hao. 1999. Landscapes and the maximal constraint satisfaction problem. In *European Conference on Artificial Evolution*. Springer, 242–253.

[5] Arie Ben-David. 2008. Comparison of classification accuracy using Cohens Weighted Kappa. *Expert Systems with Applications* 34, 2 (2008), 825–832.

[6] Girish Chandrashekar and Ferat Sahin. 2014. A survey on feature selection methods. *Computers & Electrical Engineering* 40, 1 (2014), 16–28.

[7] Thomas M Cover and Joy A Thomas. 2012. *Elements of information theory*. John Wiley & Sons.

[8] Ruiquan Ge, Manli Zhou, Youxi Luo, Qinghan Meng, Guoqin Mai, Dongli Ma, Guoqing Wang, and Fengfeng Zhou. 2016. McTwo: a two-step feature selection algorithm based on maximal information coefficient. *BMC bioinformatics* 17, 1 (2016), 1.

[9] David E. Goldberg. 1989. *Genetic algorithms in search, optimization and machine learning*. Addison-Wesley.

[10] Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *The Journal of Machine Learning Research* 3 (2003), 1157–1182.

[11] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explor. Newsl.* 11, 1 (Nov. 2009), 10–18. https://doi.org/10.1145/1656274.1656278

[12] Anil Jain and Douglas Zongker. 1997. Feature selection: Evaluation, application, and small sample performance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 19, 2 (1997), 153–158.

[13] Motoo Kimura. 1983. *The neutral theory of molecular evolution*. Cambridge University Press.

[14] Kenji Kira and Larry A Rendell. 1992. The feature selection problem: Traditional methods and a new algorithm. In *AAAI*, Vol. 2. 129–134.

[15] Ron Kohavi and George H John. 1997. Wrappers for feature subset selection. *Artificial intelligence* 97, 1 (1997), 273–324.

[16] M. Lichman. 2013. UCI Machine Learning Repository. (2013). http://archive.ics.uci.edu/ml

[17] Katherine M Malan and Andries P Engelbrecht. 2013. A survey of techniques for characterising fitness landscapes and some possible ways forward. *Information Sciences* 241 (2013), 148–163.

[18] Melanie Mitchell. 1998. *An introduction to genetic algorithms*. MIT press.

[19] Thuy TT Nguyen and Grenville Armitage. 2008. A survey of techniques for internet traffic classification using machine learning. *IEEE Communications Surveys & Tutorials* 10, 4 (2008), 56–76.

[20] Erik Pitzer and Michael Affenzeller. 2012. A comprehensive survey on fitness landscape analysis. In *Recent Advances in Intelligent Engineering Systems*. Springer, 161–191.

[21] Pavel Pudil, Jana Novovičová, and Josef Kittler. 1994. Floating search methods in feature selection. *Pattern recognition letters* 15, 11 (1994), 1119–1125.

[22] Christian M Reidys and Peter F Stadler. 2001. Neutrality in fitness landscapes. *Appl. Math. Comput.* 117, 2-3 (2001), 321–350.

[23] Willem Abraham van Aardt, Anna Sergeevna Bosman, and Katherine Mary Malan. 2017. Characterising neutrality in neural network error landscapes. In *Evolutionary Computation (CEC), 2017 IEEE Congress on*. IEEE, 1374–1381.

[24] Miles N Wernick, Yongyi Yang, Jovan G Brankov, Grigori Yourganov, and Stephen C Strother. 2010. Machine learning in medical imaging. *Signal Processing Magazine, IEEE* 27, 4 (2010), 25–38.

[25] Yun-Chi Yeh, Liuh-Chii Lin, Mei-Chen Liu, and Tsui-Shiun Chu. 2016. Feature Selection Algorithm for Motor Quality Types Using Weighted Principal Component Analysis. In *Proceedings of the 3rd International Conference on Intelligent Technologies and Engineering Systems (ICITES2014)*. Springer, 151–157.

[26] Lei Yu and Huan Liu. 2003. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *ICML*, Vol. 3. 856–863.