Solutions to Instability Problems with Sequential Wrapper-based Approaches to Feature Selection

Kevin Dunne* Padraig Cunningham† Francisco Azuaje‡

Machine Learning Group
Department of Computer Science
Trinity College, Dublin

Abstract

It is generally accepted that Wrapper approaches will outperform Filter-based approaches to feature selection, particularly in situations where an adequate amount of data is available. What is often overlooked is that Wrapper approaches can be unstable. For instance, different partitionings of the training data can result in different routes through the search space and thus in different feature subsets being selected. In this paper we illustrate examples of this problem and a solution based on the aggregation of several runs of a sequential search is suggested. This is essentially an ensemble solution to instability in feature subset selection and it does seem to stabilise the process.

1 Introduction

When using machine learning techniques to classify or analyse data, we need to identify the specific features or attributes of the available data which are helpful in making a decision. The term *feature subset selection* [1] is given to this process of identifying relevant features.

Feature subset selection enables a classifier to selectively focus its attention on relevant attributes whilst ignoring the (possibly misleading) contribution of irrelevant features. From the point of view of computational efficiency, it is beneficial to have a parsimonious set of features involved in the classification process as many learning algorithms can scale quickly (e.g. $O(N^2)$ or worse) with additional features. The other main advantage is that by concentrating on predictive features only and not considering the irrelevant ones, the accuracy of the classifier may be higher and the association between attributes and target class may be easier to learn.

*Kevin.Dunne@cs.tcd.ie
†Padraig.Cunningham@cs.tcd.ie
‡Francisco.Azuaje@cs.tcd.ie

Another application of feature selection is as a stage in a knowledge-discovery task where identification of strongly correlated features can be used to direct future research. In particular, newly emerging disciplines such as bioinformatics (e.g. gene expression analysis) can definitely benefit from additional insights into the available datasets, some of which can be extremely large both in terms of the number of features and number of samples considered. Identification of a strongly relevant feature can suggest new metabolic pathways or uncover hitherto unrecognised connections between specific cellular processes. Other domains which exploit data mining techniques (for example e-commerce) can also benefit from additional indicators of specific behaviours occurring in large volumes of predominantly numerical data.

One issue that can arise is that different sets of training examplars may lead to different feature masks being suggested. This paper identifies this concern, proposes a framework for estimating and comparing the stability of different sets of masks and suggests a mechanism based on frequency-based aggregation to form a stable set of predictive features. We evaluate the stability and performance of standard feature selection algorithms and the aggregated approach over four datasets, using a k-nearest neighbour algorithm as the classifier. We show that the aggregation approach produces impressive stability improvements over standard wrapper-based feature selection algorithms. We also show that the classification performance of the feature subsets selected are probably slightly better that those from the standard approach. This is presented in section 5. Before that we provide a short introduction to wrapper and filter-based feature subset selection techniques in the next section. In section 3 we present an example of the instability of the wrapper approach and in section 4 we discuss possible solutions to the instability problem.

2 Wrapper v's Filter-Based Approaches

There are two main classes of feature selection approaches [1]:

- Wrapper-based approaches.
- Filter-based approaches.

The distinction is based on whether the feature selection phase uses the same prediction algorithm as the final classifier (wrapper) or if instead a distinct technique is used to separately generate the feature subset before invoking the classifier (filter). Researchers working with the feature selection problem have found that wrapper approaches generally outperform the filter-based ones [1, 5].

The strength of the Wrapper idea derives from two factors:

1. Features are evaluated in context, i.e. in the presence of other features. Thus dependencies and correlations between features are considered.

2. Since Wrapper approaches focus directly on optimizing the performance of the prediction algorithm, the *bias* of the prediction algorithm is considered.

3 Instability Problems

As mentioned above, it is generally accepted that wrapper-based approaches will outperform filter-based approaches. However, one aspect that needs to be addressed is the stability of such techniques. Stability, in this context, is taken as the property of selecting the same set of features irrespective of variation in the partitioning of the training data. When knowledge discovery is a prime concern, the fact that potentially different sets of features will be produced, given a variation in training and test data, can prove problematic. How do we choose the most suitable features from those presented by different runs of the selection algorithm? Since feature selection is essentially a search problem, the initial position in the search space is highly significant in determining the final subset selected. Two possible solutions are mentioned in the next section of this paper.

3.1 Sequential Selection Wrappers

The most straightforward search strategies are based on stepwise addition or elimination of features. A sequential selection wrapper proceeds by adding or removing features from the current mask to form a new candidate. This is then evaluated using some validation metric (for example leave-one-out accuracy) and if it is superior then it replaces the current best mask and the algorithm continues. The process terminates when no more valid operations (addition/removal) can be performed or if no candidate mask exceeds the performance of its predecessor.

Forward sequential selection, FSS, starts with an empty mask of features and attempts to add one feature on each iteration. If we have added all the features or there is no improvement accrued from adding any further features, the search stops and returns the current set of features. The strictly monotonic addition of features implies that the forward selection search has a maximum search length of N cycles, where N is the number of features present in the dataset being used. The goal of the search is to add only the predictive and relevant features to the mask whilst ignoring the contribution of irrelevant features.

Backward sequential selection, BSS (also referred to as backward elimination), begins with a full mask i.e. including all features and attempts to remove one feature per cycle. The resulting mask, with one less feature than

the current one, is evaluated and the algorithm selects the highest performing candidate (again subject to a specified validation metric). If we reach an empty mask (unlikely) or the subsequent removal of any feature only deteriorates the current performance, we cease the search. Again, due to the monotonic removal of features, the search is guaranteed to terminate by N iterations. The goal of backward elimination is to consider the contribution of all features initially and then try to remove the most irrelevant ones, leaving a smaller and more predictive subset. Since backward elimination starts with all of the features present in the mask, it it more computationally expensive than FSS. However, the fact that initially more features are evaluated in combination normally yields a higher accuracy in classification [5].

A more general search strategy can be formed by considering either the addition or the removal of a feature at each stage in the search. This forms a *hill-climbing* search where we consider the optimum perturbation (toggling of a feature in the mask) on each iteration (an alternative approach would be to accept any perturbation that increases the accuracy.) We initialize the search with a random mask and then proceed to assess the effect of toggling the current status of each feature in the mask. Again we choose the optimum modified mask and we continue. We set a limit on the number of iterations to continue the search for and if this limit is reached, we return the current mask.

3.2 Examples Of Instability

An example of the instability problem can be seen by studying the masks generated by a sequential selection technique running on a large dataset. Figure 1 represents the first twenty masks produced by forward sequential selection in a one hundred run trial on a gene expression profile dataset [2] with 20 features and 997 samples distributed unevenly over 13 classes.

The histogram presented in Figure 2 indicates the frequency of occurrence of each feature in the resulting 200 masks obtained from 200 resampled trials. In this particular experiment, the number of times that a particular feature is selected can vary from 3 times for feature number 11 up to all 200 times for features 5 and 20.

$$\begin{pmatrix}
0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\
1 & 1 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\
0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\
0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\
0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \\
0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\
0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\
0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\
0 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\
0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1
\end{pmatrix}$$

Figure 1: Matrix of feature masks from multiple feature selection trials

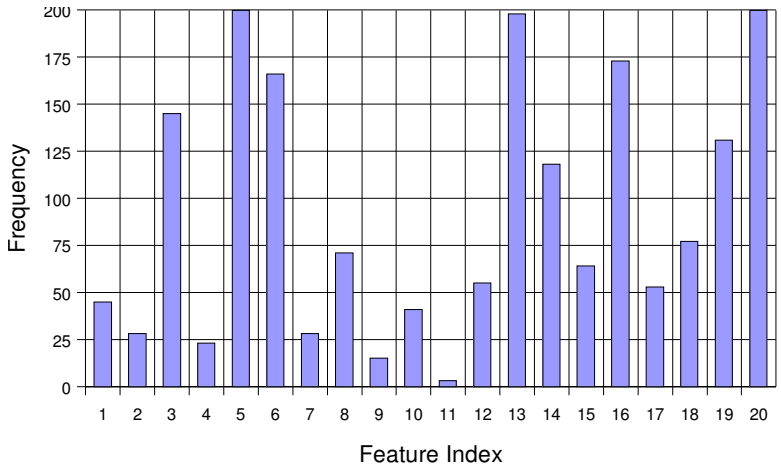

Figure 2: Feature Frequency Histogram - Feature Index vs. Frequency of Occurence

4 Solutions

4.1 Aggregation: The Wrapper-2 Approach

One way of addressing the instability of sequential feature selection techniques is by using statistical aggregation of multiple trials to form a better representation of relevant features - essentially wrapping the wrapper. The process involves running the feature selection algorithm a significant number of times and recording the selected set of features on each run. Since the simple sequential selection schemes (forward selection and backward elimination) are both deterministic, a set of trials with the same training data will yield equivalent masks. This paper addresses the problem of algorithmic instability caused by differences in the training data. In our trials, this instability is generated by using a re-sampled subset of the available training data on each iteration (the amount of data to re-sample is a parameter which can be varied). Alternative selection algorithms that start in different locations in the search space (e.g. random hill climb) can either use this mechanism to introduce variability or they can just choose a different initial mask to begin the search process from.

A procedural description of the approach is as follows:

1. Select the elements of the training data that will be present in this trial by random sampling.

2. Run the feature selection process using the resampled data as training data and either the surplus data (formed from the remainder of the original training data) or a seperate body of data as the test set.

    - The accuracy of a k-nearest neighbour classifier tested on the surplus data is used to gauge the "goodness" of the current feature mask.

3. Store the generated feature subset as the next row in the overall matrix of results.

4. After the trials are finished, use the matrix of collected masks to form the aggregated frequency histogram (AFH) of feature occurences - i.e. the total number of times that a given feature was included in the mask emitted by the feature selection algorithm.

5. Use this histogram to select the most frequent features. Selection can be via:

    - Add features in rank order (as indicated by histogram peaks) to a mask and evaluate this mask on a holdout set - stop introducing additional features when the validation accuracy starts to decrease.

- Add the top T features where T is a configurable parameter - T may be set *a priori* or alternatively may be estimated by considering the histogram profile. A threshold value for sufficient number of occurences could be set by studying the distribution of the sum total of features present in masks. Each feature that exceeded this quota would be included in the final mask.

4.2 Intensive Search Approaches

Alternative search strategies that may offer a solution to the problem of stability include parallel-search strategies and genetic algorithms. Since the problem of instability is primarily caused by the selection search process reporting potentially different local maxima, a search approach which increases the breadth of search, by considering more candidate masks at each decision point in the algorithm, would be likely to improve the stability measure.

Another technique which also exploits parallel searching is that of genetic algorithms (Holland [4]). Genetic algorithms (GA's) work by mimicing the process of natural selection of evolutionary development in a computer algorithm. Instead of explicitly coding the steps required to find the goal state, a population of candidates is created each with a random initial configuration. In the case of feature selection, the initial elements in the population will be random bit-strings of length N, containing either 1 or 0. The highest performing elements (according to a specified objective or goal metric) are then selected for reproduction with an associated chance of either genetic crossover between elements or random mutation of some part of the elements. The process continues for a given duration and the final population is used to form a "best" candidate answer to the search problem.

Because of the initial variability of the population of candidate solutions, and the additional novelty introduced by the application of crossover and mutation operators, a GA-based approach can perform a more thorough search. However, there are also stability problems with GA's and the choice of parameter settings (crossover vs. mutation probabilities, population size, bit-string representation and exact behaviour of operators) can have a huge impact on performance. An additional concern is the heavy run-time computational expense of a genetic algorithm with a significantly-sized population and reasonable number of generations.

5 Evaluation

The evaluation of the aggregated Wrapper-2 approach involved firstly specifying a metric for the assessment of stability and then running a series of trials to gauge the performance of different selection techniques, with and without the aggregation process.

5.1 Definition of Stability Metric $\hat{H}$

To assess the stability of a feature-selection technique, we need to determine how much variation there is in the distribution of features present in the subsets selected under different starting conditions. One measure that could be used is the Hamming distance between two masks. Given a pair of feature masks, m_i and m_j, we define the Hamming distance between them as follows:

$$H(m_i, m_j) = \sum_{k=1}^{N} |m_{ik} - m_{jk}| \tag{1}$$

where N is the total number of features in the dataset and m_{ik} denotes the k-th feature of mask m_i. Each mask can either include feature k, indicated by a 1 at position k in the mask, or omit it in which case the entry is 0.

We can then use this Hamming distance to yield a measure of the overall variation of a set of W feature masks generated by W runs of the feature selection algorithm. First we compute the total Hamming distance, H_t, by summing the individual Hamming distances between each pair of distinct masks:

$$H_t = \sum_{i=1}^{W} \sum_{j=i+1}^{W} H(m_i, m_j) \tag{2}$$

where W is the total number of masks.

This sum, H_t, is computed over P pairs of masks where P is $W(W-1)/2$

By dividing the total distance by this number of pairs, P, we can form the average Hamming distance, $\overline{H}$, as follows:

$$\overline{H} = \frac{H_t \cdot 2}{W \cdot (W-1)} \tag{3}$$

This average Hamming distance depends directly on the length of the masks, N. We can compute $\hat{H}$, the Average Normalized Hamming Distance (ANHD) by dividing the $\overline{H}$ value by N:

$$\hat{H} = \frac{\overline{H}}{N}$$

$$
\begin{aligned}
&= \frac{H_t \cdot 2}{N \cdot W \cdot (W-1)} \\
\hat{H} &= \frac{2}{N \cdot W \cdot (W-1)} \sum_{i=1}^{W} \sum_{j=i+1}^{W} \sum_{k=1}^{N} |m_{ik} - m_{jk}| \qquad (4)
\end{aligned}
$$

The experimental evaluation performed in this study used this measure to assess the stability of the masks generated by a series of selection techniques.

5.2 Expected Value of $\hat{H}$

A test was performed on a randomly generated set of masks to gauge the expected value of $\hat{H}$ for random data. In the test, a set of 100 masks, each of length 20, was created, forming a matrix of values. Each entry in this matrix was randomly set to either 1 or 0 and $\hat{H}$ was calculated. This process was repeated 10000 times with each value of $\hat{H}$ recorded. The mean and standard deviation of these values are recorded below:

Maximum value	0.503929
Minimum value	0.491384
Mean value	0.500001659
Standard Deviation	0.001598177

Table 1: $\hat{H}$ Attributes for a randomly generated set of masks

For the case of a set of alternating full and empty masks, the value of $\hat{H}$ will be:

$$
\begin{aligned}
\hat{H}_{0\to1} &= \frac{NW}{2 \cdot N \cdot (W-1)} \\
&= \frac{W}{2 \cdot (W-1)} \qquad (5) \\
\lim_{W\to\infty} \hat{H}_{0\to1} &= 1/2 \qquad (6)
\end{aligned}
$$

6 Results

6.1 Datasets Used for Evaluation

Four datasets were used to assess the performance of the aggregated wrapper approach:

- Saccharomyces Yeast gene expression dataset with 997 samples, 20 features and 13 classes - from Ideker et al. [2]
- Leukeamia dataset with 72 samples, 50 features, 25 positive & 47 negative examplars - from Golub et al. [3]
- Hand recognition dataset with 63 samples, 13 features and 7 classes (each with 9 examplars) - locally gathered
- Ear recognition dataset with 126 samples, 10 features and 7 classes (each with 18 examplars) - locally gathered

The largest dataset (Saccharomyces expression profile) formed the main testbed for analysis. The hand and ear data were collected locally for use in a separate feature selection project. The values of the features indicate biometric distances. Figure 3 indicates what the feature values in the Hand and Ear datasets represent.

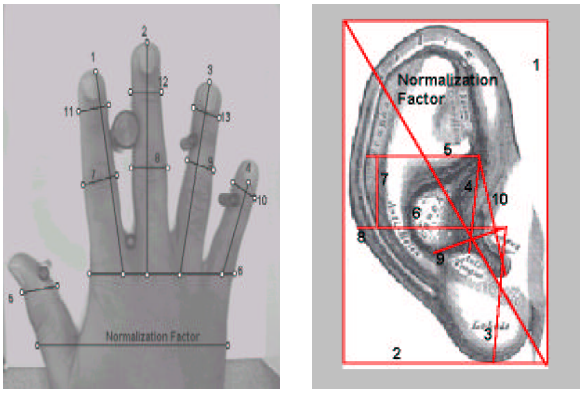

Figure 3: Hand & Ear Data Description

6.2 Procedure

The average normalised Hamming distance measure of mask difference, $\hat{H}$, was calculated over a number of trials with varying parameters as follows:

- The number of re-sampling trials performed, W.

R	Type	$\overline{\overline{H}}$	$\hat{H}$
50%	FSS	6.16222	0.308111
60%	FSS	5.99758	0.299879
70%	FSS	5.40768	0.270384
80%	FSS	5.63212	0.281606
50%	BSS	5.27354	0.263677
60%	BSS	4.36687	0.218343
70%	BSS	4.41273	0.220636
80%	BSS	4.02889	0.201444
50%	RHC	8.44061	0.422030
60%	RHC	8.35172	0.417586
70%	RHC	8.09667	0.404838
80%	RHC	7.89677	0.394838

Table 2: Saccharomyces Dataset Stability - $W = 100$, $k = 3$

- The number of neighbours used to form a classification, k.
- The percentage of data elements re-sampled, R, for each trial (compared to training dataset size).
- The dataset chosen (see section 6.1).
- The feature selection scheme i.e. forward sequential selection (FSS), backward sequential selection / elimination (BSS) or a random hill-climbing search (RHC).

Table 2 records the stability measure of each standard feature selection technique on the Saccharomyces dataset. It is clear that there is more instability in the process as the overlap in the datasets is reduced. Figures 4, 5 and 6 depict the corresponding histograms for the different search techniques used (BSS, FSS, RHC).

Table 3 records the stability measures for multiple experiments using the four datasets. The table columns represent the dataset name, the number of trials performed in the aggregation process, the number of features in the dataset and the type of feature selection used. Stability measures were calculated over the entire set of masks and between aggregated subsets - the final two column records the number of aggregated masks compared (A) and the number of features taken (F). The A masks used to form the aggregated masks were taken in order from the W available trials - a value of A that divides evenly into W was taken.

				Standard		Aggregated			
Dataset	W	N	Type	$\overline{H}$	$\hat{H}$	$\overline{H}$	$\hat{H}$	A	F
Sacc.	200	20	FSS	5.581	0.279	2.844	0.142	20	10
Sacc.	100	20	BSS	4.029	0.201	4.733	0.237	10	10
Sacc.	100	20	BSS	4.029	0.201	3.133	0.157	10	15
Sacc.	100	20	FSS	5.632	0.282	3.244	0.162	10	10
Sacc.	100	20	FSS	5.632	0.282	2.4	0.12	20	10
Sacc.	100	20	RHC	7.897	0.395	5.978	0.299	10	10
Leuk.	200	50	BSS	11.769	0.235	6.578	0.132	20	40
Leuk.	200	50	FSS	7.925	0.158	4.778	0.096	20	40
Leuk.	200	50	RHC	24.842	0.497	12.556	0.251	20	40
Hand	200	13	BSS	3.873	0.298	0.4	0.031	20	4
Hand	200	13	FSS	2.961	0.228	0.867	0.067	20	6
Hand	200	13	RHC	4.085	0.314	1.956	0.150	20	5
Ear	200	10	BSS	1.756	0.176	0.556	0.056	20	5
Ear	200	10	FSS	1.940	0.194	0	0	20	5
Ear	200	10	RHC	2.451	0.245	1.089	0.109	20	5

Table 3: Overall Stability Results

For this table, a value of $R = 80\%$ was used. These results show improvements due to aggregation that are impressive provided enough masks are used. For Saccharomyces it is clear that 10 is not enough but 20 roughly halves the Hamming distance measure of instability.

Table 4 indicates the average normalised Hamming distance from the aggregation of each of the above runs for the Saccharomyces dataset, with parameter F indicating how many features were taken (in rank order).

6.3 Prediction Accuracy

The prediction performance of the aggregated mask was compared with the average performance of a set of masks generated. The evaluation involved a set of 10 trials, where 80% of the training data was used as the case base and the remaining 20% was set aside as the holdout test set. A k-nearest neighbour algorithm with a value of $k = 3$ was used to predict the class of the holdout set with the feature mask determining which of the features were included in the calculation of the Euclidean distance metric between training examplars and the query case. In the tables that follow, W represents the number of masks used (number of trials) and R indicates the amount of training data resampled on each trial to form feature subset. The Type column specifies the feature selection mechanism used (FSS/BSS/RHC).The Bits per mask column represents the average number of features present in each mask over the total aggregation.

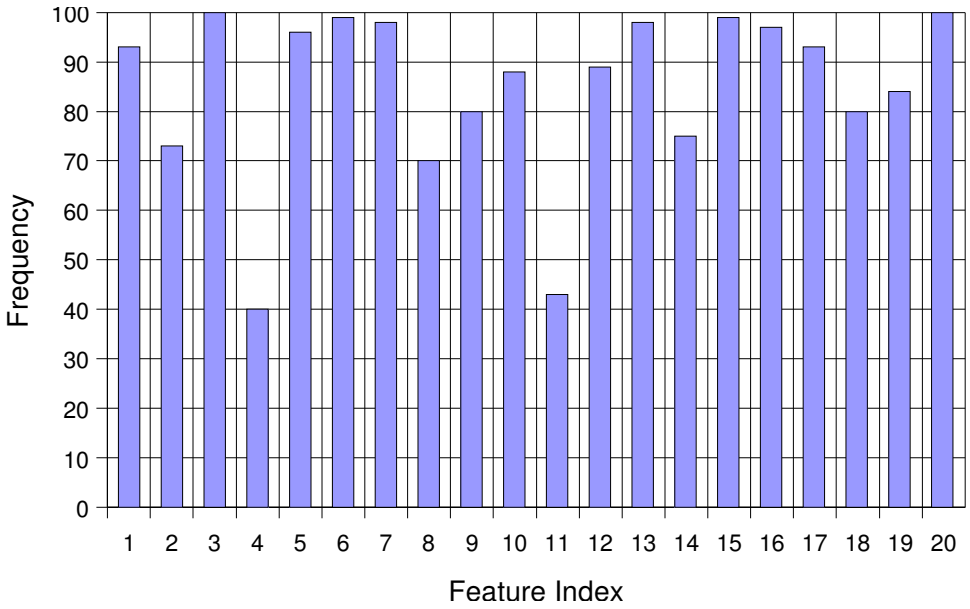

Figure 4: Feature Frequency Histogram from 100 Aggregated Trials on Saccharomyces dataset with $R = 80\%$, Backward Sequential Selection

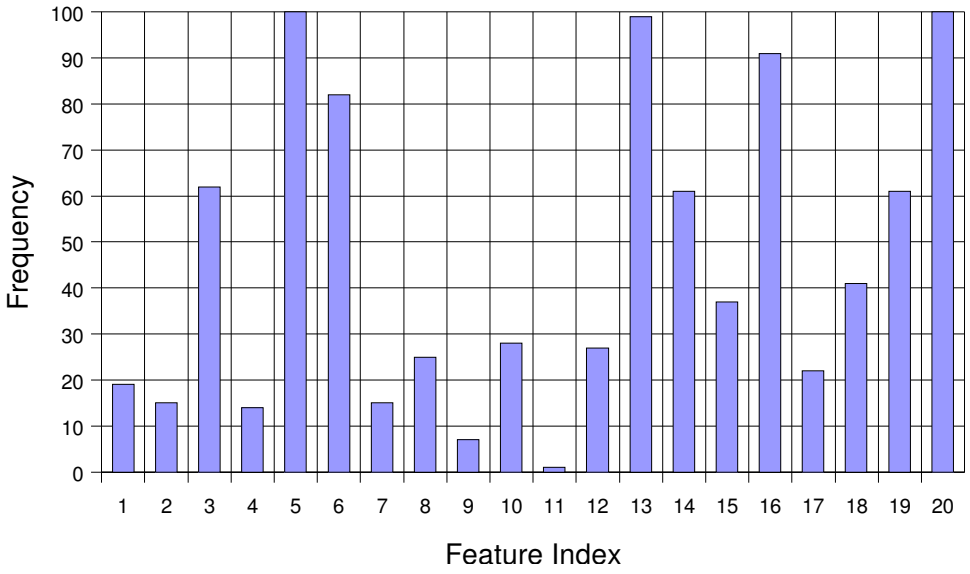

Figure 5: Feature Frequency Histogram from 100 Aggregated Trials on Saccharomyces dataset with $R = 80\%$, Forward Sequential Selection

F	$\overline{H}$	$\hat{H}$
1	1.12121	0.056061
2	1.60606	0.080303
3	2.42424	0.121212
4	2.87879	0.143939
5	2.78788	0.139394
6	3	0.15
7	2.63636	0.131818
8	2.75758	0.137879
9	3.27273	0.163636
10	3.27273	0.163636
11	3.72727	0.186264
12	4.09091	0.204545
13	3.81818	0.190909
14	3.75758	0.187879
15	2.72727	0.136364
16	2.87879	0.143939
17	2.54545	0.127273
18	1.60606	0.080303
19	0.33333	0.016667
20	0	0

Table 4: Inter-Trial Aggregated Stability - This table shows the values of $\hat{H}$ over the top F features from each of the 100-element trials performed on the Saccharomyces dataset

Type	Accuracy	Std. Dev.	Mean Bits per Mask
FSS	84.05%	2.24%	9.01
BSS	85.36%	2.25%	16.95
RHC	84.50%	2.27%	12.51

Table 5: Average Mask Prediction Performance - Saccharomyces dataset with $W = 100$, $R = 80\%$

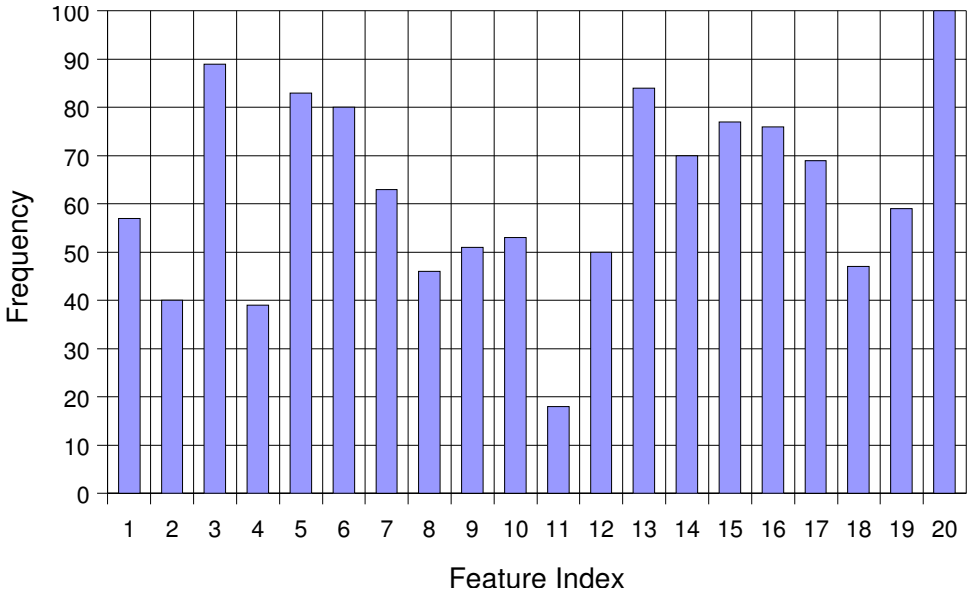

Figure 6: Feature Frequency Histogram from 100 Aggregated Trials on Saccharomyces dataset with $R = 80\%$, Random Hill-climbing

Table 6 shows the results of using the aggregated feature histogram to form the mask. The rows indicate the performance when the specified number of features, added in rank order, were included in the mask. Again the evaluation was carried out using 10-fold resampling of the training data with 80/20 split of training to test data. The average accuracy over the 10 trials along with the standard deviation is included. As can be seen from the table, both the aggregated mask performance for backward selection and the random hill-climb search is maximised when 14 features are included in the mask and 17 features represents the optimium for the forward sequential aggregation. Tables 7 and 8 show the equivalent results for the Ear dataset, tables 9 and 10 for the Hand dataset and tables 11 and 12 represent the Leukaemia dataset. Figures 7, 8 and 9 show the corresponding histograms for the Ear, Hand and Leukaemia datasets respectively.

	FSS		BSS		RHC	
#Features	Acc.	Std. Dev	Acc.	Std. Dev	Acc.	Std. Dev
1	34.05%	3.91%	22.40%	1.87%	31.00%	3.21%
2	51.50%	1.39%	38.55%	1.85%	38.15%	2.95%
3	77.15%	1.62%	64.35%	3.10%	71.15%	2.17%
4	78.80%	2.00%	72.80%	2.83%	76.65%	1.33%
5	81.70%	1.46%	75.60%	3.91%	79.20%	3.58%
6	80.65%	2.06%	77.50%	3.80%	81.00%	3.68%
7	85.50%	3.70%	79.90%	3.06%	81.75%	3.43%
8	83.90%	1.66%	82.45%	2.89%	85.00%	2.90%
9	83.35%	2.90%	83.70%	2.98%	83.25%	2.36%
10	84.60%	2.61%	82.80%	3.26%	85.35%	2.26%
11	84.15%	1.94%	83.95%	1.85%	85.50%	1.94%
12	84.75%	1.72%	86.00%	2.70%	84.60%	2.86%
13	84.15%	3.29%	86.05%	2.07%	85.90%	2.01%
14	84.60%	2.29%	86.40%	1.47%	86.80%	1.89%
15	85.80%	3.07%	85.65%	2.19%	86.60%	1.66%
16	85.95%	2.11%	85.85%	1.73%	86.75%	2.46%
17	86.70%	2.25%	85.80%	1.70%	86.55%	2.31%
18	83.85%	2.01%	84.35%	3.25%	85.95%	2.27%
19	85.60%	1.56%	84.55%	2.11%	85.40%	2.38%
20	84.70%	2.70%	83.75%	3.96%	85.50%	2.30%

Table 6: Aggregated Mask Prediction Performance - Sacc. Dataset

Type	Accuracy	Std. Dev.	Mean Bits per Mask
FSS	76.81%	7.40%	4.47
BSS	77.70%	7.07%	4.69
RHC	75.95%	7.26%	4.33

Table 7: Average Mask Prediction Performance - Ear dataset with $W = 200$, $R = 80\%$

	FSS		BSS		RHC	
#Features	Acc.	Std. Dev	Acc.	Std. Dev	Acc.	Std. Dev
1	30.77%	4.05%	37.69%	9.39%	36.92%	4.51%
2	50.38%	6.14%	64.23%	5.75%	59.62%	6.08%
3	75.38%	9.28%	79.62%	8.51%	80.77%	9.07%
4	76.15%	9.03%	80.77%	7.02%	79.23%	7.52%
5	81.54%	5.06%	80.00%	8.85%	78.08%	7.49%
6	75.38%	4.87%	73.46%	8.78%	75.77%	7.03%
7	67.31%	9.64%	75.38%	7.94%	63.46%	7.31%
8	58.85%	8.12%	60.00%	9.46%	61.15%	8.78%
9	52.31%	6.59%	52.31%	7.30%	54.62%	5.06%
10	43.08%	7.21%	46.92%	7.21%	42.69%	5.86%

Table 8: Aggregated Mask Prediction Performance - Ear Dataset

Type	Accuracy	Std. Dev.	Mean Bits per Mask
FSS	92.61%	6.16%	6.19
BSS	87.98%	7.92%	7.21
RHC	90.12%	7.05%	6.59

Table 9: Average Mask Prediction Performance - Hand dataset with $W = 200$, $R = 80\%$

	FSS		BSS		RHC	
#Features	Acc.	Std. Dev	Acc.	Std. Dev	Acc.	Std. Dev
1	39.23%	6.74%	42.31%	6.54%	36.92%	7.07%
2	63.85%	12.59%	65.38%	8.31%	66.15%	9.73%
3	73.08%	7.48%	82.31%	7.30%	69.23%	13.07%
4	93.08%	5.68%	94.62%	7.30%	94.62%	6.33%
5	94.62%	5.19%	92.31%	7.25%	96.15%	5.44%
6	92.31%	7.25%	97.69%	5.19%	96.15%	5.44%
7	93.08%	6.74%	88.46%	5.44%	93.08%	5.68%
8	90.00%	6.33%	86.92%	8.15%	87.69%	8.27%
9	90.77%	6.07%	80.77%	11.61%	92.31%	7.25%
10	82.31%	12.59%	85.38%	12.27%	84.62%	12.56%
11	76.15%	14.71%	80.77%	10.42%	84.62%	8.88%
12	80.00%	9.73%	86.92%	10.29%	74.62%	11.50%
13	40.00%	13.47%	39.23%	12.79%	33.08%	8.15%

Table 10: Aggregated Mask Prediction Performance - Ear Dataset

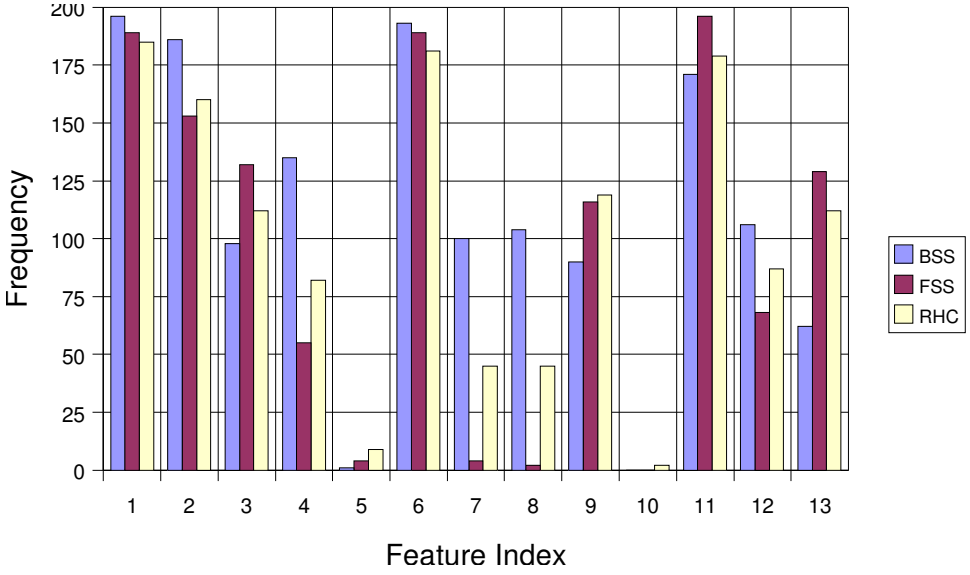

Figure 7: Feature Frequency Histogram from 200 Aggregated Trials on Hand dataset with $R = 80\%$

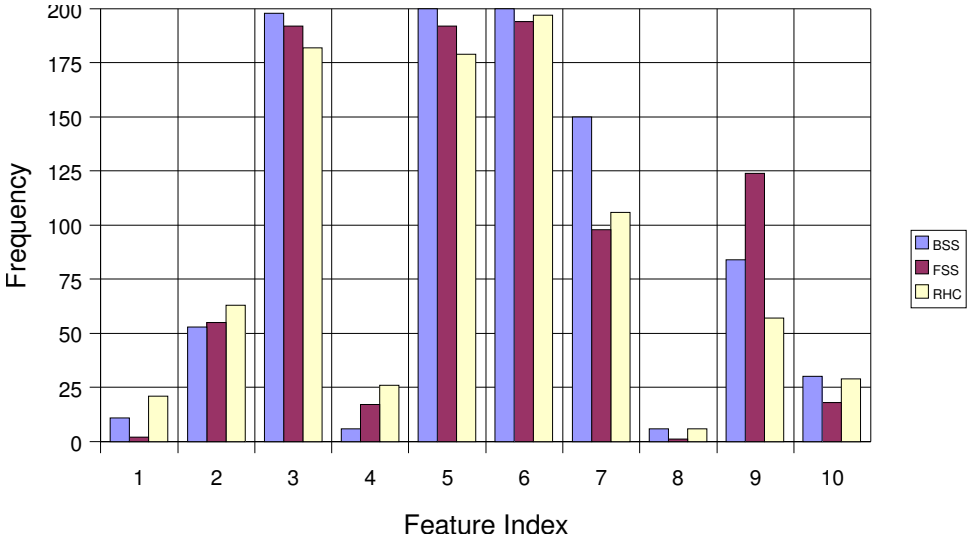

Figure 8: Feature Frequency Histogram from 200 Aggregated Trials on Ear dataset with $R = 80\%$

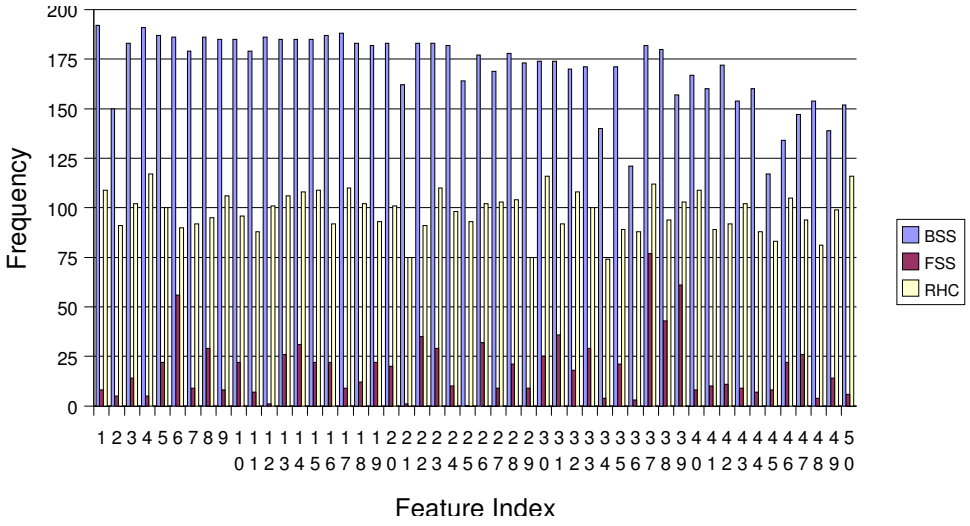

Figure 9: Feature Frequency Histogram from 200 Aggregated Trials on Leukaemia dataset with $R = 80\%$

Type	Accuracy	Std. Dev.	Mean Bits per Mask
FSS	79.06%	8.78%	4.69
BSS	79.28%	8.77%	42.67
RHC	79.50%	8.63%	24.47

Table 11: Average Mask Prediction Performance - Leukaemia dataset with $W = 200$, $R = 80\%$

	FSS		BSS		RHC	
#Features	Acc.	Std. Dev	Acc.	Std. Dev	Acc.	Std. Dev
1	81.33%	9.32%	60.00%	9.43%	60.67%	12.75%
2	80.00%	7.70%	62.00%	8.92%	70.67%	8.43%
3	74.67%	6.13%	77.33%	10.98%	74.00%	11.09%
4	83.33%	11.44%	70.67%	9.00%	73.33%	8.89%
5	83.33%	6.48%	73.33%	9.94%	76.00%	9.53%
6	84.67%	5.49%	74.00%	9.14%	76.00%	7.17%
7	80.00%	7.03%	73.33%	7.70%	75.33%	12.19%
8	82.67%	7.83%	78.67%	8.78%	71.33%	7.73%
9	*86.00%*	8.58%	74.67%	6.13%	76.67%	10.54%
10	75.33%	8.34%	75.33%	10.45%	78.67%	6.89%
11	83.33%	7.20%	73.33%	10.42%	80.00%	8.31%
12	82.00%	8.92%	73.33%	12.96%	79.33%	6.63%
13	77.33%	7.83%	72.00%	12.09%	79.33%	7.98%
14	76.67%	9.56%	82.67%	7.17%	80.67%	8.58%
15	77.33%	7.17%	78.67%	10.33%	77.33%	13.41%
16	78.00%	9.96%	78.67%	6.89%	76.67%	9.03%
17	81.33%	5.26%	84.67%	6.32%	74.00%	7.98%
18	76.67%	7.86%	80.00%	8.31%	82.00%	9.45%
19	79.33%	7.98%	*85.33%*	4.22%	81.33%	8.78%
20	76.00%	8.43%	79.33%	10.63%	82.00%	9.96%
21	78.00%	8.92%	80.67%	9.66%	74.67%	8.78%
22	80.00%	8.31%	78.67%	7.57%	76.00%	7.17%
23	83.33%	9.03%	83.33%	9.03%	75.33%	9.96%
24	76.67%	10.06%	80.67%	9.14%	77.33%	12.65%
25	81.33%	6.89%	85.33%	8.20%	83.33%	7.20%
26	76.00%	7.17%	83.33%	3.51%	80.67%	9.14%
27	76.67%	10.06%	80.67%	10.63%	80.00%	7.03%
28	80.00%	8.31%	83.33%	6.48%	80.00%	4.44%
29	79.33%	4.92%	80.67%	8.58%	80.67%	8.58%
30	76.67%	9.56%	79.33%	7.34%	86.00%	5.84%
31	76.00%	7.17%	84.00%	10.52%	81.33%	5.26%
32	78.00%	12.19%	81.33%	6.13%	81.33%	6.89%
33	78.67%	8.78%	82.00%	7.73%	79.33%	9.66%
34	79.33%	6.63%	80.67%	9.66%	84.00%	10.52%
35	82.00%	6.32%	82.00%	6.32%	84.00%	7.83%
36	86.00%	7.34%	79.33%	9.66%	84.67%	7.73%
37	84.00%	6.44%	82.67%	9.00%	78.67%	8.20%
38	80.67%	6.63%	75.33%	7.06%	80.67%	11.09%
39	80.00%	8.89%	80.00%	8.31%	80.00%	7.03%
40	80.00%	8.89%	82.67%	9.00%	82.67%	6.44%
41	78.00%	11.78%	74.67%	10.33%	84.67%	7.06%
42	84.67%	7.73%	80.00%	7.70%	79.33%	5.84%
43	72.00%	10.80%	82.00%	10.91%	80.00%	8.31%
44	84.67%	7.73%	77.33%	10.52%	78.67%	12.49%
45	76.67%	9.03%	82.00%	8.34%	84.67%	5.49%
46	*86.00%*	7.98%	72.67%	10.16%	81.33%	8.20%
47	77.33%	7.17%	78.67%	9.84%	78.00%	8.34%
48	76.67%	10.54%	77.33%	8.43%	78.67%	8.78%
49	74.00%	9.66%	75.33%	5.49%	80.00%	8.31%
50	74.67%	12.09%	74.67%	10.80%	76.67%	11.86%

Table 12: Aggregated Mask Prediction Performance - Leukaemia Dataset

7 Conclusion

As evidenced by the histograms generated by running the feature selection techniques over many trials, the same set of features is not always selected. The choice of relevant features, as gauged by the selection process, depends strongly on the data samples present in the training set. Different partitioning of the training data will lead to a different set of features being identified as being the most predictive of a target class.

The aggregation process specified in this paper suggests a way of tackling this instability issue. Ensembling techniques have proven superior to single classifiers by countering the potential inaccuracy of a single trial with the more balanced and stable viewpoint of a virtual "committee of experts". Aggregation of feature subsets is a form of ensembling where the aim is to ensure that only the features selected in a significant number of trials are proposed as most likely indicators of sample class. We have found that this aggregation approach does improve the stability of the feature selection process and the resulting classifiers are not worse and probably better that those produced from standard wrapper approaches. From a knowledge discovery perspective the improvements in stability are the most significant.

By studying the aggregated histogram, we can gain some insight into the data domain under study. The dominant peaks in the histogram signify features that occur most often in the generated feature masks and are likely to be the most relevant features in the dataset. Likewise, troughs with values close to zero are likely to indicate features that are not strongly predictive and are good candidates for pruning. The overall histogram profile is a potential indicator of the general quality of the features available in the data domain. An approximately smooth histogram with few spikes and troughs is indicative of a reasonable set of features, each one to some extent plausible for the purposes of classification. On the other hand, a highly irregular histogram with large swings between high and low peaks suggests that some of the features are significantly different when used in a classification task.

In this instance, a comparison can be drawn between the histograms for the Hand and Ear datasets and that of the Leukaemia data. The Hand and Ear datasets contain mostly strong peaks and very small troughs, indicating an easily-identifible set of relevant features and some irrelevant ones. On the other hand, the Leukaemia dataset has a pretty level histogram, indicating that most features are predictive. By studying the average number of bits in the resulting masks from the Leukaemia data, we can also see that the forward selection approach has terminated quite soon, after adding a small number of features. This would seem to suggest that although the majority of the features are relevant, there is a large degree of redundancy in the dataset.

Another use would be to compare histograms between population samples to investigate changes in the source data - if all the sets of training

data available are representative of the underlying data distribution then it would be reasonable to expect that the histogram shape should remain approximately invaraint across multiple feature selection experiments. Other information can be extracted from the matrix of masks procuced over the trials. For example, the covariance and correlation statistical measures could be used to establish pairwise dependence of features or if the predictive capability of one feature is comprised by the addition of another.

Finally, in section 4 we proposed more intensive search as an alternative to the aggregation approach evaluated here. We have done some preliminary evaluation with beam search and with genetic algorithms and surprisingly are getting quite unstable solutions. This evaluation is still at an early stage however. An issue that may be significant is the occurence of several local minima with very similar fitness - this would adversely affect stability.

References

[1] Kohavi, R., John G. H., *Wrappers for Feature Subset Selection*, Artificial Intelligence 1997, Vol. 97, No. 1-2, p273-324.

[2] Ideker T., Thorsson V., Ranish J.A., Christmas R., Buhler J., Eng J.K., Bumgarner R., Goodlett D.R., Aebersol R., Hood L. *Integrated genomic and proteomic analyses of a systematically perturbated metabolic network*, Science 2001; 292:929-933.

[3] Golub T.R., Slonim D.K., Tamayo P., Huard C., Gassenbeck M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A., Bloomfield C.D., Lander E.S.. *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.* Science 1999; 286:531-537.

[4] Holland, J. H., *Adaption in Natural and Artificial Systems*, University of Michigan Press, 1975.

[5] Aha, D. W. & Bankert, R. L. (1994), *Feature selection for case-based classification of cloud types: An empirical comparison*, in Working Notes of the AAAI-94 Workshop on Case-based Reasoning, pp 106-112.