Electronic Journal of SADIO

http://www.dc.uba.ar/sadio/ejs/

vol. 8, no. 1, pp. 12–24 (2008)

A hybrid wrapper/filter approach for feature subset selection

Ronaldo C. Prati *Gustavo E. A. P. A. Batista* *Maria Carolina Monard*

Department of Computer Science (SCC)
Institute of Mathematics and Computer Science (ICMC)
University of São Paulo at São Carlos (USP/São Carlos)
P.O. Box 668, Zip Code 13561-970
São Carlos, SP, Brazil
{prati,gbatista,mcmonard}@icmc.usp.br

Abstract

This work presents a hybrid wrapper/filter algorithm for feature subset selection that can use a combination of several quality criteria measures to rank the set of features of a dataset. These ranked features are used to prune the search space of subsets of possible features such that the number of times the wrapper executes the learning algorithm for a dataset with M features is reduced to $O(M)$ runs. Experimental results using 14 datasets show that, for most of the datasets, the AUC assessed using the reduced feature set is comparable to the AUC of the model constructed using all the features. Furthermore, the algorithm archieved a good reduction in the number of features.

Keywords. Feature Subset Selection, Wrapper, Filter, Machine Learning, Data Mining

This research was supported by the Brazilian Research Councils CNPq, FAPESP and Fundação Parque Tecnológico Itaipu — FPTI/Br

1 Introduction

Feature subset selection is a prevalent problem in Machine Learning and Data Mining, in particular for application areas in which datasets have a great number of features. In such cases, feature subset selection plays an important role and is often applied as a pre-processing step to reduce the number of features given to a learning algorithm. Feature subset selection (FSS) is a search problem, whereby each search state specifies a subset of possible features of the task at hand. Exhaustive evaluation of all feature subsets is generally intractable, and heuristic methods are often used.

The aim of feature selection is three-fold Guyon and Elisseeff (2003): (1) improving the prediction performance of models, (2) providing smaller and more cost-effective models and (3) enhancing understanding of the underlying concept which generated the data. Some methods for feature selection give more emphasis on one aspect than another, although improving prediction performance is by far the most studied.

In this work, we propose a hybrid wrapper/filter FSS algorithm for supervised classifications tasks that uses a combination of several quality criteria measures to rank the set features of a dataset, although this algorithm can also be applied whenever various methods that rank features using different quality criteria are available. In order to validate our algorithm, we carried out an empirical evaluation on 14 datasets from UCI Newman et al. (1998). Results show that, for most of the datasets, this hybrid approach is able to keep the AUC comparable to the AUC of the model constructed using all the features, as well as achieving a good reduction in the number of features.

This work is organised as follows: Section 2 describes some feature selection methods. Section 3 presents the hybrid wrapper/filter algorithm. Section 4 presents a way to construct a new measure for feature quality criteria based on the combination of other measures. Section 5 presents the experimental results and Section 6 concludes.

2 Feature Subset Selection

Supervised learning algorithms take as input a training set of N classified instances $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}$ for some unknown function $y = f(\mathbf{x})$, where the $\mathbf{x}_i$ values are typically vectors of the form $(x_{i1}, x_{i2}, ..., x_{iM})$, and x_{ij} denotes the value of the j-th feature (or attribute) X_j of $\mathbf{x}_i$. For classification purposes, the y values are drawn from a discrete set of N_{Cl} classes, *i.e.* $y \in \{C_1, C_2, ..., C_{N_{Cl}}\}$. From that training set, a learning algorithm induces a *classifier*, which is a hypothesis $\mathbf{h}$ about the true unknown function f. Although a greater number M of features should provide a better discriminating power, this is not the case for irrelevant and/or redundant features, which frequently confuse the learning algorithm.

FSS can be formalised as follows Yu and Liu (2004): let $X' \subset X$ be a subset of features and $f'(\mathbf{x}')$ the value associated to instances described by

features in X'. The objective of FSS is to select a minimum feature subset X' such that $\mathbf{P}(C|Y = f'(\mathbf{x}')) \approx \mathbf{P}(C|y = f(\mathbf{x}))$, where $\mathbf{P}(C|Y = f'(\mathbf{x}')))$ and $\mathbf{P}(C|y = f(\mathbf{x}))$ are the probability distributions of the N_{C_l} possible classes given the feature values in X' and X, respectively. This minimum subset X' is named the *optimal* subset. In this context, the FSS problem can be characterised under two aspects: how features are evaluated, and how the feature selection algorithm and the learning algorithm that use the selected features will interact.

Ideally, the best subset contains the least number of features that most contribute for building the model, and we can discard the remaining, unimportant features. As aforementioned, feature subset selection is generally cast as a search problem, and several methods were developed to find the "optimal" subset. Unfortunately, trying all possible subsets leads to a combinatorial explosion as the number of candidate features grows. Two main approaches are generally used to avoid the exhaustive search Vafaie and Jong (1993). The first one is to construct ad-hoc strategies to prune the feature space to a manageable size. The second one uses generic heuristics (primarily greedy forward or backward hill-climbing algorithms) when domain knowledge is costly or unavailable. As the usage of ad-hoc approaches are constrained to a specific domain, they are of less interest than general methods of feature subset selection.

Algorithms for feature subset selection which use generic heuristics may be categorised as wrapper, filter and embedded approaches. The wrapper approach Kohavi and John (1997) takes into account the accuracy of the classifier induced with a feature subset as a heuristic to guide the selection of features. When testing to retain or discard a feature, depending on whether the search is forwards or backwards, such a feature is retained if it improves the accuracy of the model, otherwise it is discarded. On the other hand, the filter approach Duch (2006) uses heuristics based on training data characteristics to select features. The most common approach is to rank features using some quality criterion, and filter out the lower ranked features. Finally, the embedded approach Lal et al. (2006) is an indirect method which selects features through a learning algorithm that internally implements some feature selection method. Thus, the embedded approach is intrinsic to some learning algorithms, and consequently it is restricted to specific algorithms that were projected with this characteristic.

Another approach to reduce the number of features is the elimination of redundant features. In a general way, selection of relevant features and elimination of redundant features are orthogonal approaches: the former aims to find a feature subset with high correlation between features and the class (relevance); and the latter aims to reduce the correlation among these features. For instance, if various copies of a highly relevant feature are present in a data set, feature selection using an importance criterion will select all of them, while feature redundancy will eliminate all but one feature.

3 A hybrid wrapper/filter approach for FSS

As mentioned in the previous section, wrapper and filter are quite general procedures for feature subset selection. The filter approach is usually computationally less intensive, although the wrapper approach frequently produces the best results Appice et al. (2004). Moreover, the wrapper approach has a greater computational complexity, and the results are optimised to the learning algorithm used as a wrapper. On the other hand, filters are very flexible methods since any learning algorithm can use the selected features. However, unlike the wrapper approach, there is not an established procedure to decide how many features should be selected, and this number is left to the user as a parameter (either by explicitly defining the number of features or specifying some arbitrarily chosen threshold in the quality measure). In this work, we propose a hybrid wrapper/filter algorithm aiming to explore the qualities of both strategies and try to overcome some of their deficiencies.

In the standard wrapper approach, in order to select the feature to be removed, for each feature in the feature set a classifier is induced using all but the corresponding feature, and a performance metric is calculated (*e.g.* AUC or error rate). The feature which leads to the lowest performance is the feature candidate for removal. Differently from this standard wrapper approach, where the process of selecting a feature to be removed is based on the same learning algorithm that the wrapper is based on, we proposed to use a measure to rank all the features. At each iteration, the lowest ranked feature is the candidate for removal. Thus, our approach can be understood as a hybrid solution between a filter and a wrapper approach.

The proposed wrapper uses a backward best-first search strategy and was constructed as follows: the relevant feature set is initialised with all dataset features. At each iteration, one feature is removed from the feature set until a stop criterion is met. The removed feature is the one with the lowest score given by a filter measure. The Naïve Bayes learning algorithm is applied to the reduced feature set and if the AUC assessed in this reduced set is not lower than 95% of the AUC calculated using all the dataset features, the removed feature is discarded and the process is repeated with the remaining features. Otherwise, the feature is considered relevant and is included back into the relevant feature set. In this case, the search process terminates and the relevant feature set is returned.

The purpose of this hybrid approach is two-fold. First, this modification saves computational time. As each state in the search space specifies a subset of possible features, the size of the search space for M features is 2^M. Thus, even using a simple best-first search strategy, the number of runs of the learning algorithm used to guide the search in the wrapper is $O(M^2)$. As in our approach the features are ranked before the wrapper execution, then this number of runs is $O(M)$. Second, this approach overcomes the problem of previously deciding how many features should be selected (or which threshold should be applied) by the filter approach.

Despite these advantages, this hybrid approach introduces a bias due to the

choice of the measure used to rank the features. The filter approach usually uses some measure as a heuristic to guide the selection process, and numerous measures to guide this selection are proposed in the literature. As stated earlier, this approach is widely used since it is less computationally expensive and also because it is independent of the learning system. However, since the search is not exact and the measures used as a heuristic focus on some data characteristic, it is not possible to know beforehand which measure is the most appropriate for a given domain. To ease this problem, we propose to use a combination of measures as described next.

4 Combining FSS methods using rankings

The filter approach uses a measure to quantify some data characteristic in the available set of examples, and uses this information as a heuristic to guide the search for the best feature subset. However, each heuristic emphasasis one data characteristic, and the search might be biased by the chosen heuristic. In this work, we propose using different measures to quantify data characteristics which are finally combined into one measure using a simple ranking aggregation function.

A problem that frequently arises when combining different measures is that each of them might use a different scale. For instance, measure A assigns a rating in the interval $[0, 1]$, while another measure B assigns a rating in the interval $[-1, 1]$. Furthermore, even though both measures use the same absolute scale, *i.e.*, both of them assign a rating in the interval $[0, 1]$, relative scales might be different. In other words, a score of 0.8 for measure A might have a different weight if compared with a score of 0.8 for measure B.

An approach to overcome these scale issues is to consider only the rank given by different heuristics to each feature. The proposed method, CFSS (for Combined Filter Subset Selection), is based on this approach, and works as follows:

- various different measures are used to evaluate each feature;
- the score given by each measure is used to rank the features. The best evaluated feature is ranked first, the next is ranked second, and so forth. In case of ties, *i.e.*, two features obtaining the same score, a mean rank is assigned. For instance, if there is a tie between two features at the third rank position, these two features are assigned to the "mean rank" 3.5 ($\frac{3+4}{2}$).
- any set of measures of interest can be combined into a final rank by computing the mean rank of each feature, instead of considering the scores of each feature.

For instance, consider the problem of selecting two features given a set of five features (A, B, C, D and E) and three different measures. The first measure

Table 1: Description of the datasets used in the experiments

Dataset name	Number of features	Number of instances	Number of classes	Maj. Class (%)
anneal	38 (32,6)	898	2	76,2
audiology	69 (69,0)	226	24	25,2
coil2000	85 (85,0)	9822	2	94,0
crx	15 (9,6)	690	2	55,5
ionosphere	32 (0,32)	351	2	64,1
lymphography	18 (18,0)	148	4	57,4
mushroom	22 (22,0)	8124	7	51,8
primary-tumor	17 (17,0)	339	21	24,8
promoters	57 (57,0)	106	2	50,0
soybean-large	35 (35,0)	683	19	13,0
vehicle	18 (0,18)	846	4	75,0
voting	16 (16,0)	435	2	61,4
wdbc	20 (0,20)	569	3	62,7
zoo	16 (16,0)	101	7	40,6

assigns the scores (0.9, 0.3, 0.8, 0.5 and 0.2), the second measure assigns (0.7, -0.6, 0.1, 0.4, and -0.3), and the third measure assigns (0.9, 0.1, 0.5, 0.6, 0.7). Thus, the three measures rank the five features as follows: (A, C, D, B, E) for the first, (A, D, C, E, B) for the second, and (A, E, D, C, B) for the third measure. Considering the two top ranked features, the first measure would select features A and C, the second measure A and D, and the third measure A and E. Therefore, all three measures selected feature A. However, they disagree regarding which feature should be selected as second. The mean rank position for each feature is $A = 1(\frac{1+1+1}{3})$, $B = 4.6(\frac{4+5+5}{3})$, $C = 3(\frac{2+3+4}{3})$, $D = 2.6(\frac{3+2+3}{3})$, and $E = 3.6(\frac{5+4+2}{3})$. Thus, the final ranking for the proposed approach is (A, D, C, E and B) and the selected features would be A and D. It can be observed that feature D was ranked second for the second measure and was ranked third for the first and third measures.

5 Experimental Evaluation

To validate our proposal, we carried out a series of experiments using 14 datasets from UCI Newman et al. (1998). We selected datasets having at least 15 features and without unknown feature values (to avoid a possible contamination of such unknown values in the results analysis). An overview of datasets characteristics is shown in Table 1. This table summarises, for each dataset, the original number of features (as well as the number of categorical and continuous features, in this order, in brackets), the number of instances, the number of classes as well as the percentage of instances belonging to the majority class.

The experiments were carried out using a standard implementation of the Naïve Bayes learning algorithm available in the Orange Data Mining frame-

work Demšar and ang G. Legan (2004). Furthermore, all FSS and performance measures are also used as implemented in this framework. All the results reported in this paper were obtained using 10-fold cross validation with paired folds, *i.e.*, the same training and testing partitions were used for all compared methods. For problems with more than two classes, the weighed one-against-all method was used to calculate the area under the ROC curve (AUC) Fawcett (2006)[1]. As some of the measures (described next) are only applicable to categorical features, continuous features were previously categorized using equally distance cutoffs with 6 bins (also using the implementation available in the Orange framework).

The experiments can be divided into two independent steps:

1. first, we use a hybrid wrapper/filter approach to select a subset of features as described in Section 3;
2. we use the features selected in the previous step to build a classifier and compute its AUC using an independent test set.

For the filter part of the hybrid wrapper/filter approach, we have experimented with five measures, alongside the combination of them (CFSS) as described in Section 4. These five measures are briefly described next.

Relevance is a measure that discriminates between features on the basis of their potential value in the formation of decision rules Baim (1988).

Gini index is a measure of inequality of a distribution often used in economy. It was first introduced in machine learning by Breiman Breiman et al. (1984) as a decision tree feature splitting criterion.

Information Gain is one of the most popular measures to estimate the expected decrease of Entropy.

Gain ratio was introduced by Quinlan Quinlan (1993) in order to avoid overestimation of multi-valued features. It is computed as information gain divided by the entropy of the feature's value.

Relief was first developed by Kira and Rendell Kira and Rendell (1992) and then substantially generalised and improved by Kononenko Kononenko (1994). It measures the usefulness of features based on their ability to distinguish between very similar examples belonging to different classes.

Tables 2 and 3 show, respectively, the average AUC values and the number of selected features for the hybrid wrapper/filter approach using these individual measures as well as their combination using CFSS to rank the features. For the sake of visualisation, the highest AUC value (not considering the model induced using all the features) as well as the lowest number of selected features are highlighted in gray. In both tables, the numbers in brackets correspond to standard deviations.

[1] In the weighed one-against-all method the AUC is calculated for each class taking one class as positive and the other classes as negative. The overall AUC is the weighed average of all these partial AUCs, using the class priors as weight.

Table 2: Average AUC values. The second column corresponds to the calculated AUC using all features. The following columns correspond to the wrapper/filter approach using the corresponding measure to rank the features. The last column corresponds to the CFSS combination of measures as described in Section 4. Numbers in brackets correspond to standard deviations.

Data Set	All features	Relevance	Gini Index	Infor. Gain	Gain Ratio	Relief	CFSS
anneal	94.9(1.7)	91.4(1.4)	91.8(2.0)	91.2(1.5)	92.3(1.1)	91.3(2.2)	90.6(2.0)
audiology	79.7(1.5)	72.9(2.8)	75.4(2.9)	75.0(2.6)	68.4(2.2)	73.6(2.2)	72.4(3.0)
coil2000	68.8(0.9)	66.0(1.2)	66.7(1.2)	67.0(1.7)	66.9(1.0)	64.9(1.1)	68.2(1.7)
crx	87.6(1.1)	86.5(0.6)	86.5(0.6)	86.3(0.3)	86.2(0.0)	86.6(0.7)	86.5(0.6)
ionosphere	89.3(1.9)	85.4(2.2)	85.5(2.7)	85.7(2.5)	85.1(1.9)	84.4(3.0)	84.7(4.0)
lymphography	71.3(4.0)	78.3(7.6)	75.5(5.5)	74.3(5.6)	79.9(2.5)	76.3(5.0)	77.2(5.2)
mushroom	93.8(0.1)	91.4(0.5)	92.0(0.1)	90.8(1.0)	89.7(0.5)	90.8(0.5)	92.0(0.1)
primary-tumor	65.3(1.0)	65.2(1.3)	64.1(1.5)	66.3(1.4)	66.0(1.6)	65.4(1.6)	66.0(1.8)
promoters	84.0(4.3)	78.1(3.5)	78.2(3.0)	78.5(2.9)	77.9(2.2)	77.6(2.3)	78.5(2.9)
soybean-large	98.8(0.7)	94.6(0.6)	96.3(1.0)	95.7(1.0)	95.4(0.5)	94.7(1.8)	95.3(0.9)
vehicle	82.4(0.6)	78.7(0.9)	78.7(0.6)	78.8(0.5)	78.9(1.1)	79.5(1.2)	79.1(0.5)
voting	97.3(0.5)	96.7(0.0)	96.7(0.0)	96.7(0.0)	96.7(0.0)	96.7(0.0)	96.7(0.0)
wdbc	97.6(0.4)	93.8(2.0)	93.6(1.6)	93.7(1.1)	94.2(1.0)	93.2(1.6)	93.2(1.7)
zoo	95.6(1.1)	92.1(3.7)	91.4(3.5)	91.4(4.0)	91.5(1.5)	92.1(4.5)	91.4(4.0)

Table 3: Average number of selected features using the hybrid wrapper/filter approach. The second column shows the original number of features as a reference. Numbers in brackets indicate standard deviations.

Data Set	All	Relevance	Gini Index	Infor. Gain	Gain Ratio	Relief	CFSS
anneal	38	2.20(0.52)	2.20(0.52)	2.20(0.52)	2.35(0.75)	8.80(1.51)	2.00(0.32)
audiology	69	39.45(28.11)	27.65(31.58)	31.95(31.63)	66.15(12.75)	19.35(21.92)	52.30(29.68)
coil2000	85	44.15(9.92)	1.05(0.22)	1.15(0.37)	7.55(2.37)	41.65(24.79)	2.35(1.04)
crx	15	1.10(0.31)	1.10(0.31)	1.10(0.31)	1.10(0.31)	1.25(0.72)	1.10(0.31)
ionosphere	32	4.30(6.09)	2.35(0.88)	2.35(0.88)	2.65(0.99)	8.35(5.54)	2.80(1.47)
lymphography	18	5.75(6.67)	4.70(6.84)	4.70(6.84)	6.95(6.19)	6.20(7.29)	4.70(6.84)
mushroom	22	3.00(0.00)	3.00(0.00)	2.55(0.51)	6.15(0.37)	3.85(0.37)	3.00(0.00)
primarytumor	17	3.80(1.32)	4.60(1.19)	3.05(0.76)	4.15(0.75)	3.85(1.14)	3.30(0.86)
promoters	57	3.85(12.51)	3.70(12.07)	3.70(12.07)	1.00(0.00)	1.00(0.00)	3.80(12.52)
soybeanlarge	35	4.35(0.59)	5.70(0.92)	3.15(1.42)	8.75(0.97)	12.70(3.92)	2.45(0.51)
vehicle	18	5.00(1.69)	6.60(0.88)	7.00(0.79)	7.90(1.02)	9.10(1.12)	6.40(1.43)
voting	16	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)	1.00(0.00)
wdbc	20	1.40(0.50)	1.10(0.31)	1.05(0.22)	1.85(0.37)	1.25(0.44)	1.45(0.51)
zoo	16	7.50(6.28)	7.05(6.28)	6.55(6.61)	6.70(6.26)	6.45(6.24)	7.20(6.78)

In general, the AUC values presented in Table 2 for the FSS methods are slightly lower than the AUC values obtained using all the features. This result is somewhat expected as the stopping criteria used in these approaches are quite lenient, favouring a large reduction of features despite an improvement on the AUC. Observe that a feature is discarded if the AUC assessed in the reduced dataset is not lower than 95% of the AUC calculated using all the dataset features, as described in Section 3. Only for the lymphography dataset, all FSS methods were able to obtain an improvement in terms of AUC. Other improvements (although in a lower degree) were obtained for the primary tumor dataset using Information Gain, Gain Ration and CFSS. In four datasets (voting, vehicle, promoters and crx), all FSS methods performed quite similarly.

As shown in Table 3, it can be observed that the reduction in the number of features is very high, with the average number of selected features lower than 30% of the original features with a few exceptions, as in all FSS methods for datasets audiology and zoo and for two FSS methods (Relevance and Relief) for the dataset coil2000.

The performance of FSS algorithms in which the aim is to reduce the number of features for learning, must consider at least two aspects simultaneously: the reduction in the number of features *versus* the quality of the induced classifier using the subset of selected features. In other words, this evaluation is multicriteria. To this end, in Lee et al. (2006) an evaluation model for FSS algorithms performance is proposed that considers accuracy as the quality measure of the induced classifier. This model consists of a graph, where the x-axis is related to the classifiers predictive performance and the y-axis is related to the percentage of the selected features by the FSS algorithm. Using this model, in this work the x-axis represents the maximum percentage of AUC degradation we are prepared to accept considering the reduction in the number of features represented in the

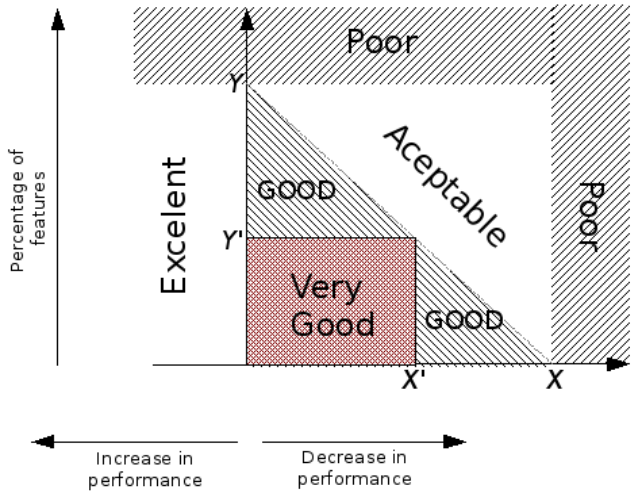

Figure 1: The multicriteria evaluation model

y-axis.

Figure 1 illustrates this performance evaluation model. Each FSS algorithm performance is placed into five categories: excellent (★★), very good (★), good (▲), acceptable (◊) and poor (–). These categories are defined by the user as follows: the value X in Figure 1 represents the maximum percentage of decrease in AUC to be considered as a good result. The value Y represents the maximum percentage of features the FSS should select to be considered a good reduction in the number of features. Taking into account these two measures, the FSS methods can be classified into the five categories previously listed. If there is an increase in classification performance and a minimum reduction in the number of features, the method is considered excellent. Otherwise, two other values X' and Y' falling inside X and Y range, respectively, are defined by the user so that FSS in this X' and Y' rectangle are considered very good. If a FSS method is below the line connecting X and Y although it is not inside the X' and Y' rectangle, it is considered good. Falling above that line but inside the X and Y rectangle, it is considered acceptable. Otherwise, it is considered very poor.

In our evaluation, X was set to 5% (*i.e.*, we considered it would be acceptable up to a maximum reduction of 5% of the AUC using all features) and Y was set to 50% (*i.e.*, we considered it would be acceptable any method that uses at the most half of the features). Thus, all FSS methods that select at most 50% of the original features and improves the original AUC were considered excellent. Furthermore, X' was set to 2.5% and Y' to 25% so that all FSS methods that selected at most 25% of the original features and the AUC values decreased at most 2.5% were considered very good, and so forth.

Table 4 summarises our results using this multicriteria evaluation model. Except for datasets audiology and promoters, for which results were considered

Table 4: Summary of the results using the multicriteria evaluation model

Data Set	Relevance	Gini Index	Infor. Gain	Gain Ratio	Relief	CFSS
anneal	▲	▲	▲	▲	▲	▲
audiology	–	–	–	–	–	–
coil2000	–	▲	▲	▲	◊	★
crx	★	★	★	★	★	★
ionosphere	▲	▲	▲	▲	–	▲
lymphography	★★	★★	★★	★★	★★	★★
mushroom	▲	★	▲	▲	▲	★
primary-tumor	★	▲	★★	★★	★★	★★
promoters	–	–	–	–	–	–
soybeanlarge	▲	★	▲	▲	▲	▲
vehicle	▲	▲	▲	◊	–	▲
voting	★	★	★	★	★	★
wdbc	▲	▲	▲	▲	▲	▲
zoo	◊	◊	◊	◊	▲	◊

poor for all FSS methods, the other datasets presented satisfactory results. In particular, excellent results were obtained for dataset lymphography for all FSS methods and for dataset primary tumor for four out of six FSS methods. In general, all FSS methods presented a similar behaviour although CFSS presents the largest number of excellent and very good rates.

6 Concluding Remarks

This paper presented a hybrid wrapper/filter approach for the feature subset selection problem. This hybrid approach can use different measures to quantify data characteristics which are combined into a single measure using a ranking aggregation function. This hybrid approach aims to save computational time (as pure wrapper approaches are generally computationally expensive) as well as automatically selecting an appropriate number of features (as pure filter based approaches do not present this characteristic).

The approach was validated on 14 datasets from UCI. Results show that, for most of the datasets, this hybrid approach is able to achieve a good reduction in the number of features as well as keeping the AUC comparable to the AUC of the model constructed using all the features in the dataset

As future work, we plan to investigate other ways to set the minimum AUC value used as stopping criterion in our hybrid wrapper/filter approach. For instance, we can include a statistical test to decide whether a reduction in AUC performance is significant or not. In other words, the algorithm continues removing features while this reduction of performance is not significant. We also plan to compare our approach with other hybrid approaches proposed in the literature, such as Zhu et al. (2007); Das (2001).

References

Appice, A., Ceci, M., Rawles, S., and Flach, P. A. (2004). Redundant feature elimination for multi-class problems. In *International conference on Machine Learning (ICML'2004)*. http://doi.acm.org/10.1145/1015330.1015397.

Baim, P. W. (1988). A method for attribute selection in inductive learning systems. *IEEE Trans. Pattern Anal. Mach. Intell.*, 10(6):888–896.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth & Books, Pacific Grove, CA.

Das, S. (2001). Filters, wrappers and a boosting-based hybrid for feature selection. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, pages 74–81.

Demšar, J. and ang G. Legan, B. Z. (2004). Orange: From experimental machine learning to interactive data mining. Technical report, Faculty of Computer and Information Science, University of Ljubljana. White Paper (www.ailab.si/orange).

Duch, W. (2006). Feature extraction, foundations and applications. chapter Filter Methods, pages 89–118. Physica-Verlag, Springer.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

Guyon, I. and Elisseeff, A. (2003). An introduction of variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182.

Kira, K. and Rendell, L. A. (1992). A practical approach to feature selection. In *Proceedings of the Ninth International Workshop on Machine Learning (ML 1992)*, pages 249–256. Morgan Kaufmann.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324.

Kononenko, I. (1994). Estimating attributes: Analysis and extensions of relief. In *European Conference on Machine Learning (ECML'1994)*, volume 784 of *Lecture Notes in Computer Science*, pages 171–182. Springer.

Lal, N., Chapelle, O., Weston, J., and Elisseeff, A. (2006). *Feature Extraction, Foundations and Applications*, chapter Embedded methods, pages 139–167. Physica-Verlag, Springer.

Lee, H. D., Monard, M. C., Voltoline, R. F., Prati, R. C., and Chung, W. F. (2006). A simple evaluation model for feature subset selection algorithms. *Inteligencia Artificial*, 10(32):9–17. Special issue with best ASAI'2006 papers.

Newman, D., Hettich, S., Blake, C., and Merz, C. (1998). UCI repository of machine learning databases. `http://www.ics.uci.edu/~mlearn/MLRepository.html`.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann.

Vafaie, H. and Jong, K. D. (1993). Robust feature selection algorithms. In *IEEE Int. Conf. on Tools with AI*, pages 356–363. IEEE Computer Society Press.

Yu, L. and Liu, H. (2004). Efficient feature selection via analysis of relevance and redundance. *Journal of Machine Learning Research*, 5:1205–1224.

Zhu, Z., Ong, Y.-S., and Dash, M. (2007). Wrapper-filter feature selection algorithm using a memetic framework. *IEEE Transactions On Systems, Man and Cybernetics*, 37(1):70–76. Part B.