Heuristic Ensemble of Filters for Reliable Feature Selection

Ghadah Aldehim, Beatriz de la Iglesia and Wenjia Wang
School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, U.K.

Keywords: Feature selection, Ensemble, Classification, Heuristics.

Abstract: Feature selection has become ever more important in data mining in recent years due to the rapid increase in the dimensionality of data. Filters are preferable in practical applications as they are much faster than wrapper-based approaches, but their reliability and consistency vary considerably on different data and yet no rule exists to indicate which one should be used for a particular given dataset. In this paper, we propose a heuristic ensemble approach that combines multiple filters with heuristic rules to improve the overall performance. It consists of two types of filters: subset filters and ranking filters, and a heuristic consensus algorithm. The experimental results demonstrate that our ensemble algorithm is more reliable and effective than individual filters as the features selected by the ensemble consistently achieve better accuracy for typical classifiers on various datasets.

1 INTRODUCTION

With the rapid advance in computer and database technologies, datasets with hundreds or thousands of features are now ubiquitous. However, most of the features in enormous datasets may be irrelevant or redundant, which can cause poor efficiency and overfitting in the learning algorithms. Therefore, it is necessary to employ some feature selection methods to select the most relevant features from the dataset. This should lead to improve efficiency and generate more accurate models (Saeys et al., 2007).

Methods for feature selection are roughly divided into two main categories: filters and wrappers. The core of wrapper approach is to employ a model trained from the given data to evaluate the discriminative power of features. It is generally more accurate but highly model-dependent and very time consuming (Kohavi and John, 1997). On the other hand, the filter method is more efficient as it uses general characteristics, such as relevance or correlation, of the data to select certain features without involving any learning algorithm (Blum and Langley, 1997).

There are, however, many different types of filters and their performance in terms of accuracy, consistency and reliability varies considerably from one dataset to another. It is not clear when a particular filter should be used for a given dataset. Hence, it is logical and often necessary to employ an ensemble approach in feature selection. As ensembles have demonstrated to be successful in classification problems. Certain concepts and methods for feature selection ensembles have been proposed (Yang et al., 2011; Saeys et al., 2008; Wang et al., 2010b), but they have only been investigated and tested with limited ranking filters and the bootstrap method or a simple arithmetic mean as consensus strategies. In this paper, we propose a heuristic ensemble method that combines the results of multiple filters through heuristic rules. The method is implemented and tested on various benchmark datasets and the results are promising.

2 RELATED WORK

Vast amount of literature exists in the feature selection research field, and as our study aims to develop a fast and reliable filter-based ensemble for feature selection, we focus our review only on filters.

Filter methods typically fall into two categories: rank and subset in terms of the format of output. Rank filters (RF) evaluate one feature at a time and the outputs are ranked by their individual discrimination power (Kira and Rendell, 1992; Kononenko, 1994), whereas subset filters (SF) evaluate subsets of features and output the best subset (Yu and Liu, 2003; Hall, 1999; Sun et al., 2012; Zhang and Zhang, 2012). Many researchers have pointed out the key drawback of feature ranking methods, that is, they assess features on an individual basis and thus do not consider

Aldehim G., de la Iglesia B. and Wang W..
Heuristic Ensemble of Filters for Reliable Feature Selection.
DOI: 10.5220/0004812401750182
In *Proceedings of the 3rd International Conference on Pattern Recognition Applications and Methods* (ICPRAM-2014), pages 175-182
ISBN: 978-989-758-018-5

possible relationships among features. Most RF tend to select those features that are identified as being individually relevant to the target class, even when they may be highly correlated to each other. Then it is possible that "the selected m best features are not the best m features" (Zhang et al., 2003). With SF, the number of candidate features subsets increases exponentially with feature dimensionality and it is not feasible to carry out an exhaustive search even for a medium-sized dataset (Zhang and Zhang, 2012). Thus, the use of subset filters entails a trade-off between computational cost and the quality of the selected feature subset, and this must be considered when developing an efficient and effective feature selection method (Yu and Liu, 2004).

(Saeys et al., 2008) proposed an ensemble built with four feature selection techniques: two filter methods (Symmetrical Uncertainty and ReliefF) and two embedded methods (Random Forests and linear SVM). For each of the four feature selection techniques, an ensemble version was created by using bootstrap aggregation. For each of the bags, a separate feature ranking was performed, and the ensemble was formed by aggregating the single ranking by weighted voting, using linear aggregation. (Olsson and Oard, 2006) studied ensembles of multiple feature ranking techniques in order to resolve text classification problems. They used three filter-based feature ranking techniques: document frequency thresholding, information gain, and the chi-square method (χ^2max and χ^2avg).

(Wang et al., 2010a) also studied the ensembles of commonly used filter-based rankers but they used six filters and increased to 18 later (Wang et al., 2012). The combining methods used in that study included arithmetic mean, where each features score is determined by the average of the ranking scores of the features in each ranking list. The highest ranked attributes are then selected from the original data to form the training dataset. They examined the performance of models with selected features using 17 different ensembles of rankers. The results show that an ensemble of very few rankers usually performs similarly or even better than ensembles of many or all rankers (Wang et al., 2012).

Recently, (Yang et al., 2011) used ReliefF (Robnik-Šikonja and Kononenko, 2003) and tuned ReliefF (TuRF) (Moore and White, 2007) for identifying SNP-SNP interactions. However, they observed that the 'unstable' results from the multiple runs of these algorithms can provide valuable information about the dataset. They therefore hypothesized that aggregating the results derived from the multiple runs of a single algorithm may improve filtering performance.

In summary, the review found that these studies were predominantly limited to using one type of filters, i.e. rank filters, as the member components of ensemble, which produces a ranking of features. Then some additional work needs to be performed to decide a cutting off point to produce a subset of selected features. In this study, we propose an ensemble framework that combines two types of filters - SF and RF by means of heuristic rules to utilise their advantages. The detail is explained in the following section.

3 HEURISTIC ENSEMBLE OF FILTERS (HEF)

3.1 Proposed Heuristic Ensemble of Filters (HEF)

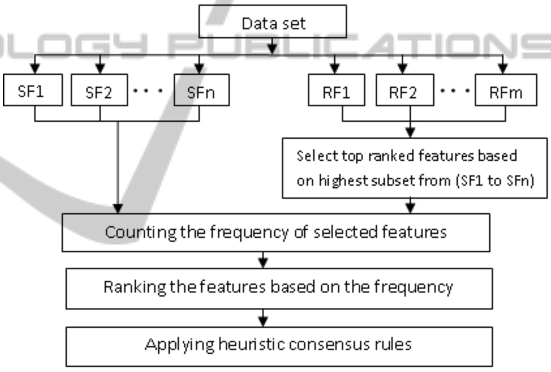
Data set
SF1
SF2
• • •
SFn
RF1
RF2
• • •
RFm
Select top ranked features based on highest subset from (SF1 to SFn)
Counting the frequency of selected features
Ranking the features based on the frequency
Applying heuristic consensus rules

Figure 1: Framework of heuristic ensemble of filters (HEF) for feature selection.

The proposed heuristic ensemble of filters (HEF), as shown in Fig.1, is composed of two types of filters: SF and RF as its members, and a heuristic algorithm as its consensus function. The idea of combining subset filters and rankers is to exploit the advantages of each. Firstly, rank filters usually assess individual features and assign their weights according to their degree of relevance. But this does not ensure conditional independence among the features, and may lead to selecting features that are redundant or have less discriminative ability. Subset filters take into account the existence and effect of redundant features, which to some extent approximate the optimal subset. However, this method entails high computational cost in terms of the subset searches, making subset filters inefficient for high dimensional data. As a result, to obtain the benefits of subset filtering without suffering the high computational cost, we choose fast subset filters, as described in section 3.2.

The process of the proposed heuristic ensemble of filters starts by running SF and RF. After that, a consensus number of features selected by the subset filters (SF) is taken as a cut-off point for the rankings generated by the ranking filters (RF). By running this heuristic step, we can obtain quick answers for cutting off the number of features in the ranker, which will accelerate the ensemble algorithm. Therefore, we will not need to select various feature numbers to test the performance or to use a wrapper to choose the appropriate number of features. The next step aggregates the results from the above sets. A heuristic consensus rule is applied to produce the final output of the ensemble.

The proposed ensemble framework is implemented in Java, primarily based on the modules provided in Weka and other standalone filter software.

3.2 Choice of Individual Filters

In principle, any filters of each type can be used as the member filters of our HEF. However, some factors should be considered when choosing the filters, which include speed, reliability and scalability. In terms of determining the number of member filters, we followed the guideline given in (Wang et al., 2010b), that is, an ensemble of very few carefully selected filters is similar to or better than ensembles of many filters. So, in this concept demonstration study, we choose four filters which are briefly described as follows to give an idea why they are selected in this study.

CSF: Correlation-based Feature Selection (Hall, 1999) is a simple filtering algorithm that ranks feature subsets according to a correlation-based heuristic evaluation function. The key idea of this algorithm is that it employs a heuristic evaluation that assesses the efficacy of individual features in terms of predicting the class. It also assesses how far the features are intercorrelated. In order to avoid high computational cost, we use liner forward selection (LSF) as a search method together with CSF instead of using Best First search. LSF is a simple complexity optimization of sequential forword selection(SFS). It entails firstly creating a filter ranking and selection of the K first features; then, the SFS algorithm is run over the selected features (Gutlein et al., 2009).

FCBF: Fast Correlation Based Filter (Yu and Liu, 2004) starts by sorting features through correlation with a response using symmetric uncertainty, optionally removing the bottom of the list according to some user-specified threshold. Then, the feature that is most correlated with the response is selected. After that, all features that have correlation with the selected feature higher than its correlation with the response are considered redundant and removed. Then, the feature is added to the minimal subset and the search starts again with the next feature.

Relief: This was first proposed by Kira and Rendell (Kira and Rendell, 1992) and then improved by Kononenko (Kononenko, 1994) to handle noise and multi-class datasets. The key idea of Relief is that it searches for the nearest neighbours of a sample of each class label, and then weights the features in terms of how well they differentiate samples for different class labels. This process is repeated for a prespecified number of instances.

Gain Ratio: this is one of the simplest and fastest feature ranking methods. It incorporates split information of features into an Information Gain statistic. The split information of a feature is obtained by measuring how broadly and uniformly the data are split. Generally, Gain Ratio evaluates the value of a feature by measuring the gain ratio with respect to the class (Quinlan, 1993).

3.3 The Heuristic Consensus Rules

The outputs of the different filters need to be aggregated through a consensus function to generate the final feature selection output of the ensemble. A consensus function can be defined from different perspective, e.g. as simple as counting the frequency of selected features (Saeys et al., 2008) to some sophisticated weighting algorithms. In this work, we focus on ensemble features selection techniques that work by aggregating the feature subsets provided by the different filters into a final consensus subset. The most frequently selected features are placed at the top, while the least frequently selected features are placed at the bottom. However, aggregating the outputs by counting the most frequently selected features may produce a high number of selected features. In order to address this issue and also to get more important features, a heuristic consensus rule is applied to produce the final output of the HEF.

Various heuristic rules can be derived based on the purpose of the analysis. Some example are described bellow:

R0	$\longmapsto$	remove nothing from the HEF.
R1	$\longmapsto$	remove features selected by only one filter.
R2	$\longmapsto$	remove features selected by only two filters.
.		
.		
RQ	$\longmapsto$	remove features selected by Q filters. $\forall Q < n+m$

Where Q is the heuristic consensus rule, n the number of subset filters and m the number of ranking filters. The heuristic rule, $R0$, uses all the features selected by any of the four filters while rule $R1$ removes any features selected by only one filter. Other heuristic rules can be defined but in this paper, $R0$ and $R1$ are implemented in two ensembles named HEF-R0 and HEF-R1 respectively. A good feature set requires some diversity, but having more agreement among the filters may decrease diversity.

4 EXPERIMENTS

4.1 Data

Table 1: Description of the testing datasets.

Dataset	# Features	# Instances	# classes
Zoo	17	101	7
Dermatology	34	366	6
Promoters	57	106	2
Splice	61	3,191	3
M-feat-factors	216	2,000	10
Arrhythmia	279	452	13
Colon	2,000	62	2
SRBCT	2,308	83	4
Leukaemia	7,129	72	2
CNS	7,129	60	2
Ovarian	15,154	253	2

Eleven benchmark datasets (shown in table 1) selected from different domains were used in our experiments to test the performance of our proposed heuristic ensemble of filters. Six of them, Zoo, Dermatology, Promoters, Splice, Multi-feature-factors and Arrhythmia, are from the UCI Machine Learning Repository[1], two others (Colon and Leukaemia) from the Bioinformatics Research Group[2], and the final three (SRBCT, Central Nervous System (CNS) and Ovarian) from the Microarray Datasets website[3]. Note that these datasets differ greatly in sample size (ranging from 60 to 3,191) and number of features (ranging from 17 to 15,154). Also, they include binary-class and multi-class classification problems. This should provide for testing and should be well suited to the feature selection methods under differing conditions.

4.2 Experiment Design and Procedure

As it is generally accepted that the effectiveness of feature selection can be indirectly evaluated through measuring classification performance of classifiers that are trained with the selected features, we thus conducted several series of experiments with a variety of datasets to empirically evaluate the performances of the HEFs and compare them with each individual filter used in this study, and also the full feature set without any feature selection performed. The classification performance may be dependent on types of classifiers used even under the exactly same conditions, same subset of features and samples, and training procedure. To verify the consistency of the feature selection methods, in our experiments, we used three types of classifiers: NBC (Naive Bayesian Classifier)(John and Langley, 1995), KNN (*k*-Nearest Neighbor)(Aha et al., 1991) and SVM (Support Vector Machine)(Platt, 1999). These three algorithms were chosen because they represent three quite different approaches in machine learning and they are state-of-the-art algorithms that are commonly used in data mining practice.

The parameters of classifiers and filters for each experiments are set to the default value of weka. For each dataset, the experiments are carried out in two phases: feature selection phase and evaluation phase. The first phase to run HEF to produce a subset of ranked features, as well as the subsets selected by each individual filters. The second phase is to evaluate the effectiveness of the selected features with three kinds of models: NBC, KNN and SVM. Specifically, it firstly trains the model of each type with the full set of features and the subsets produced by FCBF, CFS, ReliefF, Gain Ratio, HEF and HEF-R1, using the 10-fold cross validation strategy for each classifier. Each experiment is then repeated ten times with different shuffling random seeds in order to assess the consistency and reliability of the results. The statistical significance of the results of multiple runs for each experiment is calculated and the comparison between accuracies is done with Students paired two-tailed *t*-test with a significance level of 0.05. In total, 23,100 models were built for the experiments$(7(FS+ensemble) \times 11(datasets) \times 3(classifiers) \times 10(run) \times 10(folds))$.

5 RESULTS AND ANALYSIS

5.1 Results of Feature Selections

Table 2 lists the number of features selected by each filter in addition to two heuristic ensembles: HEF(HEF-R0) and HEF-R1. We observe from the table that the average number of selected features dramatically reduced the dimensionality of the data by selecting only a small proportion of the original features in those datasets. Although HEF represents the

[1] http://repository.seasr.org/Datasets/UCI/arff/

[2] http://www.upo.es/eps/aguilar/datasets.html

[3] http://csse.szu.edu.cn/staff/zhuzx/Datasets.html

Table 2: Number of the features selected by four individual filters and two ensembles for each dataset.

Dataset	All features	FCBC	CSF	ReliefF	Gain Raito	HEF	HEF-R1
Zoo	17	7	10	10	10	11	11
Dermatology	34	16	19	19	19	28	24
Promoters	57	6	6	6	6	7	6
Splice	61	22	22	22	22	29	25
M-feat-factors	216	38	47	47	47	82	62
Arrhythmia	279	12	21	21	21	52	17
Colon	2,000	14	23	23	23	50	21
SRBCT	2,308	82	77	82	82	177	92
Leukaemia	7,129	51	52	52	52	111	58
CNS	7,129	28	36	36	36	60	37
Ovarian	15,154	30	36	36	36	76	43
Average	3,125.18	27.81	31.72	32.18	32.18	62.09	36
St.Dv.	4,829.8	22.59	20.76	21.88	21.88	49.33	25.92

Table 3: The accuracies of the NB Classifiers trained by the selected features and all the features.

Dataset	All features	FCBC	CSF	ReliefF	Gain Raito	HEF	HEF-R1
Zoo	93.96	93.56	94.25	92.27-	**95.24+**	95.05	95.05
Dermatology	97.43	97.86	**98.55+**	96.06-	85.32-	98.20+	98.52+
Promoters	90.19	**94.62+**	94.52+	93.86+	**94.62+**	93.71+	94.57+
Splice	95.41	96.16+	96.16+	96.24+	95.98+	96.04+	**96.33+**
M-feat-factors	92.47	93.60+	**93.68+**	87.16-	89.98-	92.53	92.98
Arrhythmia	62.39	65.86+	68.93+	65.66+	53.25-	68.87+	**69.60+**
Colon	55.81	84.67+	85.00+	85.80+	83.06+	**85.86+**	85.55+
SRBCT	99.04	99.63	**100+**	**100+**	82.00	**100+**	**100+**
Leukaemia	98.75	**99.44+**	98.61	95.97-	95.97-	98.61	98.61
CNS	61.00	76.49+	76.66+	75.00+	72.33+	74.83+	**77.33+**
Ovarian	92.41	**99.92+**	99.84+	98.34+	98.02+	98.81+	98.81+
Average	85.35	91.07	91.47	89.67	87.57	91.14	**91.58**
St.Dv.	16.74	10.99	10.28	10.69	13.93	10.39	9.94
W/T/L		8/3/0	9/2/0	7/0/4	6/1/4	8/3/0	8/3/0

Table 4: The accuracies of the KNN Classifiers trained by the selected features and all the features.

Dataset	All features	FCBC	CSF	ReliefF	Gain Raito	HEF	HEF-R1
Zoo	96.14	96.04	96.04	**97.03+**	96.04	96.04	96.04
Dermatology	94.64	95.57+	**97.10+**	94.29	86.45-	95.54+	96.91+
Promoters	79.71	**91.13+**	**91.13+**	89.99+	**91.13+**	90.19+	**91.13+**
Splice	74.43	81.21+	81.21+	80.52+	**82.06+**	79.59+	80.46+
M-feat-factors	96.03	96.36+	**96.44+**	93.48-	95.32+	96.31+	96.34+
Arrhythmia	53.20	69.82+	61.39+	57.76+	43.52-	57.52+	**61.88+**
Colon	76.83	78.38+	81.45+	85.8+	77.74	**86.3+**	80.71+
SRBCT	82.39	99.87+	**100+**	**100+**	**100+**	**100+**	**100+**
Leukaemia	84.39	**99.58+**	97.49+	95.41+	94.44+	98.48+	98.77+
CNS	59.50	83.66+	76.5+	76.50+	**84.83+**	80.17+	82.83+
Ovarian	94.86	**100+**	99.96+	99.13+	98.85+	**100+**	**100+**
Average	81.52	89.23	88.97	87.77	86.39	89.10	**89.55**
St.Dv.	14.85	12.47	12.34	12.74	15.92	12.82	11.91
W/T/L		10/1/0	10/1/0	9/1/1	7/2/2	10/1/0	10/1/0

total number of features selected from all the four filters, it is still less than the average full set by up to 71 times for genetic datasets.

5.2 Feature Selection Evaluation with Different Classifiers

For comparison, all the original features for each dataset are also used in testing. For each dataset,

Table 5: The accuracies of the SVM Classifiers trained by the selected features and all the features.

Dataset	All features	FCBC	CSF	ReliefF	Gain Raito	HEF	HEF-R1
Zoo	**96.24**	96.03	96.13	95.24	95.14-	95.45-	95.45-
Dermatology	96.04	97.67+	**98.06**+	95.63	88.71-	**98.06**+	98.01+
Promoters	91.03	92.83+	92.83+	91.98	92.83+	91.86	**92.86**+
Splice	93.13	95.92+	95.91+	**95.98**+	95.95+	94.15+	94.30+
M-feat-factors	**97.70**	97.15-	97.26-	96.12-	96.91-	97.62	97.43
Arrhythmia	**71.06**	58.6-	67.83-	68.36-	59.13-	69.62-	61.86-
Colon	84.52	88.7+	88.22+	87.42+	83.06	**88.93**+	86.69+
SRBCT	99.63	99.63	99.87	**100**+	98.67-	**100**+	**100**+
Leukaemia	98.04	**99.3**+	97.49	97.22-	97.08-	98.32	98.32
CNS	67.16	**90.50**+	88.5+	76.83+	87.33+	89.17+	88.83+
Ovarian	99.96	**100**+	**100**+	99.56-	99.56-	**100**+	**100**+
Average	90.41	92.39	92.92	91.30	90.40	**93.02**	92.16
St.Dv.	11.44	11.80	9.24	10.04	11.58	8.71	10.94
W/T/L		6/3/2	5/4/2	3/4/4	3/1/7	4/5/2	5/3/3

with each selection, and for each type of 3 models (NBC, KNN and SVM), 100 models (ten runs of ten-fold cross validations) are generated and their average testing accuracies are calculated.

Table 3 shows the results on the eleven datasets with the Naive Bayesian Classifier. The notations + or - denote that the result of the classification of the models trained with the features selected with the current selector is significantly better or worse than that of models trained with all the original features in the statistical test mentioned earlier. The bold value in each row shows the best classification result. The last three rows in each table show the average accuracies, the standard deviations for the accuracies and W/T/L (which summarizes the wins/ties/losses in accuracy by comparing the models trained with all the features and the features selected by other).

As expected, each single filter performed well in some datasets (in bold) but poorly in others. That confirms the perception that the performance of individual filters is inconsistent and unreliable and there is no meaningful pattern can be extracted to indicate when they do better and when they do not. Nevertheless, The NB classifiers trained with the features selected by HEF-R1 have a higher average accuracy for all the datasets and a lower standard deviation, which indicates that HEF-R1 are not only more reliable and consistent but also more accurate than the individual filters in feature selection. In addition, HEF-R1 achieves the highest accuracy on four datasets. Comparing the results for this classifier using the full feature set with others, it can be observed that in most cases, the accuracy is increased in HEF-R1, HEF, CSF and FCBC, while in the rank filters, the performance is poorer than in the others but still better than full feature set.

The results from the KNN (k = 1) classifiers in table 4, show similar patterns to those in Table 3 with lower accuracy in general than NBC, but again the individual filters demonstrate to be less reliable compared with HEF-R1.

The results from SVM classifiers in table 5, show that ensembles performed consistently, This time HEF is the overall winner as it has a marginally higher average accuracy and a lower standard deviation than all the others, although two subset filters produced similar performance under this experimental conditions. A different phenomenon is the SVM models trained with the full feature set, as they performed not as bad as the other two types (NB and KNN) of models and even gave the highest accuracy on three datasets (Zoo, Multi-Feature Factor and Arrhythmia). The average accuracy of SVM models trained with all the features is the same as that trained with features selected by Gain Ratio filter, not much worse than the rest in terms of accuracy, but SVMs using the full features are less efficient than the SVMs using fewer features. So, feature selection is still beneficial with SVM as a classifier.

In general, the most important benefit of using all ensemble is to achieve high consistency and reliability as well as a relatively high accuracy. So, we wish for an ensemble to be comparable to the ”best” member in an ensemble in accuracy but more reliable than the ”best” members. In our experimental results CFS indeed is comparable to HEF in some cases but it did not do well in others. Therefore, in general HEF is better than CSF.

6 CONCLUSIONS AND FUTURE WORK

In this paper, a framework of heuristic ensemble of filters (HEF) has been proposed to overcome the weaknesses of single filters. It combines the outputs from two types of filters, SF and RF, with heuristic rules as consensus functions to improve the consistency and effectiveness in feature selection. The proposed HEF and HEF-R1 have been tested on 11 benchmark datasets with the number of features varied from 17 to as many as 15,154. The statistical analysis on the experimental results show that the ensemble technique performed more consistently and in some cases even more accurate than individual filters. Specifically,

1. HEF-R1 performed best for NBC and KNN, while HEF performed best when using the SVM classifier, which demonstrates that our proposed ensemble is more reliable and consistent than using single filters.
2. There is no single best approach for all the situations. In other words, the performance of the single filters varies from dataset to dataset and also was influenced by the type of models chosen as classifier. Thus, one filter may perform well in a given dataset for a particular classifier but perform poorly when used on a different dataset or with a different type of classifier.
3. Among the four filters we used in our heuristic ensemble of filters, the subset filters (FCBF and CSF) were more frequently better and less frequently worse on average than the rank filters.
4. The experimental results show that the ensemble technique performed better overall than any individual filter in terms of reliability, consistency and accuracy.

Future work may include additional experiments measuring the stability of our approach, which would represent an additional way to evaluate our results. In addition, investigations could be conducted on different numbers and types of filters. Finally, we plan to use ensemble classification to overcome the differentials between the individual classifiers.

REFERENCES

Aha, D. W., Kibler, D., and Albert, M. K. (1991). Instance-based learning algorithms. *Machine learning*, 6(1):37–66.

Blum, A. L. and Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97(1):245–271.

Gutlein, M., Frank, E., Hall, M., and Karwath, A. (2009). Large-scale attribute selection using wrappers. In *Computational Intelligence and Data Mining, 2009. CIDM'09. IEEE Symposium on*, pages 332–339. IEEE.

Hall, M. A. (1999). *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, https://www.lri.fr.

John, G. H. and Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 338–345. Morgan Kaufmann Publishers Inc.

Kira, K. and Rendell, L. A. (1992). The feature selection problem: Traditional methods and a new algorithm. In *Proceedings of the National Conference on Artificial Intelligence*, pages 129–129. John Wiley & Sons Ltd.

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

Kononenko, I. (1994). Estimating attributes: analysis and extensions of relief. In *Machine Learning: ECML-94*, pages 171–182. Springer.

Moore, J. H. and White, B. C. (2007). Tuning relieff for genome-wide genetic analysis. In *Evolutionary computation, machine learning and data mining in bioinformatics*, pages 166–175. Springer.

Olsson, J. and Oard, D. W. (2006). Combining feature selectors for text classification. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 798–799. ACM.

Platt, J. C. (1999). 12 fast training of support vector machines using sequential minimal optimization.

Quinlan, J. R. (1993). *C4. 5: programs for machine learning*, volume 1. Morgan kaufmann.

Robnik-Šikonja, M. and Kononenko, I. (2003). Theoretical and empirical analysis of relieff and rrelieff. *Machine learning*, 53(1-2):23–69.

Saeys, Y., Abeel, T., and Van de Peer, Y. (2008). Robust feature selection using ensemble feature selection techniques. In *Machine Learning and Knowledge Discovery in Databases*, pages 313–325. Springer.

Saeys, Y., Inza, I., and Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517.

Sun, X., Liu, Y., Li, J., Zhu, J., Chen, H., and Liu, X. (2012). Feature evaluation and selection with cooperative game theory. *Pattern Recognition*, 45(8):2992–3002.

Wang, H., Khoshgoftaar, T., and Gao, K. (2010a). Ensemble feature selection technique for software quality classification. In *Proceedings of the 22nd International Conference on Software Engineering and Knowledge Engineering*, pages 215–220.

Wang, H., Khoshgoftaar, T. M., and Napolitano, A. (2010b). A comparative study of ensemble feature selection techniques for software defect prediction. In

Machine Learning and Applications (ICMLA), Ninth International Conference on, pages 135–140. IEEE.

Wang, H., Khoshgoftaar, T. M., and Napolitano, A. (2012). Software measurement data reduction using ensemble techniques. *Neurocomputing*, 92:124–132.

Yang, P., Ho, J., Yang, Y., and Zhou, B. (2011). Gene-gene interaction filtering with ensemble of filters. *BMC bioinformatics*, 12(Suppl 1):S10.

Yu, L. and Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Machine Learning International Workshop*, volume 20, page 856.

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

Zhang, L.-X., Wang, J.-X., Zhao, Y.-N., and Yang, Z.-H. (2003). A novel hybrid feature selection algorithm: using relieff estimation for ga-wrapper search. In *Machine Learning and Cybernetics, 2003 International Conference on*, volume 1, pages 380–384. IEEE.

Zhang, Y. and Zhang, Z. (2012). Feature subset selection with cumulate conditional mutual information minimization. *Expert Systems with Applications*, 39(5):6078–6088.