Accepted Manuscript

On feature selection protocols for very low-sample-size data

Ludmila I. Kuncheva, Juan J. Rodríguez

PII: S0031-3203(18)30102-X
DOI: 10.1016/j.patcog.2018.03.012
Reference: PR 6490

To appear in: *Pattern Recognition*

Received date: 29 September 2017
Revised date: 28 February 2018
Accepted date: 11 March 2018

Please cite this article as: Ludmila I. Kuncheva, Juan J. Rodríguez, On feature selection protocols for very low-sample-size data, *Pattern Recognition* (2018), doi: 10.1016/j.patcog.2018.03.012

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Highlights

- Feature selection with very few instances, possibly high-dimensional.
- Widely used protocol: 1) feature selection, 2) cross-validation to test a classifier.
- Alternative, proper, protocol includes both steps in a single cross-validation loop.
- Experiment using 24 datasets, 3 feature selection methods and 5 classifier models.
- The proper protocol accuracy is significantly closer to the true accuracy.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

On feature selection protocols for very low-sample-size data

Ludmila I. Kuncheva[a], Juan J. Rodríguez[b]

[a] *Bangor University, Dean Street, Bangor Gwynedd, LL57 1UT, United Kingdom*
[b] *Universidad de Burgos, Escuela Politécnica Superior, Avda. de Cantabria s/n, 09006 Burgos, Spain*

Abstract

High-dimensional data with very few instances are typical in many application domains. Selecting a highly discriminative subset of the original features is often the main interest of the end user. The widely-used feature selection protocol for such type of data consists of two steps. First, features are selected from the data (possibly through cross-validation), and, second, a cross-validation protocol is applied to test a classifier using the selected features. The selected feature set and the testing accuracy are then returned to the user. For the lack of a better option, the same low-sample-size dataset is used in both steps. Questioning the validity of this protocol, we carried out an experiment using 24 high-dimensional datasets, three feature selection methods and five classifier models. We found that the accuracy returned by the above protocol is heavily biased, and therefore propose an alternative protocol which avoids the contamination by including both steps in a single cross-validation loop. Statistical tests verify that the classification accuracy returned by the proper protocol is significantly closer to the true accuracy (estimated from an independent testing set) compared to that returned by the currently favoured protocol.

Keywords: feature selection, wide datasets, experimental protocol, training/testing, cross-validation,

1. Introduction

Selecting a feature subset of low cardinality and high discrimination power has been a centre-stage quest since the dawn of pattern recognition [1, 2, 3, 4]. Feature selection from high-dimensional data has been extensively studied [5, 6, 7, 8, 9, 10, 11, 12]. In many cases, feature selection is sought as the end goal of the data analysis. For example, the user may wish to know which combination of genes out of several thousand genes forms a distinctive signature for a particular disorder [13]. In neuroscience, the user may be interested in the multi-voxel patterns of brain activation which discriminate between different cognitive states. Finding such multi-voxel patterns can be cast as a feature selection problem [14].

Wide datasets are characterised by a large number of features (high dimensionality) and a small number of objects. Such wide datasets are common in many areas, examples of which are neuroimaging, bioinformatics,

Email addresses: l.i.kuncheva@bangor.ac.uk (Ludmila I. Kuncheva), jjrodriguez@ubu.es (Juan J. Rodríguez)

ACCEPTED MANUSCRIPT

psychology, and sport sciences. What if the available sample has only a couple of dozens examples? This may happen when data does not exist in abundance, for example, studies of rare diseases or extraordinary athletes. Sometimes collecting of such data is prohibitively expensive or destructive. How reliable are any conclusions drawn from such datasets? In particular, how meaningful is feature selection? Ultimately, we can offer the user a subset of the original features, together with a trained classifier model, and an estimate of the classification accuracy. The classification accuracy in itself is a gauge of how good the returned feature set is. Here we argue that, quite often, we are misleading the user by returning to them an optimistically biased estimate of the classification accuracy. One reason for this bias is the so called "*peeking* phenomenon", which has already been brought to the attention of the community [15, 16], especially in the light of experimenting with high-dimensional data [17, 18]. The "peeking" happens if the data for testing the model is seen during some part of the training. Peeking usually happens when there is a preliminary training stage, for example data quantisation, feature selection, or parameter tuning. The effect is that the estimate of the classification accuracy which we return to the user may be optimistically biased. More importantly, the returned feature set may also be an artefact rather than a highly discriminative set.

While the caution of overfitting in feature selection has been raised several times over the past years [16, 15], it does not seem to have been properly addressed by the larger community, and especially in applications which are most vulnerable. Curiously, a comprehensive recent survey by Li et al. [19] does not even mention the issue, while another one ~~uses~~ considers only the training data for feature selection [20][1]. Instead, these studies review elaborate methods for stable, sparse, and multi-source feature selection from wide data. All these developments critically depend on using the correct training/testing protocol, and may not be adequate at all for very small sample size data. The difficulty in offering a stable and unbiased estimate of the classification accuracy may render the selected feature subsets no better than chance. In addition to joining the appeal for clean, non-contaminated feature selection protocols, here we set out to address two further issues. First, we demonstrate the deficiency of the widely used (flawed) protocol using 24 high-dimensional datasets, three feature selection methods and 5 classifiers. Second, we propose a clean protocol and show that its accuracy matches significantly closer the accuracy estimated from a properly sized datasets. The rest of the paper is organised as follows. Related work is presented in Section 2. Section 3 discusses the right and the wrong protocol for feature selection, and gives an example of the optimistic bias which the wrong protocol is prone to. Section 4 reports and discusses our experimental results, followed by our recommendations in Section 5.

[1] https://doi.org/10.1145/3136625, https://arxiv.org/abs/1601.07996

ACCEPTED MANUSCRIPT

2. Related work

2.1. Many studies are unaware of the overfitting caveat

Smialowski et al. [16] and Reunanen [15] warn about the optimistic bias of an improperly applied feature selection protocol, and emphasise the importance of using testing data unseen at *any* part of the feature selection and the classifier training. In spite of this warning, "peeking" is still widely present, casting doubts in the findings of the respective studies. Sometimes it is not clear whether the training/testing protocol has been applied only to testing the classifier or to the feature-selection-classifier-training together. A keyword search for the joint term 'feature selection' on Web-of-Science[2], carried out on the 11 January 2018 returns over 2,300 articles since 2017. A thorough systematic analysis of these publications in the light of our research question is infeasible, hence we opted for a small set of random examples. We selected these examples blindly, without specifically looking for articles which will confirm our concern about the wrong protocol. Out of the 17 papers we picked, 6 apply the wrong protocol, 4 do not give explicit details to judge either way, and 7 apply the training/testing correctly [21, 22, 23, 24, 25, 26, 27]. We took our motivation from the alarmingly high proportion of studies oblivious to the overfitting caveat. These findings make our message even more important because the comparisons in these studies (not cited for obvious reasons), and the related claims, may be compromised by using a flawed evaluation metric.

2.2. Peaking and peeking

We should be cautious not to confuse "peeking" with "peaking". The "peaking phenomenon", also called in the past "peak-effect" or "The Hughes paradox" [28, 29, 30] is now well documented. The paradox is that by discarding information (features), we may obtain a better classifier. There are at least two causes for this phenomenon. First, the classifier model is never the perfect (Bayes) classifier. If, hypothetically, we knew the exact probability distributions of the classes, all relevant features will be suitably exploited, and all irrelevant ones, ignored. There will be no decline in the accuracy if more features are included, be they relevant or irrelevant. Since the ideal classifier is only a fiction, a substitute is usually chosen from the large toolbox of pattern recognition and machine learning. For some of these models, irrelevant features may spoil the performance (for example, the k-nearest neighbour (k-nn)). Second, the fact that the dataset is finite, precludes estimating the parameters of the classifier to arbitrarily precision. This in itself could contribute to the peaking effect. The peak identifies the optimal number of features for the chosen classifier model and feature selection procedure. Note, therefore, that "peeking" is quite different from "peaking". "Peeking" is an oversight on the experimenter's part while "peaking" can be described as a data/model quirk.

[2] http://wok.mimas.ac.uk/

ACCEPTED MANUSCRIPT

2.3. The effect of the small sample size

Wide datasets with low sample size are typically too small to allow for a split into training and testing. Take for example The Great British Medallists Research Project which is an in-depth study of 32 former GB athletes from Olympic sports[3] Selecting the most important traits and practices may inform further training and selection decisions for boosting the performances of elite athletes. The dataset limits come from the fact that there are simply no more instances to add. Nonetheless, the hold-out protocol where relatively small-size data sets are split randomly into a training and a testing part is still used in feature selection [31].

The problem of an inadequately small sample size has been flagged in the past [32]. However, here we are interested in extreme cases of very small-size data, which have not been considered before.

While concerns have been raised before, to the best of our knowledge, there is no comprehensive experimental study which clearly demonstrates the extent of the problem of overfitting in feature selection for very low-sample-size data. To illustrate this point, we replicated results due to Raunannen, 2003 [15]. The problem of the overfitting has been aptly exemplified by a sequential forward selection (SFS) on the 'sonar' data from the UCI repository [33] (2 classes containing respectively 97 and 111 instances, and 60 features). Half of the data was used for training, and the other half for testing. The feature selection was carried out through the leave-one-out cross-validation protocol (LOO) on the training part of the data. The nearest neighbour classifier (1-nn) was used as the classifier of choice in the wrapper approach. Thus the accuracy of the classifier with the selected feature subset is directly the output from the SFS procedure. The "proper" testing accuracy was subsequently estimated on the testing data for all feature set sizes. Figure 1 shows the training (LOO) accuracy and the testing classification accuracy for 10 splits into halves, and the accuracy averaged across the 10 splits. The axes are formatted to match exactly Figure 1 in the original paper.

Both curves match the ones in the original paper. We further carried out experiments where instead of 50% (104 instances), the training data contained 20% (42 instances) and 10% (21 instances) of the data. Again, 10 runs with different random splits into training and testing were carried out, and the accuracy curves were averaged across the 10 runs. Figure 2 shows the averaged accuracies as functions of the cardinality of the feature set.

To highlight the severity of the problem, we showed the discrepancy between predicted and actual accuracy by joining the corresponding values for 50% split of the data. The figure shows that the gap between these accuracies increases dramatically for smaller training sizes considered here.

Here we examine experimentally the inadequacy of the flawed protocol and propose an alternative.

[3] http://ipep.bangor.ac.uk/medalists_research.php.en

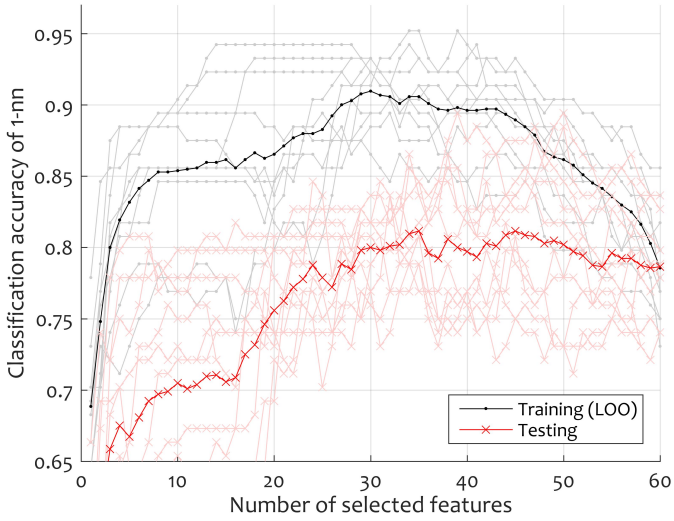

Figure 1: Results for the 'sonar' data set replicating the SFS illustrative experiment of Reunanen, 2003 [15]. The subsets of features were obtained from SFS. The leave-one-out accuracy of 1-nn was used as the as the feature subset evaluation criterion.

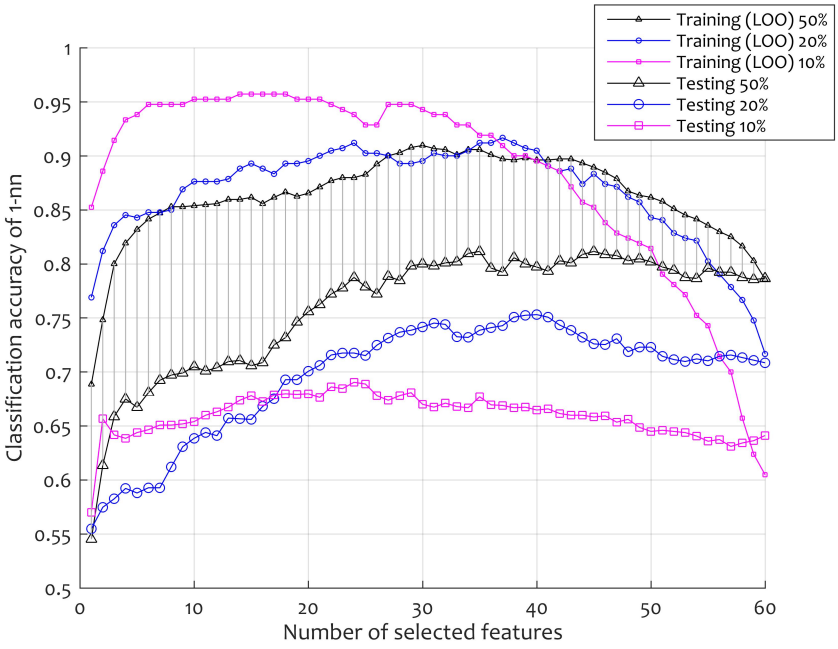

Figure 2: Comparison of accuracies for different sample sizes for the 'sonar' data set. Small markers show the training (predicted) accuracy, and large markers, the testing (actual) accuracy of the 1-nn classifier. The subsets of features were obtained from SFS. The leave-one-out accuracy of 1-nn was used as the as the feature subset evaluation criterion.

ACCEPTED MANUSCRIPT

3. Methods

3.1. Feature selection approaches and their applicability to small-sample-size data

While the field abounds with feature selection methods, little will be suitable for the wide datasets considered here. The way of traversing the possible candidate subsets of features will be no different from the ways adopted in the conventional approaches. However, the criterion for evaluating these subsets must be chosen with caution. Consider the three established approaches: wrapper, filter, and embedded selection [34]. In the wrapper approach, a classifier is trained using the candidate subset of features, and a discrimination measure (usually the classification accuracy) is subsequently calculated. The filter approach, on the other hand, uses a proxy for the desired discrimination measure, and avoids training a classifier. While in the former two approaches the classifier model is not directly responsible for selecting or ranking the features, some classifier models allow for this combined process (embedded approach). Examples of such models are the decision tree classifier, the linear SVM classifier, and the random forest classifier ensemble [35].

It is universally accepted that wrapper methods give better results than filter or embedded methods. For wide datasets, however, the drawbacks of the wrapper approach are amplified into major flaws. The first flaw is the lack of fidelity. In a dataset with N objects, a leave-one-out (N-fold) cross-validation will give only $N + 1$ possible distinct values for the accuracy. The feature sets of interest will likely take an even more limited set of values corresponding to the higher spectrum of the accuracies. Thus, they may not be distinguishable from one another or from other, less valuable, feature sets. The second flaw is the increased risk of overfitting compared to the filter or the embedded approach. Thus, we propose to use the state-ot-the-art filter and embedded methods for evaluating the candidate subsets for extreme wide datasets. In the experiments further on, we apply the Fast Correlation-Based Filter (FCBF) [6], ReliefF [36, 37], and the Symmetrical Uncertainty [38].

3.2. The right and the wrong protocols

Here we argue the main point of this study. A remarkably large number of studies in feature selection, including some quite influential ones, use a flawed (contaminated) protocol, which openly or subtly includes peeking. This protocol is illustrated in Figure 3. First, the feature selector $\mathcal{F}$ is applied to the data, and a set of features S is selected. Next, classifier models $\mathcal{C}$ are evaluated on the same data, possibly using cross-validation, and the best classifier is returned to the user along with the estimate of the classification accuracy from the cross-validation experiment, A_{LOO}.

The caveat here is that the dataset is used twice: once for finding S through $\mathcal{F}$, and once for evaluating $\mathcal{C}$. Thus the classifier's testing data have already been used for selecting S. Hence, a positive bias can be expected due to this "peeking". How can this be done without peeking? Figure 4 shows one possible answer in the form of a non-contaminated protocol, which will be called the "proper" protocol.

ACCEPTED MANUSCRIPT

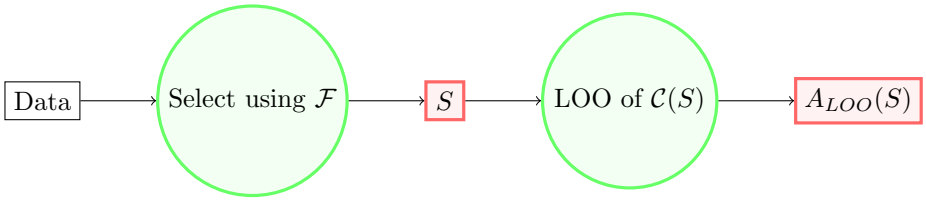

Figure 3: Diagram of the widely used but incorrect (contaminated) protocol for feature selection. Boxes represent inputs and outputs; shaded boxes represent output returned to the user; and circles represent procedures. S is the selected subset of features; $A_{LOO}(S)$ is the classification accuracy predicted through leave-one-out cross-validation for the chosen classifier $\mathcal{C}$, and $\mathcal{F}$ is the chosen feature selection method.

In the proper protocol, the cross-validation loop includes the feature selector $\mathcal{F}$. A feature set (or ranking) S_i is obtained for each cross-validation fold using the respective training data. Then the chosen classifier $\mathcal{C}(S_i)$ is trained on the same training data using the selected features. Finally the testing data for the fold is used to evaluate the accuracy of $\mathcal{C}(S_i)$. By averaging the accuracies for the cross-validation folds, we obtain one final value, A_{PRO}, which estimates the accuracy of the whole process (feature selection followed by classification). At no point in this training process is the testing data seen by the feature selector or classifier. At the end, the output returned to the user is the feature set S obtained from the whole dataset through $\mathcal{F}$. Interestingly, in most cases, this is the same set obtained from the wrong protocol. The difference is in the classification accuracy which accompanies this set. Our hypothesis is that, due to the peeking, A_{LOO} is optimistically biased, and therefore misleading, while A_{PRO} is closer to the true accuracy, which can be estimated from a previously unseen testing set.

3.3. An example of the optimistic bias of the wrong protocol

Figure 5 shows an example of the above argument. We chose the arrhythmia dataset from the UCI repository [33]. The data contains 279 features (attributes) and 452 objects (instances). We grouped the class labels into two classes: (1) normal (207 objects, 45.8%) and (2) arrhythmia (245 objects, 54.2%). Extreme wide datasets were sampled 100 times, with 10 objects in each class. We chose the ReliefF feature ranker as $\mathcal{F}$, and the linear discriminant classifier (LDC) with a diagonal covariance matrix and uniform priors as $\mathcal{C}$. A feature ranking was obtained for each of the 100 runs. The incorrect protocol illustrated in Figure 3 was applied to derive the predicted accuracy $A_{LOO}(S)$ for feature subsets of increasing cardinality, labelled 'LOO' in the figure. In this example, we set the maximum cardinality to 40% of the cardinality of the feature set. The "proper" protocol was applied as well, giving accuracy A_{PRO}, which is labelled as 'Proper' in the figure.

The accuracy of $\mathcal{C}$ trained on the whole wide dataset of 20 objects, A_T (labelled 'Test' in the figure), was evaluated using the remaining 432 objects left aside for testing. We treat this value as the desired quantity, which A_{LOO} and A_{PRO} strive to approximate. For comparison, for every run, we calculated the accuracy of

ACCEPTED MANUSCRIPT

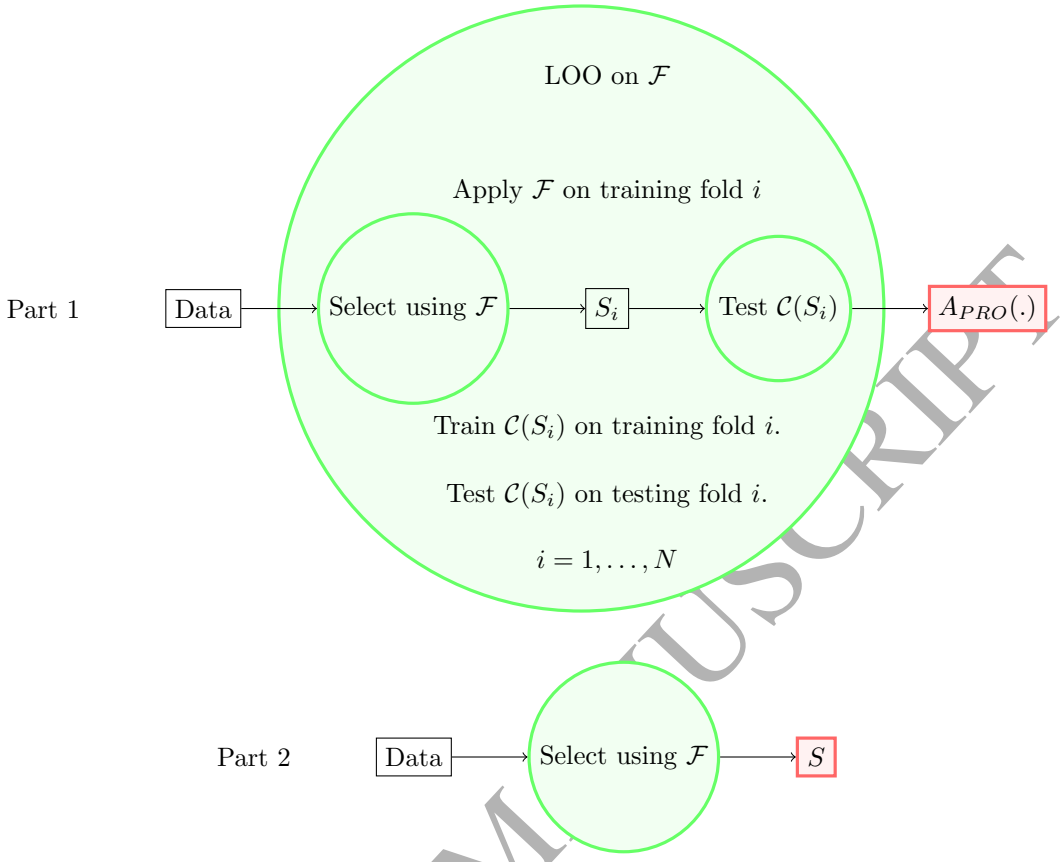

Figure 4: Diagram of the "proper" protocol for feature selection. Boxes represent inputs and outputs; shaded boxes represent output returned to the user; and circles represent procedures. S is the selected subset of features; $A_{PRO}(S)$ is the classification accuracy predicted through leave-one-out cross-validation for the chosen classifier $\mathcal{C}$, and $\mathcal{F}$ is the chosen feature selection method.

a random permutation of the features instead of the ranking offered by $\mathcal{F}$. As there is no selection method to cross-validate in the random approach, the "wrong" and the "proper" protocols both amount to evaluating the LOO accuracy using the training data, denoted R_{LOO}, and labelled as 'Random LOO' in the figure. Again, we subsequently calculated the quantity which R_{LOO} attempts to predict by evaluating a $\mathcal{C}$ trained on the whole training data (with the respective random subset of features) using the testing part of the data. This value, R_T, is labelled in the figure as 'Random Test'.

The graph shows exactly where the problem lies. We have shaded the gap between A_{LOO} and A_T in blue, and the gap between A_{PRO} and A_T in red. Clearly, A_{LOO} is heavily optimistically biased, whereas A_{PRO} is a lot closer to A_T. The large optimistic bias is caused by using the wrong protocol (peeking), which, unfortunately, is the standard practice in many studies, even very highly valued ones. However, A_{PRO} is not a perfect solution to this problem either. There is a visible pessimistic bias of A_{PRO}. One possible reason

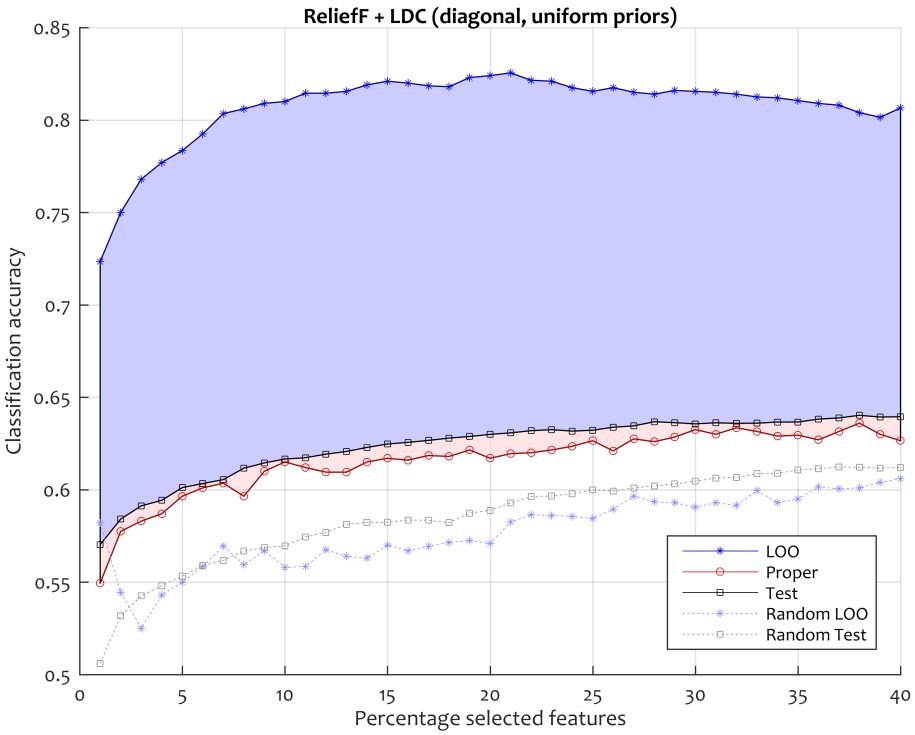

Figure 5: An example of the optimistic bias of the wrong feature selection protocol using the arrhythmia data from UCI.

for this bias is that when we evaluate $\mathcal{C}$ in Part 1 of the right protocol in Figure 4, the classifier is built on $N-1$ objects, and for the testing accuracy, we build the classifier on all N objects. Given that N is quite small, the difference of one object is noticeable, even for a stable classifier models as LDC. Still, we argue that this bias is smaller than the bias of A_{LOO}, and is better suited as a guarantee returned to the user.

The random curves, expectedly, run under the curves using a proper feature selector, showing lower classification accuracy. The argument why R_{LOO} is worse than R_T is the same as above. To obtain R_{LOO}, we train $\mathcal{C}$ on $N-1$ objects, and for R_T, on N objects.

The classification accuracy as a function of the number of features will not behave in the same way for all classifiers. There could be idiosyncratic pockets of features which perform excellently for a specific classifier and are largely overlooked by most other classifiers. The peak-effect may be strongly or less strongly pronounced depending on the classifier. The same argument holds for the feature selection method $\mathcal{F}$. There could be "lucky pairings" between $\mathcal{F}$ and $\mathcal{C}$ for the dataset of interest, giving high accuracy with fewer features, but this cannot be known in advance.

ACCEPTED MANUSCRIPT

4. Experimental study

The purpose of the experiment is to verify our hypothesis that the proper feature selection protocol gives a closer estimate of the testing accuracy than the widely used contaminated protocol. In addition, we will seek to answer the following questions:

1) Does the protocol choice have the same impact over different feature selection methods?

2) Does the protocol choice have the same impact over different classifier models?

4.1. Data

The characteristics of the 24 datasets used here are presented in Table 1. They were taken from the repository[4] [20]. Some of the datasets within the collection are from the UCI Machine Learning Repository [33].

4.2. Feature selectors and classifiers

The experiments for this part of the study were carried out in Weka [39]. We experimented with the following choices of feature selection methods $\mathcal{F}$ and classifier models $\mathcal{C}$ implemented in Weka:

Symmetrical Uncertainty [38] (SU) is a measure of correlation between two nominal features based on their individual and joint entropies. When one of the two features is the class variable, we have a measure of the worth of the paired feature. This measure can be used for ranking the features. It does not take into account any interaction between them. To apply this measure to continuous-valued features, they are first discretised.

Fast correlation-based filter (FCBF) [6] also uses SU. Unlike SU, however, it takes into account the correlation between the features. The method aims at selecting features which have high correlation with the class variable and low correlation among themselves.

ReliefF [37], a variant of Relief [36], is an instance-based feature ranking method. A subset of the instances is randomly selected multiple times and the feature weights are updated based on the proximity of the instances from the same classes in the selected sample.

The classifiers used were: the nearest neighbour (1-NN), the decision tree classifier (J48) [40], the linear discriminant classifier (LDC)[5], the naïve Bayes classifier (NB) and the Random Forest classifier ensemble (RF) [41].

[4] The repository is available at `http://featureselection.asu.edu/datasets.php`

[5] pseudo-linear LDA is implemented in Weka.

ACCEPTED MANUSCRIPT

Table 1: Characteristics of the high-dimensional datasets.

Dataset	Instances	Features	Classes
ALLAML	72	7129	2
arcene	200	10000	2
BASEHOCK	1993	4862	2
Carcinom	174	9182	11
CLL_SUB_111	111	11340	3
COIL20	1440	1024	20
colon	62	2000	2
gisette	7000	5000	2
GLI_85	85	22283	2
GLIOMA	50	4434	4
Isolet	1560	617	26
leukemia	72	7070	2
lung	203	3312	5
lymphoma	96	4026	9
madelon	2600	500	2
PCMAC	1943	3289	2
Prostate_GE	102	5966	2
RELATHE	1427	4322	2
SMK_CAN_187	187	19993	2
TOX_171	171	5748	4
USPS	9298	256	10
warpAR10P	130	2400	10
warpPIE10P	210	2420	10
Yale	165	1024	15

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

4.3. Calculation of the criteria values

We carried out 10 runs for each data set. For FCBF, the number of features was determined within the algorithm. For the ranker methods, the number of features was varied as $\{1, 2, ..., 9, 10, 15, 20, ...100\}$. In each run, the dataset was randomly split into a training part of $10 \times c$ instances for training, and the remaining instances for testing, where c is the number of classes. Denote the training part (10 instances per class) by $\mathcal{D}_{\text{train}}$, and the testing part by $\mathcal{D}_{\text{test}}$. For a chosen feature selection method $\mathcal{F}$ and a chosen classifier model $\mathcal{C}$ (for a fixed number of features or number determined by $\mathcal{F}$), we calculated the following criteria of interest:

- A_{LOO}. Apply $\mathcal{F}$ on $\mathcal{D}_{\text{train}}$ to obtain feature set S. Denote by $\mathcal{D}_{\text{train}}(S)$ the restriction of $\mathcal{D}_{\text{train}}$ on the feature subspace S. Evaluate $\mathcal{C}$ on $\mathcal{D}_{\text{train}}(S)$ using leave-one-out cross-validation. This accuracy is A_{LOO}.

- A_{PRO}. Organise a leave-one-out loop on $\mathcal{D}_{\text{train}}$. For each training fold, i, apply $\mathcal{F}$ (with or without cross-validation) to obtain feature set S_i. Test $\mathcal{C}$ using S_i on the remaining testing instance. The averaged accuracy on the testing instances is A_{PRO}.

- A_T. Apply $\mathcal{F}$ on $\mathcal{D}_{\text{train}}$ (with or without cross-validation) to obtain feature set S. (This is the same step as in calculating A_{LOO}.) Evaluate $\mathcal{C}$ on $\mathcal{D}_{\text{test}}(S)$. This accuracy is A_T.

4.4. Protocol and results

To enable statistical analyses, we need to determine a suitable number of features for the rankers. We tried two approaches:

- Maximum. For each run, identify the maximum of the curve A_{LOO} and store the smallest number for which this maximum is achieved, N_M. In the same way, determine N'_M, for which the curve of the proper protocol peaks.

- Parabola. Assuming that there is a peak effect as described in Section 2, we fit a parabola $y = ax^2 + bx + c$ (through least squares) to A_{LOO} and A_{PRO}. If the parabola is convex ($a < 0$), we return the position of the maximum $N_P = -\frac{b}{2a}$ (similarly for N'_P). If the parabola is concave, the Maximum method above is applied to determine N_P (N'_P).

In this way, we may have different feature sets and different cardinalities by LOO and Proper. Denote by A^Y_X the accuracy A_X measured for a feature set of cardinality Y. If our hypothesis is correct, $A^{N'}_{PRO}$ will be closer to $A^{N'}_T$ than A^N_{LOO} is to A^N_T. In other words, we would expect the following inequality to hold:

$$|A^{N'}_{PRO} - A^{N'}_T| < |A^N_{LOO} - A^N_T|. \quad (1)$$

ACCEPTED MANUSCRIPT

Tables 2–6 show the results for the individual data sets, and the feature selection methods: FCBF, ReliefF/Maximum, ReliefF/Parabola, Symmetric Uncertainty/Maximum, and Symmetric Uncertainty/Parabola. We show the classification accuracies A_{LOO}, A_{PRO}, and the respective A_T, averaged across the 10 runs and the 5 classifiers. Given in the tables are also the averaged cardinality of the selected feature subset, $|S|$, for each dataset. We denote the difference of interest by $\Delta_X = A_X - A_T$, where X stands for LOO or PRO. The columns with the differences are shown in boldface in the table. For each dataset, the smaller one of the two differences Δ – by absolute value – is shown in a box. Since the values of the classification accuracies are not commensurable across datasets, nor are the differences thereof, only the sign rank statistical test is applicable. The p-values from the sign test comparing the paired values of $|\Delta|$ are given in the respective table caption. For all feature selection methods, we found significant difference at level 0.01. This supports our hypothesis that the proper protocol gives closer estimates of the true accuracy compared to the peeking protocol for very small-size data.

Next we ran the sign test for the paired observations separately for each classifier and feature selection method. Each test was calculated from 240 pairs of values (24 data sets, 10 runs). We ran the right-tailed sign test with null hypothesis: $|A_{PRO} - A_T| \geq |A_{LOO} - A_T|$ (LOO is equivalent or better than the proper protocol). All p-values, with one exception, were under 0.00005, strongly rejecting the null hypothesis, thereby landing further support to our claim. The only relatively larger p-value of 0.0274, still under 0.05, was observed for the FCBF feature selector and the J48 classifier.

To examine further the effect of the classifier model on the differences between the predicted and true accuracy, we plot in Figure 6 $|A_{LOO} - A_T|$ versus $|A_{PRO} - A_T|$ for the five classifier models. Each point on the plot comes from one run, a given feature selection method and the classifier specified in the title of the sub-figure. Thus, each plot contains 24 datasets $\times$ 10 runs $\times$ 5 feature selection methods/variants = 1,200 points. Out of these, we calculated the percentage where $|A_{LOO} - A_T| > |A_{PRO} - A_T|$, supporting our hypothesis, shown as "win" (W) in the title of the sub-figure. All such points are above the diagonal line of the square. We also show in the title the percentage of draws (D), where $|A_{LOO} - A_T| = |A_{PRO} - A_T|$, and the percentage of losses (L), where $|A_{LOO} - A_T| < |A_{PRO} - A_T|$. It sign test is applied to any of the data subsets in the 5 sub-figures, the hypothesis in eqn (1) is strongly supported.

Finally, we illustrate the reduction of optimistic bias when using the correct protocol in Figure 7. We chose one example of feature selection method (ReliefF/Max) and classifier (Random Forest) but we note that all such plots look similar. A_{LOO}, A_{PRO} and the respective A_T are averaged across the 10 runs for each data set. A dot marker represents (A_T,A_{LOO}) for a given data set, and a triangle marker, represents (A_T,A_{PRO}). The markers for the same data sets are joined by an arrow from LOO to PRO. The downward tendency of the arrows shows the reduction of the optimistic bias by applying the correct protocol.

In summary, we confirm that using the proper protocol for feature selection from very wide datasets gives more truthful results compared to the currently favoured protocol, which we termed here "the wrong"

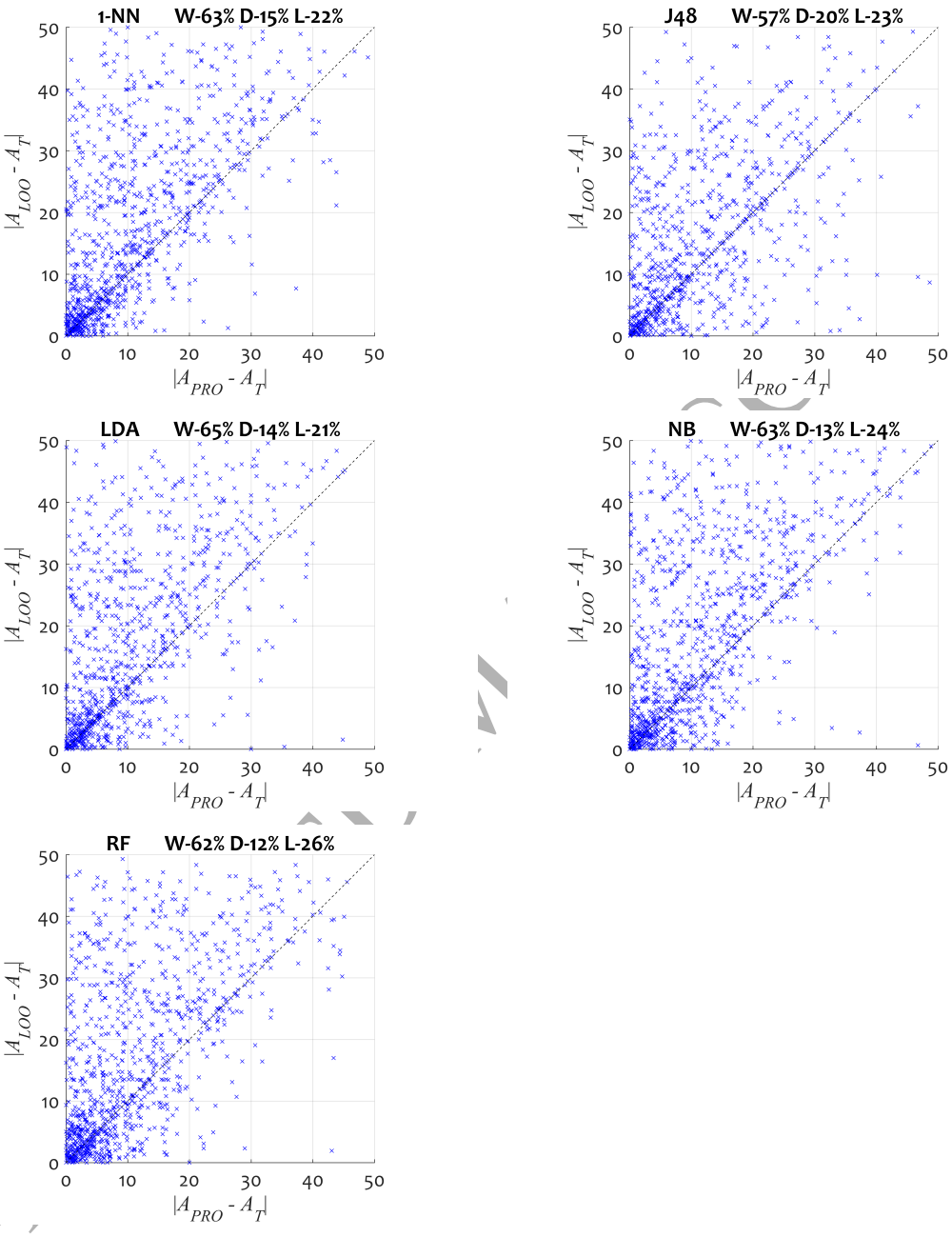

Figure 6: Scatteplot of $|A_{LOO} - A_T|$ versus $|A_{PRO} - A_T|$ the five classifier models. Each point on the plot comes from one run, a given feature selection method and the classifier specified in the title of the sub-figure. W/D/L mean win/draw/loss, where W is the percentage of points where (1) holds.

ACCEPTED MANUSCRIPT

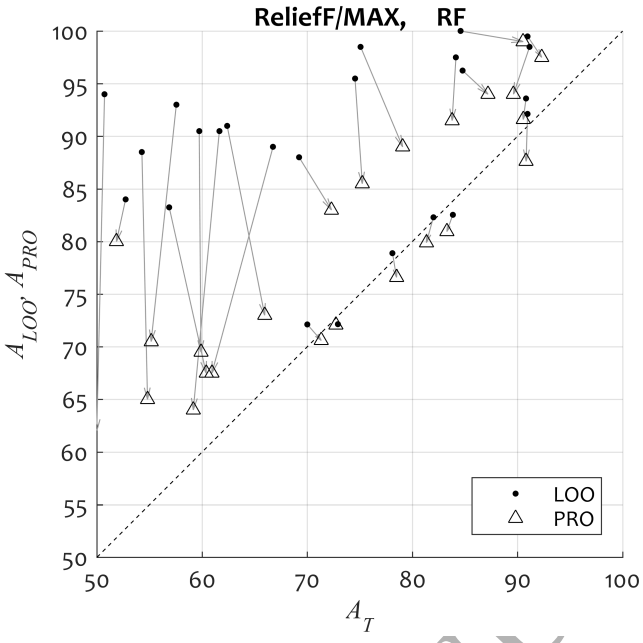

Figure 7: Scatterplot of A_{LOO} and A_{PRO} versus A_T. A dot marker represents (A_T,A_{LOO}) for a given data set, and a triangle marker, (A_T,A_{PRO}). The markers for the same data sets are joined by an arrow from LOO to PRO.

protocol or the "contaminated" protocol. Our results also suggest that the bias is likely universally present across many feature selection methods and classifier models.

5. Conclusions

This paper demonstrates the importance of applying a clean (non-contaminated) protocol for feature selection for wide datasets with a very low sample size. While the set of features returned to the user may be the same from both protocols, the estimate of the classification accuracy, which must be returned too, will likely be misleading if the wrong protocol is used. Running an experimental study with 24 datasets, we found statistically significant differences between the biases of the wrong and the proper protocols for all classifier models and feature selection methods we tested.

Based on these results, we recommend using the proper protocol (Figure 4) instead of the popular alternative (Figure 3).

Further on, the ranker methods, which are suitable for this type of data, need additional analysis for choosing the cardinality of the feature set to be returned. We examined two simple variants: maximum and parabola, and found that the conclusions applied to both. As a future line of research, we are planning to investigate other methods for determining the cardinality of the best feature subset using a stability index [42, 43]. Ensembles of ranker methods are also a good way forward [44] for very small-size data. In

ACCEPTED MANUSCRIPT

addition to a more stable ranking, they offer further possibilities to use stability for obtaining the cardinality of the returned feature subset. ~~Apart from the optimistic bias considered here, feature selection from wide datasets of very small sample sizes poses further problems. The selected feature set may not be better than a randomly picked feature set of the same cardinality. A good practice to verify the merit of the returned feature set would be to compare the estimated accuracy with that of randomly generated feature subsets of the same cardinality, similarly to random permutation tests.~~ Most importantly, one should seek to increase the sample size.

Acknowledgements

This work was supported by project RPG-2015-188 funded by The Leverhulme Trust, UK and by project TIN2015-67534-P (MINECO/FEDER, UE) funded by the *Ministerio de Economía y Competitividad* of the Spanish Government and European Union FEDER funds.

References

[1] E. A. Patrick, Fundamentals of Pattern Recognition, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1972.

[2] P. A. Devijver, J. Kittler, Pattern Recognition: A Statistical Approach, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1982.

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

[4] M. Dash, H. Liu, Feature selection for classification, Intelligent Data Analysis 1 (1997) 131–156.

[5] X. Zhu, Z. Huang, Y. Yang, H. Tao Shen, C. Xu, J. Luo, Self-taught dimensionality reduction on the high-dimensional small-sized data, Pattern Recognition 46 (1) (2013) 215–229.

[6] L. Yu, H. Liu, Feature selection for high-dimensional data: A fast correlation-based filter solution, in: In Proceedings of the 20th International Conference on Machine Learning (ICML2003), Washington, DC, 2003.

[7] J. Hua, W. D. Tembe, E. R. Dougherty, Performance of feature-selection methods in the classification of high-dimension data, Pattern Recognition 42 (3) (2009) 409–424.

[8] A. Golugula, G. Lee, A. Madabhushi, Evaluating feature selection strategies for high dimensional, small sample size datasets, in: IEEE International Conference of Engineering in Medicine and Biology Society (EMBS), 2011, pp. 949–952.

[9] P. Bermejo, L. de la Ossa, J. A. Gámez, J. M. Puerta, Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking, Knowledge-Based Systems 25 (1) (2012) 35–44.

[10] E. P. Xing, M. I. Jordan, R. M. Karp, Feature selection for high-dimensional genomic microarray data, in: Proceedings of the 18th International Conference on Machine Learning, (ICML2001), 2001, pp. 601–608.

[11] Y. Saeys, I. n. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics, Bioinformatics (Oxford, England) 23 (19) (2007) 2507–2517.

[12] G. Brown, A. Pocock, M. Zhao, M. Lujan, Conditional likelihood maximisation: A unifying framework for information theoretic feature selection, Journal of Machine Learning Research 13 (2012) 27–66.

[13] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Machine Learning 46 (2002) 389–422.

[14] F. De Martino, G. Valente, N. Staeren, J. Ashburner, R. G. a E. Formisano, Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns, NeuroImage 43 (1) (2008) 44–58.

ACCEPTED MANUSCRIPT

[15] J. Reunanen, Overfitting in making comparisons between variable selection methods, Journal of Machine Learning Research 3 (2003) 1371–1382.

[16] P. Smialowski, D. Frishman, S. Kramer, Pitfalls of supervised feature selection, Bioinformatics 26 (3) (2010) 440–443.

[17] S. Diciotti, S. Ciulli, M. Mascalchi, M. Giannelli, N. Toschi, The "peeking" effect in supervised feature selection on diffusion tensor imaging data, American Journal of NeuroradiologyarXiv:http://www.ajnr.org/content/early/2013/07/18/ajnr.A3685.full.pdf.
URL http://www.ajnr.org/content/early/2013/07/18/ajnr.A3685

[18] F. Pereira, T. Mitchell, M. Botvinick, Machine learning classifiers and fMRI: a tutorial overview, NeuroImage 45 (1, Supplement 1) (2009) S199 – S209.

[19] Y. Li, T. Li, H. Liu, Recent advances in feature selection and its applications, Knowledge and Information Systems 53 (3) (2017) 551–577. doi:10.1007/s10115-017-1059-8.

[20] J. Li, K. Cheng, S. Wang, F. Morstatter, T. Robert, J. Tang, H. Liu, Feature selection: A data perspective, arXiv:1601.07996.

[21] F. Viegas, L. Rocha, M. Gonçalves, F. Mourão, G. Sá, T. Salles, G. Andrade, I. Sandin, A genetic programming approach for feature selection in highly dimensional skewed data, Neurocomputing 273 (2018) 554–569. doi:10.1016/j.neucom.2017.08.050.
URL https://doi.org/10.1016/j.neucom.2017.08.050

[22] E. Hancer, B. Xue, M. Zhang, D. Karaboga, B. Akay, Pareto front feature selection based on artificial bee colony optimization, Information Sciences 422 (2018) 462–479. doi:10.1016/j.ins.2017.09.028.
URL https://doi.org/10.1016/j.ins.2017.09.028

[23] P. P. Kundu, S. Mitra, Feature selection through message passing, IEEE Transactions on Cybernetics 47 (12) (2017) 4356–4366.

[24] S. Solorio-Fernández, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, A new unsupervised spectral feature selection method for mixed data : A filter approach, Pattern Recognition 72 (2017) 314–326. doi:10.1016/j.patcog.2017.07.020.

[25] J. Izetta, P. F. Verdes, P. M. Granitto, Improved multiclass feature selection via list combination, Expert Systems With Applications 88 (2017) 205–216. doi:10.1016/j.eswa.2017.06.043.

[26] K. Yu, X. Wu, W. Ding, Y. Mu, H. Wang, Markov blanket feature selection using representative sets, IEEE Transactions on Neural Networks and Learning Systems 28 (11) (2017) 2775–2788.

[27] R. M. O. Cruz, R. Sabourin, G. D. C. Cavalcanti, META-DES.Oracle: Meta-learning and feature selection for dynamic ensemble selection, Information Fusion 38 (2017) 84–103. doi:10.1016/j.inffus.2017.02.010.

[28] J. M. V. Campenhout, On the peaking of the Hughes mean recognition accuracy: The resolution of an apparent paradox, IEEE Transactions on Systems, Man, and Cybernetics 8 (5) (1978) 390 – 395. doi:DOI:10.1109/TSMC.1978.4309980.

[29] A. K. Jain, D. Zongker, Feature selection: evaluation, application and small sample performance, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (2) (1997) 153–158.

[30] C. Sima, E. R. Dougherty, The peaking phenomenon in the presence of feature-selection, Pattern Recognition Letters 29 (11) (2008) 1667–1674. doi:10.1016/j.patrec.2008.04.010.
URL http://dx.doi.org/10.1016/j.patrec.2008.04.010

[31] B. Ghaddar, J. Naoum-sawaya, High dimensional data classification and feature selection using support vector machines, European Journal of Operational Research 265 (3) (2018) 993–1004. doi:10.1016/j.ejor.2017.08.040.
URL https://doi.org/10.1016/j.ejor.2017.08.040

[32] S. Raudys, A. Jain, Small sample size effects in statistical pattern recognition: Recommendations for practitioners and open problems, in: Proc. 10th Int. Conf. on Pattern Recognition, Atlantic City, New Jersey, 1990, pp. 417–423.

[33] K. Bache, M. Lichman, UCI machine learning repository (2013).

ACCEPTED MANUSCRIPT

URL http://archive.ics.uci.edu/ml

[34] R. Kohavi, G. John, Wrappers for feature subset selection, Artificial Intelligence Journal 97 (1997) 273–324.

[35] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, B. P. Feuston, Random forest: a classification and regression tool for compound classification and QSAR modeling, Journal of chemical information and computer sciences 43 (6) (2003) 1947–1958.

[36] K. Kira, L. A. Rendell, A practical approach to feature selection, in: Proceedings of the ninth international workshop on Machine learning, 1992, pp. 249–256.

[37] M. Robnik-Šikonja, I. Kononenko, Theoretical and empirical analysis of relieff and rrelieff, Machine Learning 53 (1) (2003) 23–69. doi:10.1023/A:1025667309714.

URL https://doi.org/10.1023/A:1025667309714

[38] M. A. Hall, G. Holmes, Benchmarking attribute selection techniques for discrete class data mining, IEEE Transactions on Knowledge and Data Engineering 15 (6) (2003) 1437–1447. doi:10.1109/TKDE.2003.1245283.

[39] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The WEKA data mining software: an update, SIGKDD Explor. Newsl. 11 (1) (2009) 10–18. doi:10.1145/1656274.1656278.

URL http://doi.acm.org/10.1145/1656274.1656278

[40] J. Quinlan, C4.5: Programs for Machine Learning, Morgan Kauffman, 1993.

[41] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32.

[42] L. Kuncheva, A stability index for feature selection, in: Proc. IASTED, Artificial Intelligence and Applications, Innsbruck, Austria, 2007, pp. 390–395.

[43] W. Altidor, T. Khoshgoftaar, A. Napolitano, A noise-based stability evaluation of threshold-based feature selection techniques, in: Information Reuse and Integration (IRI), 2011 IEEE International Conference on, IEEE, 2011, pp. 240–245.

[44] T. Abeel, T. Helleputte, Y. Van de Peer, P. Dupont, Y. Saeys, Robust biomarker identification for cancer diagnosis with ensemble feature selection methods, Bioinformatics (Oxford, England) 26 (3) (2010) 392–3988.

Appendix

ACCEPTED MANUSCRIPT

Table 2: FCBF feature selection method. Classification accuracies A_{LOO}, A_{PRO}, and the respective A_T, averaged across the 10 runs and the 5 classifiers. $|S|$ is the averaged cardinality of the selected feature subset. $\Delta_X = A_X - A_T$, where X stands for LOO or PRO. The columns with the differences are shown in boldface. For each dataset, the smaller one of the two differences Δ – by absolute value – is shown in a box. The p-value of the sign test for equivalence of Δ_{LOO} and Δ_{PRO} is 0.0066.

Dataset	$\lvert S\rvert$	A_{LOO}	A_T	Δ_{LOO}	$\lvert S\rvert$	A_{PRO}	A_T	Δ_{PRO}
ALLAML	(2.0)	92.3	80.2	**12.1**	(2.0)	80.2	80.2	**-0.0**
BASEHOCK	(5.8)	77.6	65.0	**12.6**	(5.8)	57.2	65.0	**-7.8**
CLL_SUB_111	(44.6)	88.3	55.6	**32.7**	(44.6)	66.5	55.6	**10.9**
COIL20	(101.4)	84.1	85.0	**-0.9**	(101.4)	83.0	85.0	**-2.0**
Carcinom	(192.7)	88.7	86.8	**2.0**	(192.7)	81.6	86.8	**-5.2**
GLIOMA	(37.3)	87.7	67.5	**20.2**	(37.3)	68.1	67.5	**0.6**
GLI_85	(2.0)	94.9	70.6	**24.3**	(2.0)	65.9	70.6	**-4.7**
Isolet	(17.0)	68.2	67.5	**0.7**	(17.0)	64.9	67.5	**-2.7**
PCMAC	(4.8)	71.4	57.6	**13.8**	(4.8)	50.9	57.6	**-6.7**
Prostate_GE	(4.6)	95.1	78.3	**16.8**	(4.6)	77.8	78.3	**-0.5**
RELATHE	(6.5)	80.6	54.2	**26.4**	(6.5)	55.7	54.2	**1.5**
SMK_CAN_187	(13.3)	86.8	54.0	**32.8**	(13.3)	60.3	54.0	**6.3**
TOX_171	(40.1)	77.5	54.2	**23.3**	(40.1)	52.8	54.2	**-1.5**
USPS	(17.6)	71.1	73.0	**-1.9**	(17.6)	68.2	73.0	**-4.7**
Yale	(18.2)	63.3	56.7	**6.6**	(18.2)	58.6	56.7	**2.0**
arcene	(21.5)	84.3	54.7	**29.6**	(21.5)	51.3	54.7	**-3.4**
colon	(12.8)	86.1	62.6	**23.5**	(12.8)	65.8	62.6	**3.2**
gisette	(17.0)	91.0	70.5	**20.5**	(17.0)	67.5	70.5	**-3.0**
leukemia	(8.0)	96.2	87.2	**9.0**	(8.0)	79.2	87.2	**-8.0**
lung	(68.6)	93.9	84.8	**9.1**	(68.6)	87.0	84.8	**2.2**
lymphoma	(46.6)	97.9	91.5	**6.4**	(46.6)	90.0	91.5	**-1.5**
madelon	(8.6)	81.1	49.7	**31.4**	(8.6)	35.2	49.7	**-14.5**
warpAR10P	(23.4)	81.5	80.5	**1.1**	(23.4)	75.4	80.5	**-5.0**
warpPIE10P	(48.5)	92.5	91.3	**1.2**	(48.5)	90.7	91.3	**-0.6**

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Table 3: ReliefF feature selection method, MAXIMUM version. Classification accuracies A_{LOO}, A_{PRO}, and the respective A_T, averaged across the 10 runs and the 5 classifiers. $|S|$ is the averaged cardinality of the selected feature subset. $\Delta_X = A_X - A_T$, where X stands for LOO or PRO. The columns with the differences are shown in boldface. For each dataset, the smaller one of the two differences Δ – by absolute value – is shown in a box. The p-value of the sign test for equivalence of Δ_{LOO} and Δ_{PRO} is 0.0015.

Dataset	$\lvert S\rvert$	A_{LOO}	A_T	Δ_{LOO}	$\lvert S\rvert$	A_{PRO}	A_T	Δ_{PRO}
ALLAML	(15.2)	97.6	86.9	**10.7**	(19.2)	93.9	87.5	**6.4**
BASEHOCK	(21.6)	87.2	62.5	**24.7**	(8.2)	69.3	60.3	**9.0**
CLL_SUB_111	(30.0)	84.1	52.1	**32.0**	(22.3)	78.1	51.1	**27.1**
COIL20	(92.5)	72.1	73.0	**-0.9**	(87.7)	70.9	72.2	**-1.3**
Carcinom	(69.5)	85.3	82.7	**2.6**	(69.3)	80.8	82.6	**-1.8**
GLIOMA	(35.7)	87.5	69.1	**18.5**	(27.3)	80.5	68.2	**12.4**
GLI_85	(7.7)	96.9	75.1	**21.8**	(19.7)	86.6	77.0	**9.6**
Isolet	(89.4)	67.4	67.4	**-0.0**	(90.1)	66.9	67.3	**-0.4**
PCMAC	(20.0)	88.4	59.6	**28.8**	(15.0)	65.0	58.1	**6.9**
Prostate_GE	(16.5)	96.5	79.6	**16.9**	(12.4)	89.9	80.6	**9.3**
RELATHE	(17.0)	88.8	54.3	**34.5**	(29.3)	64.1	54.4	**9.7**
SMK_CAN_187	(13.6)	93.9	57.6	**36.3**	(18.0)	72.6	56.7	**15.9**
TOX_171	(43.2)	79.8	56.2	**23.6**	(51.2)	65.8	57.8	**8.0**
USPS	(67.0)	72.6	72.0	**0.6**	(66.1)	70.5	71.9	**-1.4**
Yale	(68.4)	63.2	59.7	**3.5**	(70.0)	60.4	60.8	**-0.4**
arcene	(19.1)	90.8	58.0	**32.8**	(26.0)	69.3	56.7	**12.6**
colon	(7.9)	90.8	62.8	**28.0**	(11.7)	75.5	64.3	**11.2**
gisette	(15.4)	95.0	72.9	**22.1**	(24.6)	83.0	74.1	**8.9**
leukemia	(12.5)	99.2	89.8	**9.4**	(15.8)	95.8	89.7	**6.1**
lung	(44.9)	92.8	81.2	**11.5**	(57.0)	89.7	82.5	**7.2**
lymphoma	(16.6)	99.1	77.5	**21.6**	(39.3)	94.7	77.8	**16.9**
madelon	(27.8)	93.5	50.8	**42.7**	(18.9)	66.4	50.1	**16.3**
warpAR10P	(62.9)	77.6	77.4	**0.2**	(60.2)	76.2	76.5	**-0.2**
warpPIE10P	(54.9)	89.9	84.9	**4.9**	(51.9)	87.8	85.3	**2.6**

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Table 4: ReliefF feature selection method, PARABOLA version. Classification accuracies A_{LOO}, A_{PRO}, and the respective A_T, averaged across the 10 runs and the 5 classifiers. $|S|$ is the averaged cardinality of the selected feature subset. $\Delta_X = A_X - A_T$, where X stands for *LOO* or *PRO*. The columns with the differences are shown in boldface. For each dataset, the smaller one of the two differences Δ – by absolute value – is shown in a box. The p-value of the sign test for equivalence of Δ_{LOO} and Δ_{PRO} is 0.0003.

Dataset	$\lvert S\rvert$	A_{LOO}	A_T	Δ_{LOO}	$\lvert S\rvert$	A_{PRO}	A_T	Δ_{PRO}
ALLAML	(43.5)	96.9	90.0	**6.9**	(47.5)	95.9	89.5	**6.4**
BASEHOCK	(33.8)	84.7	63.6	**21.1**	(15.4)	73.6	61.1	**12.5**
CLL_SUB_111	(47.9)	79.1	51.9	**27.2**	(41.0)	76.7	51.1	**25.5**
COIL20	(85.6)	69.6	71.6	**-2.0**	(85.6)	69.7	71.7	**-2.0**
Carcinom	(67.1)	79.8	81.7	**-1.9**	(71.8)	79.3	81.5	**-2.2**
GLIOMA	(60.4)	81.7	68.0	**13.7**	(53.8)	81.9	68.2	**13.8**
GLI_85	(32.9)	96.3	78.1	**18.2**	(40.8)	91.0	77.5	**13.5**
Isolet	(86.7)	65.4	66.9	**-1.5**	(88.8)	65.7	67.1	**-1.4**
PCMAC	(39.1)	84.2	59.7	**24.5**	(26.3)	72.0	58.2	**13.8**
Prostate_GE	(31.5)	95.1	80.4	**14.7**	(30.7)	90.1	80.0	**10.1**
RELATHE	(34.3)	84.5	54.0	**30.5**	(37.9)	68.9	54.6	**14.3**
SMK_CAN_187	(31.9)	91.6	58.2	**33.4**	(32.9)	78.7	56.5	**22.2**
TOX_171	(60.3)	73.0	57.1	**15.9**	(66.3)	70.3	59.0	**11.3**
USPS	(76.6)	69.2	72.1	**-2.9**	(77.6)	69.1	72.1	**-3.0**
Yale	(66.2)	59.2	61.3	**-2.1**	(72.4)	59.4	61.3	**-1.9**
arcene	(37.8)	88.4	58.2	**30.2**	(38.2)	77.5	57.0	**20.5**
colon	(17.7)	88.8	63.2	**25.6**	(23.7)	78.2	64.1	**14.1**
gisette	(32.8)	93.9	74.2	**19.7**	(31.9)	84.0	74.0	**10.0**
leukemia	(25.2)	99.0	90.0	**9.0**	(28.6)	96.4	90.2	**6.2**
lung	(71.2)	89.0	84.8	**4.3**	(75.2)	88.7	85.8	**2.8**
lymphoma	(56.1)	98.5	83.2	**15.3**	(64.9)	97.0	83.8	**13.2**
madelon	(51.6)	90.9	50.8	**40.1**	(27.3)	74.3	50.0	**24.3**
warpAR10P	(67.2)	75.1	77.3	**-2.1**	(69.8)	75.4	77.5	**-2.0**
warpPIE10P	(58.1)	87.5	85.2	**2.3**	(59.0)	87.3	85.6	**1.7**

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Table 5: Symmetrical Uncertainty feature selection method, MAXIMUM version. Classification accuracies A_{LOO}, A_{PRO}, and the respective A_T, averaged across the 10 runs and the 5 classifiers. $|S|$ is the averaged cardinality of the selected feature subset. $\Delta_X = A_X - A_T$, where X stands for LOO or PRO. The columns with the differences are shown in boldface. For each dataset, the smaller one of the two differences Δ – by absolute value – is shown in a box. The p-value of the sign test for equivalence of Δ_{LOO} and Δ_{PRO} is 0.0003.

Dataset	$\|S\|$	A_{LOO}	A_T	Δ_{LOO}	$\|S\|$	A_{PRO}	A_T	Δ_{PRO}
ALLAML	(9.6)	98.6	85.5	**13.1**	(21.8)	92.6	87.0	**5.6**
BASEHOCK	(8.7)	84.1	64.9	**19.2**	(13.4)	70.1	63.2	**6.9**
CLL_SUB_111	(18.8)	87.1	53.3	**33.7**	(28.0)	75.7	54.2	**21.5**
COIL20	(81.9)	79.0	79.2	**-0.2**	(83.7)	77.8	79.1	**-1.3**
Carcinom	(56.4)	88.0	85.2	**2.8**	(61.9)	82.5	85.0	**-2.5**
GLIOMA	(28.5)	89.3	63.8	**25.4**	(34.9)	80.3	65.8	**14.4**
GLI_85	(3.2)	98.8	72.4	**26.4**	(22.9)	83.8	75.2	**8.6**
Isolet	(86.6)	68.6	68.4	**0.2**	(86.7)	67.8	68.4	**-0.6**
PCMAC	(10.9)	80.4	59.2	**21.2**	(5.1)	69.0	60.7	**8.3**
Prostate_GE	(6.0)	98.5	81.2	**17.3**	(18.0)	89.4	81.7	**7.7**
RELATHE	(13.2)	87.6	52.9	**34.7**	(9.7)	67.0	53.0	**14.0**
SMK_CAN_187	(12.4)	95.6	55.3	**40.3**	(9.9)	73.8	55.4	**18.4**
TOX_171	(37.7)	80.3	55.8	**24.5**	(44.3)	66.0	55.3	**10.7**
USPS	(76.4)	65.5	65.4	**0.1**	(77.4)	63.4	65.1	**-1.7**
Yale	(49.0)	67.6	61.3	**6.2**	(51.4)	63.7	59.1	**4.7**
arcene	(17.2)	93.8	56.1	**37.7**	(24.1)	67.4	56.3	**11.1**
colon	(7.1)	92.3	63.2	**29.1**	(25.3)	74.7	61.3	**13.4**
gisette	(20.6)	96.1	74.2	**21.9**	(24.4)	82.2	73.9	**8.3**
leukemia	(6.5)	99.4	87.7	**11.7**	(17.6)	92.8	90.1	**2.7**
lung	(38.6)	95.8	81.2	**14.6**	(51.1)	90.0	81.9	**8.2**
lymphoma	(23.2)	99.5	80.0	**19.5**	(45.3)	93.8	82.3	**11.5**
madelon	(15.1)	92.0	50.1	**41.9**	(22.9)	59.8	50.1	**9.7**
warpAR10P	(64.5)	78.9	77.3	**1.6**	(61.6)	76.8	77.6	**-0.8**
warpPIE10P	(68.2)	90.3	87.2	**3.0**	(69.4)	88.9	86.7	**2.2**

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Table 6: Symmetrical Uncertainty feature selection method, PARABOLA version. Classification accuracies A_{LOO}, A_{PRO}, and the respective A_T, averaged across the 10 runs and the 5 classifiers. $|S|$ is the averaged cardinality of the selected feature subset. $\Delta_X = A_X - A_T$, where X stands for LOO or PRO. The columns with the differences are shown in boldface. For each dataset, the smaller one of the two differences Δ – by absolute value – is shown in a box. The p-value of the sign test for equivalence of Δ_{LOO} and Δ_{PRO} is 0.0015.

Dataset	$\lvert S\rvert$	A_{LOO}	A_T	Δ_{LOO}	$\lvert S\rvert$	A_{PRO}	A_T	Δ_{PRO}
ALLAML	(32.2)	98.2	87.5	**10.7**	(48.2)	93.9	87.7	**6.2**
BASEHOCK	(28.0)	81.3	64.6	**16.7**	(33.8)	76.1	64.4	**11.7**
CLL_SUB_111	(22.3)	85.4	53.5	**31.9**	(49.6)	75.9	54.9	**21.0**
COIL20	(73.4)	76.7	78.3	**-1.6**	(72.9)	76.3	78.4	**-2.1**
Carcinom	(62.0)	85.3	85.6	**-0.3**	(67.2)	84.4	85.8	**-1.5**
GLIOMA	(39.9)	85.7	66.8	**18.9**	(52.0)	82.7	69.1	**13.7**
GLI_85	(15.0)	98.7	75.6	**23.1**	(30.5)	86.4	75.5	**10.9**
Isolet	(90.3)	67.0	68.2	**-1.2**	(92.6)	67.1	68.4	**-1.3**
PCMAC	(25.2)	79.0	59.1	**19.9**	(14.1)	71.7	60.6	**11.1**
Prostate_GE	(18.5)	97.3	81.1	**16.2**	(38.6)	91.6	81.7	**9.9**
RELATHE	(28.4)	84.1	53.4	**30.7**	(34.3)	76.5	52.9	**23.6**
SMK_CAN_187	(20.7)	94.7	55.8	**38.9**	(14.9)	78.1	55.7	**22.4**
TOX_171	(60.1)	75.1	58.0	**17.1**	(63.4)	71.2	57.8	**13.4**
USPS	(83.1)	62.9	65.1	**-2.1**	(85.5)	62.6	65.2	**-2.6**
Yale	(56.9)	63.6	61.1	**2.5**	(56.8)	63.8	61.3	**2.4**
arcene	(33.5)	91.9	57.4	**34.5**	(41.4)	76.6	57.2	**19.4**
colon	(14.7)	90.8	62.0	**28.8**	(41.0)	79.3	61.6	**17.7**
gisette	(33.2)	95.0	74.6	**20.4**	(34.8)	84.2	73.8	**10.4**
leukemia	(19.4)	99.2	88.6	**10.6**	(36.2)	94.5	90.5	**4.0**
lung	(61.8)	94.0	82.8	**11.1**	(70.4)	91.8	83.2	**8.6**
lymphoma	(59.0)	98.2	85.6	**12.6**	(69.0)	97.5	85.5	**12.0**
madelon	(21.1)	85.9	50.4	**35.5**	(35.7)	69.2	50.1	**19.1**
warpAR10P	(60.6)	76.3	78.7	**-2.4**	(63.0)	76.6	78.5	**-1.9**
warpPIE10P	(67.7)	88.3	87.0	**1.3**	(68.3)	88.0	87.1	**0.9**

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Ludmila (Lucy) I. Kuncheva is a Professor of Computer Science at Bangor University, UK. Her interests include pattern recognition, and specifically classifier ensembles. She has published two monographs and over 150 research papers. Lucy has won two Best Paper Awards (2006 IEEE TFS and 2003 IEEE TSMC.) She is a Fellow of IAPR.

Juan J. Rodríguez is an Associate Professor of Computer Science at Universidad de Burgos, Spain. His interests include data science, pattern recognition and specifically classifier ensembles.

ACCEPTED MANUSCRIPT