On the Over-Fitting Problem of Complex Feature Selection Methods

Petr Somol and Jana Novovičová
Department of Pattern Recognition
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic
CZ18208 Prague 8
Email: {somol, novovic}@utia.cas.cz

Pavel Pudil
Faculty of Management
Prague University of Economics
Jarošovská 1117/II
CZ37701, Jindřichův Hradec
Email: pudil@fm.vse.cz

***Abstract*—One of the hot topics discussed recently in relation to machine learning is the question of actual performance of modern feature selection methods. Feature selection has been a highly active area of research in recent years due to its potential to improve both the performance and economy of automatic decision systems in various applicational fields, including medicine, image analysis, remote sensing, economics etc. The number of available methods and methodologies has grown rapidly throughout recent years while promising important improvements. Yet recently many authors put this development in question, claiming that simpler older tools are actually better than complex modern ones – which, despite promises, are claimed to actually fail in real-world applications. We investigate this question, show several illustrative examples and draw several conclusions and recommendations regarding feature selection methods' expectable performance.**

I. Introduction

Dimensionality reduction (DR) concerns with the task of finding low dimensional representation for high dimensional data. DR is an important step in data preprocessing in machine learning and pattern recognition applications.

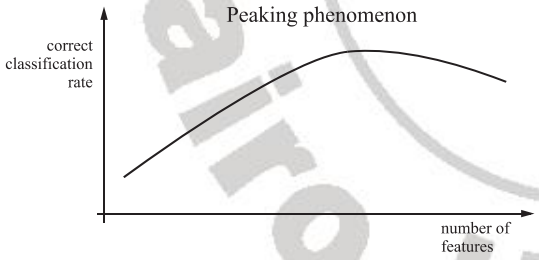

Fig. 1. With finite training data inclusion of features improves classification performance up to a certain point only

It is sometimes the case that such tasks as classification of the data represented by so called feature vectors, can be carried out in the reduced space more accurately than in the original space. In general, the decision rule in classifier must be estimated from a set of finite (usually small) number of training samples. If the number of training samples is small, problems are commonly manifested due to the so-called *peaking phenomenon* [1] (see Figure 1). This phenomenon concerns the dependence of the probability of correct recognition of patterns outside the training set and the number of features used. Initially the performance improves as new features are added, but at some point inclusion of further features may result in an actual degradation in performance. As a consequence, it is possible to improve the accuracy of the classifier's performance by deleting a feature [2].

There are two main ways of doing DR depending on the resulting features: DR by *feature selection* (FS) and DR by *feature extraction* (FE). The FS approach does not attempt to generate new features, but tries to select the "best" ones from the original set of features. The FE approach defines a new feature vector space in which each new feature is obtained by transformations of the original features. FS leads to savings in measurement cost and the selected features retain their original physical interpretation, important e.g., in medical applications. On the other hand, transformed features generated by FE may provide a better discriminative ability than the best subset of given features, but these new features may not have a clear physical meaning. A typical feature selection process consists of four basic steps: *feature subset selection*, *feature subset evaluation*, *stopping criterion*, and *result validation*. Based on the *selection criterion* choice, feature selection methods may roughly be divided into four types: the *filter* [3], [4], the *wrapper* [5], the *embedded* [6], but also [7] or [8], [9] and the *hybrid* [10], [11], [12]. The filter model relies on general characteristics of the data to evaluate and select feature subsets without involving any mining algorithm. The wrapper model requires one predetermined mining algorithm and uses its performance as the evaluation criterion. It attempts to find features better suited to the mining algorithm aiming to improve mining performance. This approach tends to be more computationally expensive than the filter approach. The embedded approach integrates the feature selection process into the model estimation process. Devising model and selecting features is thus one inseparable learning process, that may be looked upon as a special form of wrappers. Embedded methods thus offer performance competitive to wrappers, enable faster learning process, but produce results tightly coupled with particular model. The hybrid aims at combining the advantages of more than one of the listed approaches.

Certainly many key questions in FS remain unanswered and key problems remain unsolved to satisfaction. For example,

not enough is known about error bounds of many popular feature selection criteria, especially about their relation to classifier generalization performance. Despite the huge number of methods in existence, it is still a very hard problem to perform FS satisfactorily, e.g., in the context of gene expression [13] data, with enormous dimensionality and very few samples. Similarly, in text categorization [14] the standard way of FS is to completely omit context information and to resort to much more limited FS based on individual feature evaluation. In medicine these problems tend to become emphasized, as the available datasets are often incomplete (missing feature values in sample vectors), continuous and categorical data is to be treated at once, and the notion of feature itself may be difficult to interpret.

Among many criticisms of the current FS development there is one targeted specifically at the effort of finding more effective search methods, capable of yielding results closer to optimum with respect to some chosen criterion. The key argument against such methods is their often observed tendency to "over-select" features [15], or to find feature subsets fitted too tightly to training data, what degrades generalization. In other words, more search-effective methods are supposed to cause a similar unwanted effect as classifier over-training. Indeed, this is a serious problem that requires attention.

In recent literature the problem of "over-effective" FS has been addressed many times [16], [15]. Yet, the effort to point out the problem (which seems to have been ignored, or at least insufficiently addressed before) now seems to have led to the other extreme notion of claiming that most of FS method development is actually contra-productive. This is, that older simpler methods are actually superior to newer methods, mainly due to better over-fitting resistance.

The purpose of this paper is to discuss the issue of comparing actual FS methods' performance and to show experimentally what impact of the more effective search in newer methods can be expected.

A. FS Methods Overview

Before giving overview of the main methods to be discussed further we should note that it is not generally agreed in literature what the term "FS method" does actually describe. The term "FS method" is equally often used to refer to a) the complete framework that includes everything needed to select features, or b) the combination of search procedure and criterion or c) just the bare search procedure. In the following we will focus mainly on comparing the standard search procedures, which are not criterion- or classifier dependent. The widely known representatives of such "FS methods" are:

- Best Individual Features (BIF) [17],
- Sequential Forward Selection (SFS), Sequential Backward Selection (SBS), [18],
- "Plus l-take away r" Selection (+L-R) [18],
- Sequential Forward Floating Selection (SFFS), Sequential Backward Floating Selection (SFBS) [19],
- Oscillating Search (OS) [20].

Many other methods exist (in all senses of the term "FS Method"), among others the generalized versions of the ones listed above, various randomized methods, methods related to use of specific tools (FS for Support Vector Machines, FS for Neural Networks) etc. For overview see, e.g., [17], [21]. The selection of methods we are going to investigate is motivated by their interchangeability – any one of them can be used with the same given criterion, data and classifier. This makes experimental comparison easier.

II. Performance Estimation Problem

FS methods comparison seems to be understood ambiguously as well. It is very different whether we compare concrete method properties or the final classifier performance determined by use of particular methods under particular settings. Certainly, final classifier performance (preferably on independent test data) is the ultimate quality measure. However, misleading conclusions about FS may be easily drawn when evaluating nothing else, as classifier performance depends on many more different aspects then just the actual FS method used. Nevertheless, in the following we will adapt classifier accuracy as the main means of FS method assessment.

There seems to be a general agreement in the literature that wrapper-based FS enables creation of more accurate classifiers than filter-based FS. This claim is nevertheless to be taken with caution, while using actual classifier accuracy as the FS criterion in wrapper-based FS may lead to the very negative effects mentioned above (over-training). At the same time the weaker relation of filter-based FS criterion functions to particular classifier accuracy may help better generalization. But these effects can be hardly judged before the building of classification system has actually been accomplished.

In the following we will focus only on wrapper-based FS. Wrapper-based FS can be accomplished (and accordingly its effect can be evaluated) using one of the following methods:

- Re-substitution – In each step of the FS algorithm all data is used both for classifier training and testing. This has been shown to produce strongly optimistically biased results.
- Data split – In each step of the FS algorithm the same part of the data is used for classifier training and the other part for testing. This is the correct way of classifier performance estimation, yet it is often not feasible due to insufficient size of available data or due to inability to prevent bias caused by unevenly distributed data in the dataset (e.g., it may be difficult to ensure that with two-modal data distribution the training set won't by coincidence represent one mode and the testing set the other mode)
- Cross-Validation (CV) – Training data is split to several parts. Then in each FS step a series of tests is performed, with all but one data part used for classifier training and the remaining part used for testing. The average classifier performance is then considered to be the result of FS criterion evaluation. Because in each test a different part of data is used for testing, all data is eventually utilized,

without actually testing the classifier on the same data on which it had been trained. This is significantly better than re-substitution.

- Leave-one-out – A special case of CV with just one sample left for testing in each data split. This is computationally more expensive, but better utilizes the data.
- Hold-Out (HO) – Training data is randomly sampled. In each FS step a series of tests is performed, with part of the training data randomly sampled for classifier training and another part randomly sampled for testing. The average classifier performance is then considered to be the result of FS criterion evaluation. Unlike CV this may avoid possible bias caused by deterministically and evenly splitted data, but requires possibly higher number of trials than CV.

A. Feature Selection Stability

To investigate the robustness of the FS process, i.e., its dependence on particular data sampling, we repeat each FS experiment 100 times on differently sampled part of the training data. We define two measures to be called *consistency* and *weighted consistency*, that expresses the stability, or robustness of FS method with respect to various data samplings.

Let $Y = \{f_1, f_2, \ldots, f_{|Y|}\}$ be the set of all features and let $\mathcal{S} = \{S_1, S_2, \ldots, S_n\}$ be a system of $n > 1$ feature subsets $\mathrm{S}_j = \{f_{k_i} |\ i = 1, \ldots, d_j, f_{k_i} \in \mathrm{Y}, d_j \in \{1, \ldots, |\mathrm{Y}|\}\}, j = 1, \ldots, n,\ n > 1, n \in \mathbb{N}$. Denote $\mathcal{F}_f$ the system of subsets in $\mathcal{S}$ containing feature f, i.e.,

$$\mathcal{F}_f = \{S | S \in \mathcal{S}, f \in S\}. \tag{1}$$

Let F_f be the number of subsets in $\mathcal{F}_f$ and X the subset of Y representing all features that appear anywhere in system $\mathcal{S}$, i.e.,

$$X = \{f | f \in Y, F_f > 0\}. \tag{2}$$

Let N denote the number of all features in system $\mathcal{S}$, i.e.,

$$N = \sum_{g \in \mathrm{X}} F_g = \sum_{i=1}^{n} |\mathrm{S_i}|, \quad N \in \mathbb{N}, \quad N \geq n. \tag{3}$$

Definition 1: The *consistency* $C(\mathcal{S})$ of feature subsets in system $\mathcal{S}$ is defined as:

$$C(\mathcal{S}) = \frac{1}{|X|} \sum_{f \in X} \frac{F_f - 1}{n - 1}\ . \tag{4}$$

Definition 2: The weighted consistency $CW(\mathcal{S})$ of system $\mathcal{S}$ is defined as

$$CW(\mathcal{S}) = \sum_{f \in \mathrm{X}} w_f \frac{F_f - 1}{n - 1}\ , \tag{5}$$

where $w_f = \frac{F_f}{\sum_{g \in \mathrm{X}} F_g}$, $0 < w_f \leq 1$, $\sum_{f \in \mathrm{X}} w_f = 1$.

Because $F_f = 0$ for all $f \in \mathrm{Y} \setminus \mathrm{X}$, the *weighted consistency* $CW(\mathcal{S})$ can be equally expressed using notation (3) as

$$CW(\mathcal{S}) = \sum_{f \in \mathrm{Y}} \frac{F_f}{N} \cdot \frac{F_f - 1}{n - 1}\ . \tag{6}$$

The main properties of both $C(\mathcal{S})$ and $CW(\mathcal{S})$ are:

1) $0 \leq C(\mathcal{S}) \leq 1$, $0 \leq CW(\mathcal{S}) \leq 1$.
2) $C(\mathcal{S}) = 1$, $CW(\mathcal{S}) = 1$ if and only if (iff) all subsets in $\mathcal{S}$ are identical.
3) $C(\mathcal{S}) = 0$, $CW(\mathcal{S}) = 0$ iff all subsets in $\mathcal{S}$ are disjunct from each other.

It is obvious that $CW(\mathcal{S}) = 0$ iff $N = |\mathrm{X}|$, i.e., iff $F_f = 1$ for all $f \in \mathrm{X}$. This is unrealistic in most of real cases. Whenever $n > |\mathrm{X}|$, some feature must appear in more than one subset and consequently $CW(\mathcal{S}) > 0$. Similarly, $CW(\mathcal{S}) = 1$ iff $N = n|\mathrm{X}|$, otherwise all subsets can not be identical. Note that for $C(\mathcal{S}) \approx 0.5$ on average each feature present in $\mathcal{S}$ appears in about half of all subsets.

When comparing FS methods, higher stability of subsets produced during experiment trials is clearly advantageous. However, it should be considered a complementary measure only as it does not have any straight relation to the key measure of classifier generalization ability.

Remark: In experiments, if the best performing FS method also produces feature subsets with high *consistency*, its superiority can be more confidently assumed well founded.

III. EXPERIMENTS

To illustrate the differences between simpler and more complex FS methods we have collected experimental results under various settings: for two different classifiers, four FS search algorithms and seven datasets with dimensionalities ranging from 13 to 65 and number of classes ranging from 2 to 3. We used 3 different mammogram datasets as well as wine and spectf datasets from UCI Repository [22], speech data from British Telecom and sonar data [23]. For details see Tables I to VII.

Note that the choice of classifier and/or FS setup may not be optimal for each dataset, thus the reported results may be inferior to results reported in the literature; the purpose of our experiments is mutual comparison of FS methods only.

The following set-up was used in all experiments. The number of FS trials (size of evaluated system of subsets) was set to $n = 100$. From each dataset 25% of data in each class was reserved for testing and as such excluded from FS process. In each FS trial 90% of the remaining data was randomly sampled to form a trial-local data set. In *wrapper* FS setting the criterion value has been obtained as the average over 10 classification rates obtained using 10-fold hold-out, where in each loop the trial-local data had been scattered randomly to 60% training, 30% validation, and 10% unused data. In *filter* setting the criterion values have been computed from the training data part only. All reported classification rates have been obtained on independent test data.

The application of BIF, SFS and SFFS was straightforward. The OS algorithm allows various set-ups. Here we resorted to the simple deterministic version, denoted as OS(1,BIF) in the following as it is initialized by means of BIF and the oscillating cycle depth [20] parameter has been set to enable fastest search to $\Delta = 1$. The problem of determining optimal feature subset size was solved in all experiments by brute

force. All algorithms were applied repeatedly for all possible feature sizes whenever needed. The final result has been determined as that with the highest classification accuracy (and lowest subset size in case of ties).

A. Notes on Obtained Results

The examples (though naturally covering only a subset of possible types of recognition problems) illustrate several phenomena that can not be neglected in general case.

Over-fitting [15] is indeed a problem of utmost importance. It manifests itself in degraded classification rate of stronger FS methods on independent data when compared to BIF (see Table IIa, IIIb, IVa,b, Va). However, it may be also seen that in many cases it is the stronger selectors that yield best classification rate of all methods and both classifiers on independent data (see Table I, II, VI, VII). It is very difficult, if not impossible, to predict which feature selector will yield best results for particular problem. The best guess (although not valid in all cases) is to expect the overfitting problem to appear more often with too small number of sample data in relation to dimensionality.

For illustration consider the average classification rate of each method and classifier over the experiments. With gaussian classifier the ordering of methods according to achieved classification rate on independent test data is SFFS 70.9%, SFS $70,2\%$, BIF $69,6\%$ and OS(1,BIF) $68,3\%$. With 3NN classifier it is SFFS $81,6\%$, BIF $81,3\%$, SFS $80,6\%$ and OS(1,BIF) $80,2\%$. Note that SFFS is a strong feature selector, yet it is shown here to overperform other (simpler) methods. It is also shown that FS methods' performance depends on circumstances – here with gaussian classifier SFS performs better than BIF, with 3NN it is the other way round.

Besides classification performance other performance characteristics of the considered methods should not be neglected. In terms of dimensionality reduction performance (percentage of features discarded, relatively to full problem dimensionality) in the presented experiments with gaussian classifier the method ordering is BIF 74.2%, SFS $71,9\%$, OS(1,BIF) 70.5% and SFFS $69,1\%$. However, with 3NN the most effective dimensionality reducer shows to be OS(1,BIF) $67,6\%$ followed by SFS 65.2%, SFFS 63.1% and BIF 51.7%.

In terms of stability (robustness of feature preferencess) as measured by the CW measure (5) the resulting method ordering for gaussian classifier is BIF 0.559, OS(1,BIF) 0.487, SFS 0.463 and SFFS 0.458. With 3NN it is BIF 0.658, SFFS 0.547, OS(1,BIF) 0.531 and SFS 0.485.

It is clear that no unanimous winner can be pointed out among the considered methods. BIF proves to be the most stable FS method, performing well whenever over-fitting is a problem. However, in terms of classification accuracy on independent data SFFS shows to be the most recommendable.

IV. DISCUSSION AND CONCLUSIONS

With respect to FS we can distinguish the following entities which all affect the resulting classification performance: search algorithms, stopping criteria, feature subset evaluation criteria, data and classifier. The impact of the FS process on the final classifier performance (with our interest targeted naturally at its generalization performance, i.e., its ability to classify previously unknown data) depends on all of these entities.

When comparing pure search algorithms as such, then there is enough ground (both theoretical and experimental) to claim that newer, often more complex methods, have better potential of finding better solutions. This often follows directly from the method definition, as newer methods are often defined to improve some particular weakness of older ones. (Unlike BIF, SFS takes into account inter-feature dependencies. Unlike SFS, +L-R does not suffer the nesting problem. Unlike +L-R, Floating Search does not depend on pre-specified user parameters. Unlike Floating Search, OS may avoid local extremes by means of randomized initialization etc.). Better solution, however, means in this context merely being closer to optimum with respect to the adopted criterion. This may not tell much about final classifier quality, while criterion choice has proved to be a considerable problem in itself. None of applicable criteria seems to have good enough relation to classifier generalization performance.

When comparing feature selection methods as a whole (under specific criterion-classifier-data settings) the advantages of more modern search algorithms may diminish considerably. Reunanen [16] points out, and our experiments confirm, that a simple method like BIF or SFS may lead to better classifier generalization. The problem we see with the ongoing discussion is that this is often claimed to be the general case. But this is not true, as confirmed by several of our experiments.

Moreover, the possibly uneven resulting classification accuracy on independent test data in case of complex FS methods may be viewed as a direct consequence of insufficient criteria. In this view it is difficult to claim that more complex FS methods are problematic per se.

Our concluding recommendation can be stated as follows: whenever possible, try variety of methods ranging from BIF to more complex ones. If one method only has to be chosen, than we would stay with SFFS as the best general compromise between performance, generalization ability and search speed.

A. Does It Make Sense to Develop New FS Methods?

Our current experience shows that no clear and unambiguous qualitative hierarchy can be established within the existing framework of methods, i.e., although some methods perform better than others more often, this is not the case always and any method can prove to be the best tool for some particular problem. Adding to this pool of methods may thus bring improvement, although it is more and more difficult to come up with new ideas that have not been utilized before. Regarding the performance of search algorithms as such, developing methods that yield results closer to optimum with respect to any given criterion may bring considerably more advantage in future, when better criteria may have been found to better express the relation between feature subsets and classifier generalization ability.

TABLE I
CLASSIFICATION PERFORMANCE AS RESULT OF WRAPPER-BASED FEATURE SELECTION ON WINE DATA.

Wine data: 13 features, 3 classes containing 59, 71 and 48 samples, UCI Repository											
			Criterion val.		Classif. rate		Subset Size		Stability		Time
	Classifier	FS Method	Mean	*St.Dv.*	Mean	*St.Dv.*	Mean	*St.Dv.*	C	CW	h:m:s.ss
a)	Gaussian	BIF	0.601	*0.023*	0.532	*0.036*	2.41	*0.81*	0.338	0.831	00:00
		SFS	0.645	*0.025*	0.515	*0.065*	3.27	*0.77*	0.29	0.672	00:00
		SFFS	0.672	*0.013*	0.579	*0.068*	3.61	*1.18*	0.511	0.637	00:01
		OS(1,BIF)	0.718	*0.016*	0.565	*0.076*	3.03	*1.06*	0.427	0.627	00:02
b)	3-NN	BIF	0.969	*0.005*	0.959	*0.018*	9.14	*1.89*	0.7	0.884	00:00
	scaled	SFS	0.978	*0.006*	0.972	*0.021*	7.71	*1.58*	0.589	0.736	00:01
		SFFS	0.985	*0.003*	0.968	*0.022*	7.57	*1.81*	0.578	0.75	00:02
		OS(1,BIF)	0.990	*0.004*	0.974	*0.019*	7.4	*1.85*	0.565	0.717	00:02

TABLE II
CLASSIFICATION PERFORMANCE AS RESULT OF WRAPPER-BASED FEATURE SELECTION ON MAMMOGRAM DATA.

Mammogram data, 65 features, 2 classes containing 57 (benign) and 29 (malignant) samples, UCI Rep.											
			Criterion val.		Classif. rate		Subset Size		Stability		Time
	Classifier	FS Method	Mean	*St.Dv.*	Mean	*St.Dv.*	Mean	*St.Dv.*	C	CW	h:m:s.ss
a)	Gaussian	BIF	0.701	*0.019*	0.685	*0.057*	3.13	*1.97*	0.099	0.397	00:07
		SFS	0.746	*0.020*	0.671	*0.081*	5.33	*2.38*	0.078	0.236	02:17
		SFFS	0.754	*0.022*	0.648	*0.079*	5.56	*2.52*	0.080	0.203	09:21
		OS(1,BIF)	0.814	*0.019*	0.532	*0.115*	7.12	*2.54*	0.110	0.205	14:22
b)	3-NN	BIF	0.792	*0.026*	0.757	*0.068*	10.9	*7.51*	0.186	0.509	00:00
	scaled	SFS	0.872	*0.037*	0.810	*0.117*	10.1	*5.71*	0.152	0.386	00:07
		SFFS	0.909	*0.035*	0.886	*0.102*	7.28	*4.58*	0.112	0.521	00:40
		OS(1,BIF)	0.917	*0.026*	0.814	*0.109*	8.5	*5.36*	0.133	0.449	00:41

TABLE III
CLASSIFICATION PERFORMANCE AS RESULT OF WRAPPER-BASED FEATURE SELECTION ON SONAR DATA.

Sonar data, 60 features, 2 classes containing 103 (mine) and 105 (rock) samples, Gorman & Sejnowski											
			Criterion val.		Classif. rate		Subset Size		Stability		Time
	Classifier	FS Method	Mean	*St.Dv.*	Mean	*St.Dv.*	Mean	*St.Dv.*	C	CW	h:m:s.ss
a)	Gaussian	BIF	0.705	*0.016*	0.507	*0.045*	10.1	*4.51*	0.319	0.666	00:05
		SFS	0.795	*0.015*	0.551	*0.091*	14.3	*4.97*	0.23	0.363	01:39
		SFFS	0.819	*0.015*	0.548	*0.085*	16.7	*5.10*	0.272	0.404	10:53
		OS(1,BIF)	0.846	*0.014*	0.500	*0.065*	16.46	*5.64*	0.277	0.414	25:23
b)	3-NN	BIF	0.855	*0.010*	0.649	*0.084*	20.7	*7.16*	0.377	0.759	00:01
	scaled	SFS	0.885	*0.013*	0.516	*0.086*	20.7	*8.35*	0.338	0.417	00:30
		SFFS	0.906	*0.012*	0.496	*0.076*	22.9	*8.31*	0.375	0.491	02:10
		OS(1,BIF)	0.931	*0.008*	0.489	*0.064*	19.44	*6.57*	0.317	0.570	02:38

TABLE IV
CLASSIFICATION PERFORMANCE AS RESULT OF WRAPPER-BASED FEATURE SELECTION ON SONAR DATA.

Spectf data, 34 features, 2 classes containing 212 and 55 samples, UCI Repository											
			Criterion val.		Classif. rate		Subset Size		Stability		Time
	Classifier	FS Method	Mean	*St.Dv.*	Mean	*St.Dv.*	Mean	*St.Dv.*	C	CW	h:m:s.ss
a)	Gaussian	BIF	0.800	*0.001*	0.783	*0.020*	4.18	*7.59*	0.141	0.321	00:01
		SFS	0.806	*0.004*	0.758	*0.034*	14.5	*5.16*	0.322	0.406	00:22
		SFFS	0.814	*0.008*	0.746	*0.036*	12.3	*4.94*	0.272	0.35	01:33
		OS(1,BIF)	0.809	*0.006*	0.771	*0.025*	12.8	*4.26*	0.283	0.484	02:43
b)	3-NN	BIF	0.804	*0.011*	0.762	*0.037*	6.12	*7.20*	0.137	0.257	00:01
	scaled	SFS	0.846	*0.011*	0.746	*0.041*	8.31	*4.26*	0.181	0.321	00:19
		SFFS	0.859	*0.012*	0.752	*0.039*	10.1	*5.28*	0.222	0.387	01:33
		OS(1,BIF)	0.884	*0.009*	0.735	*0.040*	8.9	*3.91*	0.199	0.363	02:15

ACKNOWLEDGEMENTS

The work has been supported by projects AV0Z1075050506 of the GAAV CR, GAČR 102/08/0593, 102/07/1594 and CR MŠMT grants 2C06019 ZIMOLEZ and 1M0572 DAR.

REFERENCES

[1] J. M. V. Campenhout, "On the peaking of the Hughes mean recognition accuarcy: The resolution of an apparent paradox," *IEEE Transactions on System, Man and Cybernetics*, vol. SMC-8, pp. 390–395, May 1978.

[2] R. P. W. Duin and B. J. Kröse, "On the possibility of avoiding peaking," in *Proc 5th international conference on pattern recognition*, Miami beach, Florida, USA, December 1980, pp. 1375–1378.

[3] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proceedings of the 20th International Conference on Machine Learning*, 2003, pp. 56–63.

[4] M. Dash, K. Choi, S. P., and H. Liu, "Feature selection for clustering - a filter solution," in *Proceedings of the Second International Conference on Data Mining*, 2002, pp. 115–122.

[5] R. Kohavi and G. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, pp. 273–324, 1997.

[6] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

TABLE V
CLASSIFICATION PERFORMANCE AS RESULT OF WRAPPER-BASED FEATURE SELECTION ON MAMMOGRAM DATA.

WPBC data, 31 features, 2 classes containing 151 (nonrecur) and 47 (recur) samples, UCI Repository											
			Criterion val.		Classif. rate		Subset Size		Stability		Time
	Classifier	FS Method	Mean	*St.Dv.*	Mean	*St.Dv.*	Mean	*St.Dv.*	C	CW	h:m:s.ss
a)	Gaussian	BIF	0.756	*0.006*	0.755	*0.028*	6.39	*7.42*	0.198	0.295	00:00
		SFS	0.783	*0.012*	0.722	*0.044*	7.86	*4.52*	0.246	0.314	00:04
		SFFS	0.798	*0.011*	0.711	*0.044*	7.93	*3.40*	0.248	0.320	00:25
		OS(1,BIF)	0.799	*0.010*	0.701	*0.039*	6.11	*2.40*	0.189	0.373	00:54
b)	3-NN scaled	BIF	0.742	*0.014*	0.677	*0.076*	7.59	*9.59*	0.237	0.319	00:00
		SFS	0.780	*0.012*	0.697	*0.052*	7.9	*5.21*	0.247	0.312	00:03
		SFFS	0.791	*0.010*	0.707	*0.042*	9.85	*5.77*	0.311	0.360	00:16
		OS(1,BIF)	0.789	*0.012*	0.693	*0.040*	4.93	*3.67*	0.151	0.282	00:28

TABLE VI
CLASSIFICATION PERFORMANCE AS RESULT OF WRAPPER-BASED FEATURE SELECTION ON MAMMOGRAM DATA.

WDBC data, 30 features, 2 classes containing 357 (benign) and 212 (malignant) samples, UCI Rep.											
			Criterion val.		Classif. rate		Subset Size		Stability		Time
	Classifier	FS Method	Mean	*St.Dv.*	Mean	*St.Dv.*	Mean	*St.Dv.*	C	CW	h:m:s.ss
a)	Gaussian	BIF	0.940	*0.003*	0.940	*0.007*	22.7	*7.64*	0.753	0.817	00:00
		SFS	0.962	*0.004*	0.947	*0.014*	8.37	*3.51*	0.272	0.456	00:07
		SFFS	0.966	*0.003*	0.954	*0.016*	9.18	*3.24*	0.299	0.48	00:57
		OS(1,BIF)	0.967	*0.003*	0.951	*0.010*	8.73	*3.67*	0.284	0.510	01:20
b)	3-NN scaled	BIF	0.969	*0.002*	0.963	*0.006*	23.7	*3.71*	0.787	0.907	00:02
		SFS	0.978	*0.002*	0.956	*0.015*	11.5	*5.37*	0.376	0.509	00:30
		SFFS	0.981	*0.002*	0.958	*0.012*	13.2	*5.16*	0.435	0.547	02:10
		OS(1,BIF)	0.984	*0.002*	0.964	*0.011*	11.4	*5.01*	0.374	0.561	03:19

TABLE VII
CLASSIFICATION PERFORMANCE AS RESULT OF WRAPPER-BASED FEATURE SELECTION ON SPEECH DATA.

Speech data, 15 features, 2 classes containing 682 (yes) and 736 (no) samples, British Telecom											
			Criterion val.		Classif. rate		Subset Size		Stability		Time
	Classifier	FS Method	Mean	*St.Dv.*	Mean	*St.Dv.*	Mean	*St.Dv.*	C	CW	h:m:s.ss
a)	Gaussian	BIF	0.677	*0.026*	0.669	*0.027*	4.78	*3.87*	0.361	0.589	00:00
		SFS	0.766	*0.006*	0.749	*0.008*	6.57	*2.23*	0.542	0.791	00:01
		SFFS	0.804	*0.005*	0.774	*0.012*	8.92	*1.00*	0.633	0.813	00:08
		OS(1,BIF)	0.804	*0.006*	0.764	*0.014*	8.79	*0.85*	0.673	0.801	00:12
b)	3-NN scaled	BIF	0.921	*0.005*	0.926	*0.005*	14.3	*0.89*	0.956	0.969	00:04
		SFS	0.944	*0.003*	0.942	*0.009*	6.89	*1.35*	0.454	0.716	00:20
		SFFS	0.947	*0.003*	0.943	*0.007*	6.77	*1.36*	0.446	0.773	01:25
		OS(1,BIF)	0.947	*0.003*	0.943	*0.007*	6.63	*1.21*	0.436	0.779	03:01

[7] I. Kononenko, "Estimating attributes: Analysis and extensions of relief," in *ECML-94: Proc. European Conf. on Machine Learning*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 1994, pp. 171–182.

[8] P. Pudil, J. Novovičová, N. Choakjarernwanit, and J. Kittler, "Feature selection based on approximation of class densities by finite mixtures of special type," *Pattern Recognition*, vol. 28, pp. 1389–1398, 1995.

[9] J. Novovičová, P. Pudil, and J. Kittler, "Divergence based feature selection for multimodal class densities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 18, no. 2, pp. 218–223, 1996.

[10] S. Das, "Filters, wrappers and a boosting-based hybrid for feature selection," in *Proc. of the 18th International Conference on Machine Learning*, 2001, pp. 74–81.

[11] M. Sebban and R. Nock, "A hybrid filter/wrapper approach of feature selection using information theory," *Pattern Recognition*, vol. 35, pp. 835–846, 2002.

[12] P. Somol, J. Novovičová, and P. Pudil, "Flexible-hybrid sequential floating search in statistical feature selection," *Lecture Notes in Computer Science*, no. 4109, pp. 632–639, 2006.

[13] Y. Saeys, I. naki Inza, and P. L. naga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[14] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, March 2002.

[15] Š. J. Raudys, "Feature over-selection," in *Structural, Syntactic, and Statistical Pattern Recognition*, vol. LNCS 4109. Berlin / Heidelberg, Germany: Springer-Verlag, 2006, pp. 622–631.

[16] J. Reunanen, "Overfitting in making comparisons between variable selection methods," *J. Mach. Learn. Res.*, vol. 3, pp. 1371–1382, 2003.

[17] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 4–37, 2000.

[18] P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. London: Prentice-Hall International, 1982.

[19] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, pp. 1119–1125, 1994.

[20] P. Somol and P. Pudil, "Oscillating search algorithms for feature selection," in *Proceedings of the 15th IAPR Int. Conference on Pattern Recognition, Conference B: Pattern Recognition and Neural Networks*, 2000, pp. 406–409.

[21] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, pp. 491–502, 2005.

[22] A. Asuncion and D. Newman, "UCI machine learning repository," 2007. [Online]. Available: http://www.ics.uci.edu/$\sim$mlearn/{MLR}epository.html

[23] R. P. Gorman and T. J. Sejnowski, "Analysis of hidden units in a layered network trained to classify sonar targets," *Neural Networks*, vol. 1, pp. 75–89, 1988.