Evaluating the Stability of Feature Selectors That Optimize Feature Subset Cardinality

Petr Somol[1,2] and Jana Novovičová[1,2]

[1] Dept. of Pattern Recognition, Institute of Information Theory and Automation,
Academy of Sciences of the Czech Republic, 182 08 Prague, Czech Republic
{somol,novovic}@utia.cas.cz
http://ro.utia.cz/

[2] Faculty of Management, Prague University of Economics, Czech Republic

Abstract. Stability (robustness) of feature selection methods is a topic of recent interest. Unlike other known stability criteria, the new consistency measures proposed in this paper evaluate the overall occurrence of individual features in selected subsets of possibly varying cardinality. The new measures are compared to the generalized Kalousis measure which evaluates pairwise similarities between subsets. The new measures are computationally very effective and offer more than one type of insight into the stability problem. All considered measures have been used to compare two standard feature selection methods on a set of examples.

Keywords: Feature selection, stability, relative weighted consistency measure, sequential search, floating search.

1 Introduction

Feature selection (FS) 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. It has been pointed out recently that not only model performance but also stability (robustness) of the FS process is important. Domain experts prefer FS algorithms that perform stably when only small changes are made to the data set. However, relatively little attention has been devoted to the stability of FS methods so far. Recent works in the area of FS methods' stability mainly focus on various stability indices, introducing measures based on Hamming distance, Dunne et al. [1], correlation coefficients and Tanimoto distance, Kalousis et al. [2], consistency index, Kuncheva [3] and Shannon entropy, Křížek et al. [4]. Most of these recent works focus on the stability of single FS methods, while Saeys et al. [5] construct and study an ensemble of feature selectors. Stability of FS procedures depends on sample size, criteria utilized to perform FS and complexity of FS procedure (Raudys [6]).

To evaluate the stability of feature selectors we propose in this paper several new measures to be called the *consistency measure* (C), the *weighted consistency measure* (CW) and the *relative weighted consistency measure* (CW_{rel}). Unlike most other measures, these can be used for assessing FS methods that yield

N. da Vitora Lobo et al. (Eds.): SSPR&SPR 2008, LNCS 5342, pp. 956–966, 2008.

subsets of varying sizes. We compare the new measures to the generalized form of Kalousis measure (GK) [2]. All four measures have been used to compare the stability of Sequential Forward Selection (SFS) [7] and Sequential Forward Floating Selection (SFFS) [8] on a set of examples.

2 The Problem of Feature Selection Stability

It is common that classifier performance is considered the ultimate quality measure, even when assessing the FS process. However, misleading conclusions may be easily drawn when ignoring stability issues. Unstable FS performance may seriously deteriorate the properties of the final classifier by selecting the wrong features. In the following we focus on several new measures allowing to assess FS stability.

Let Y be the original set of features of size (cardinality) |Y|. Following [2] we define the *stability* of the FS algorithm as the robustness of the feature preferences it produces to differences in training sets drawn from the same generating distribution. FS algorithms express the feature preferences in the form of a selected feature subset $\mathrm{S} \subseteq \mathrm{Y}$. Stability quantifies how different training sets drawn from the same generating distribution affect the feature preferences.

2.1 Considered Measures of Feature Selection Stability

Let $\mathrm{Y} = \{f_1, f_2, \ldots, f_{|\mathrm{Y}|}\}$ be the set of all features and let $\mathcal{S} = \{\mathrm{S}_1, \ldots, \mathrm{S}_n\}$ be a system of $n > 1$ ($n \in \mathbb{N}$) feature subsets $\mathrm{S}_j = \big\{f_i | i = 1, \ldots, d_j, f_i \in \mathrm{Y}, d_j \in \{1, \ldots, |\mathrm{Y}|\}\big\}, j = 1, \ldots, n$ obtained from n runs of the evaluated FS algorithm. Let F_f be the number of occurrences (frequency) of feature f in system $\mathcal{S}$. Let X be the subset of Y representing all features that appear anywhere in $\mathcal{S}$:

$$\mathrm{X} = \{f | f \in \mathrm{Y}, F_f > 0\} = \bigcup_{i=1}^{n} \mathrm{S}_i, \quad \mathrm{X} \neq \emptyset. \tag{1}$$

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

$$N = \sum_{g \in \mathrm{X}} F_g = \sum_{i=1}^{n} |\mathrm{S_i}|, \quad N \in \mathbb{N}, \quad N \geq n. \tag{2}$$

Let us now introduce several measures usable for evaluating FS stability.

Definition 1. *The consistency $C(\mathcal{S})$ of system $\mathcal{S}$ is defined as*

$$C(\mathcal{S}) = \frac{1}{|\mathrm{X}|} \sum_{f \in \mathrm{X}} \frac{F_f - 1}{n - 1} \, . \tag{3}$$

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

$$CW(\mathcal{S}) = \sum_{f \in \mathrm{X}} w_f \frac{F_f - 1}{n - 1} \, , \tag{4}$$

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 $f \in \mathrm{Y}\backslash\mathrm{X}$, the *weighted consistency* $CW(\mathcal{S})$ can be equally expressed using notation (2) as

$$CW(\mathcal{S}) = \sum_{f \in \mathrm{Y}} \frac{F_f}{N} \cdot \frac{F_f - 1}{n-1} \, . \tag{5}$$

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.

Clearly, for any N, n representing some system of subsets $\mathcal{S}$ and for given Y there exists a system $\mathcal{S}_{min}$ with such configuration of features in its subsets that yields the minimal possible $CW(\cdot)$ value, to be denoted $CW_{min}(N, n, \mathrm{Y})$, being possibly greater than 0. Similarly, a system $\mathcal{S}_{max}$ exists that yields the maximal possible $CW(\cdot)$ value, to be denoted $CW_{max}(N, n)$, being possibly lower than 1.

It can be easily seen that $CW_{min}(\cdot)$ gets high when the sizes of feature subsets in system approach the total number of features $|\mathrm{Y}|$, because in such system the subsets get necessarily more similar to each other. Consequently, using measure (3) or (4) for comparison of various FS methods may lead to misleading results if the methods tend to yield systems of differently sized subsets. For this reason we introduce another measure, to be called the *relative weighted consistency*, which suppresses the influence of the sizes of subsets in system on the final value.

Definition 3. *The relative weighted consistency* $CW_{rel}(\mathcal{S}, \mathrm{Y})$ *of system* $\mathcal{S}$ *characterized by* N, n *and for given* Y *is defined as*

$$CW_{rel}(\mathcal{S}, \mathrm{Y}) = \frac{CW(\mathcal{S}) - CW_{min}(N, n, \mathrm{Y})}{CW_{max}(N, n) - CW_{min}(N, n, \mathrm{Y})} \, , \tag{6}$$

$$CW_{rel}(\mathcal{S}, \mathrm{Y}) = CW(\mathcal{S}) \quad \mathit{for} \quad CW_{max}(N, n) = CW_{min}(N, n, \mathrm{Y}).$$

Remark: The values $CW_{min}(\cdot)$ and $CW_{max}(\cdot)$ will be derived in Section 3.

It can be seen that for any N, n representing some system of subsets $\mathcal{S}$ and for given Y it is true that $0 \leq CW_{rel}(\mathcal{S}, \mathrm{Y}) \leq 1$ and for the corresponding systems $\mathcal{S}_{min}$ and $\mathcal{S}_{max}$ it is true that $CW_{rel}(\mathcal{S}_{min}) = 0$ and $CW_{rel}(\mathcal{S}_{max}) = 1$. The measure (6) does not exhibit the unwanted behavior of yielding higher values for systems with subset sizes closer to $|\mathrm{Y}|$, i.e., is independent on the size of feature subsets selected by the examined FS methods under fixed Y. We can say that this measure characterizes for given $\mathcal{S}, \mathrm{Y}$ the relative degree of randomness of the system of feature subsets on the scale between the maximum and minimum values of the weighted consistency (4).

A conceptually different measure of FS stability can be derived from the *similarity* measure between two subsets of features S_i and S_j, $\mathrm{S}_K(\mathrm{S}_i, \mathrm{S}_j)$ of arbitrary cardinality introduced by Kalousis et al. [2] as a straightforward adaptation of the Tanimoto distance measuring the amount of overlap between two sets.

Definition 4. *The similarity measure (to be called generalized Kalousis) of system $\mathcal{S}$ is the average similarity over all pairs of feature subsets in $\mathcal{S}$:*

$$GK(\mathcal{S}) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} S_K(\mathrm{S}_i, \mathrm{S}_j) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{|\mathrm{S}_i \cap \mathrm{S}_j|}{|\mathrm{S}_i \cup \mathrm{S}_j|}. \quad (7)$$

$GK(\mathcal{S})$ takes values from $[0,1]$ with 0 indicating empty intersection between all pairs of subsets $\mathrm{S}_i, \mathrm{S}_j$ and 1 indicating that all subsets are identical.

The properties of all introduced measures are discussed further in Sect. 4.

3 Relative Weighted Consistency Measure

To obtain the explicit formula for the *relative weighted consistency* measure $CW_{rel}(\mathcal{S}, \mathrm{Y})$ of system $\mathcal{S}$ for given Y as defined in Eq.(6) we derive in this Section the minimum and the maximum values for the *weighted consistency* measure (5). First we introduce supporting concepts.

3.1 Compacted Form of Arbitrary System of Feature Subsets

It follows from Eq.(5) that $CW(\cdot)$ is constant for all systems of subsets with equal N, n and identical feature frequencies. Therefore for any system $\mathcal{S}$ we can derive its *compacted form* yielding equal $CW(\cdot)$ value.

Definition 5. *The compacted form of system $\mathcal{S}$ is system $\mathcal{S}^{com}$ with equal characteristics N, n and equal feature frequencies, but with features reordered among subsets so that subset sizes are maximally equalized.*

It can be seen that for each system $\mathcal{S}$ a compacted form $\mathcal{S}^{com}$ exists yielding equal $CW(\cdot)$ value. It can also be seen that $\mathcal{S}^{com}$ consists of n_1 subsets of size $(k+1)$ and $(n - n_1)$ subsets is of size k, where

$$n_1 = N \bmod n \,, \qquad k = \frac{N - N \bmod n}{n} \,. \quad (8)$$

A compacted form is illustrated in Fig. 1.

3.2 The Impact of Feature Replacement on Consistency Value

Consider a system of subsets $\mathcal{S}$ characterized by N, n. We will now investigate how the value $CW(\mathcal{S})$ changes if one instance of some feature f_i in some subset in $\mathcal{S}$ is removed and another instance of some other feature f_j is added instead, so that system characteristics N, n remain unchanged. Let F_i and F_j denote the frequency of features f_i and f_j for all $i, j = 1, \ldots, |\mathrm{Y}|, i \neq j$ in $\mathcal{S}$.

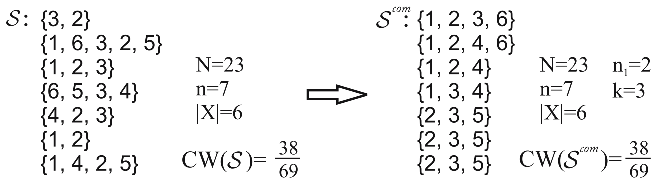

Fig. 1. Compacting system $\mathcal{S}$ to $\mathcal{S}^{com}$ does not change N, n, feature frequencies, nor the respective $CW(\cdot)$ value

Lemma 1. *Assume $F_i \leq F_j$. Replace one instance of the (equally or less frequent) feature f_i in system $\mathcal{S}$ by one new instance of the (equally or more frequent) feature f_j to obtain system $\mathcal{S}^{\mp}$. Then $CW(\mathcal{S}) < CW(\mathcal{S}^{\mp})$.*

Lemma 2. *Assume $F_i > F_j$. Replace one instance of the more frequent feature f_i in system $\mathcal{S}$ by one new instance of the less frequent feature f_j to obtain system $\mathcal{S}^{\mp}$. Then $CW(\mathcal{S}) \geq CW(\mathcal{S}^{\mp})$, with equality iff $F_i = F_j + 1$.*

Proof. (of Lemmas 1 and 2) Let us assume that $F_i = F_j + d$, where $d \in \mathbb{Z}$, $\mathbb{Z}$ is the set of integers. If $d \leq 0$ then $F_i \leq F_j$. If $d \geq 1$ then $F_i > F_j$. Let us denote by $\mathcal{S}^{\mp}$ the system in which one instance of feature f_i has been removed and one instance of feature f_j has been added, i.e., the frequency of feature f_i is $F_i - 1$ and the frequency of feature f_j is $F_j + 1$. Then we have

$$\begin{aligned}
&CW(\mathcal{S}) - CW(\mathcal{S}^{\mp}) \\
&= K \cdot \big\{F_i(F_i-1) + F_j(F_j-1) - [(F_i-1)(F_i-1-1) + (F_j+1)(F_j+1-1)]\big\} \\
&= K \cdot \big\{(F_j+d)(F_j+d-1) + F_j(F_j-1) \\
&- \big[(F_j+d-1)(F_j+d-1-1) + (F_j+1)(F_j+1-1)\big]\big\} = K \cdot 2(d-1),
\end{aligned}$$

where $K = 1/\big(N(n-1)\big)$. It immediately follows that $CW(\mathcal{S}) < CW(\mathcal{S}^{\mp})$ iff $d \leq 0$, $CW(\mathcal{S}) > CW(\mathcal{S}^{\mp})$ iff $d > 1$ and $CW(\mathcal{S}) = CW(\mathcal{S}^{\mp})$ iff $d = 1$. □

To summarize less formally – we have shown that when replacing one feature instance in system $\mathcal{S}$ by another while keeping the system characteristics N, n unchanged, it is true that: a) increasing the difference between frequencies of two features increases the value of $CW(\mathcal{S})$ defined in Eq.(5), while b) decreasing the difference between frequencies of two features decreases the value of $CW(\mathcal{S})$.

3.3 Minimum Value of Weighted Consistency

Consider an arbitrary system of subsets $\mathcal{S}$ characterized by N, n and given Y. We will now focus on finding the lower bound on $CW(\mathcal{S})$.

Definition 6. *The minimal system of subsets $\mathcal{S}_{min}$ characterized by N, n and for given* Y *is such system, where* $\max_{i,j \in \{1,\ldots,|\mathrm{Y}|\}}(|F_i - F_j|) \leq 1$.

An example of a minimal system is given in Fig. 2(a).

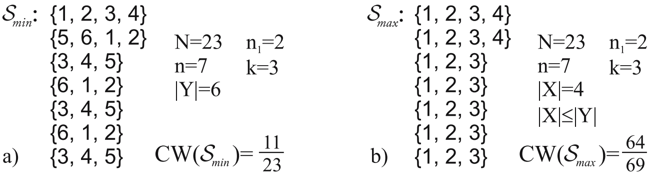

Fig. 2. (a) The system expected to yield the lowest value of $CW(\cdot)$ given N, n, Y. (b) The system expected to yield the highest value of $CW(\cdot)$ for given N, n.

Lemma 3. *Let $\mathcal{S}$ be a system of subsets characterized by N, n and given* Y. *If $\mathcal{S}_{min}$ is the minimal system with equal N, n and* Y, *then $CW(\mathcal{S}_{min}) \leq CW(\mathcal{S})$.*

Proof. Taking use of Lemma 2 as long as there exist features $f_i, f_j \in \mathrm{Y}$ such that $F_i > F_j + 1$, modify $\mathcal{S}$ so that one instance of f_i is replaced by one new instance of f_j, what decreases $CW(\mathcal{S})$. No decrease is possible iff there is no chance to take use of Lemma 2, i.e., the system conforms to Definition 6. □

Consider now for fixed N, n and given Y the compacted form of system $\mathcal{S}_{min}$. Let us denote for simplicity

$$D = N \bmod |\mathrm{Y}| \ . \tag{9}$$

It can be seen that no feature frequency in $\mathcal{S}_{min}$ can be lower than $F^{'}$, where

$$F^{'} = (N - D)/|\mathrm{Y}| \ . \tag{10}$$

Lemma 4. *The minimum value $CW_{min}(N, n, \mathrm{Y})$ of the consistency measure $CW(\mathcal{S})$ for a system $\mathcal{S}$ with characteristics N, n and for given* Y *is*

$$CW_{min}(N, n, \mathrm{Y}) = \frac{N^2 - |\mathrm{Y}|(N - D) - D^2}{|\mathrm{Y}|N(n-1)} \ . \tag{11}$$

Proof. It is obvious that in the compacted form of $\mathcal{S}_{min}$ exactly $(|\mathrm{Y}|-D)$ features occur F' times and D features occur $(F'+1)$ times. Substituting in (5) we obtain

$$CW_{min}(N, n, \mathrm{Y}) = (|\mathrm{Y}| - D)\left(\frac{F'}{(|\mathrm{Y}| - D)F' + D(F'+1)} \cdot \frac{F'-1}{n-1}\right) \tag{12}$$

$$+D\left(\frac{F'+1}{(|\mathrm{Y}| - D)F' + D(F'+1)} \cdot \frac{(F'+1)-1}{n-1}\right) = \frac{N^2 - |\mathrm{Y}|(N - D) - D^2}{|\mathrm{Y}|N(n-1)} \ .$$

□

3.4 Maximum Value of Weighted Consistency

Consider an arbitrary system $\mathcal{S}$ of subsets characterized by N, n and given Y. We will now focus on finding the upper bound on $CW(\mathcal{S})$. First denote for simplicity

$$H = N \bmod n \ . \tag{13}$$

Definition 7. *The maximal system of subsets $\mathcal{S}_{max}$ characterized by N, n is such system, where k features [defined in (8)] occur n times and, if $H > 0$, one feature occurs H times.*

An example of a maximal system is given in Fig. 2(b).

Lemma 5. *Let $\mathcal{S}$ be a system of subsets characterized by N, n. If $\mathcal{S}_{max}$ is the maximal system with the same characteristics, then $CW(\mathcal{S}) \leq CW(\mathcal{S}_{max})$.*

Proof. Taking use of Lemma 1 as long as there exist features $f_i, f_j \in \mathrm{Y}$ such that $F_i \leq F_j$, modify $\mathcal{S}$ so that one instance of f_i is replaced by one new instance of f_j, what increases $CW(\mathcal{S})$. No increase is possible only if there is no chance to take use of Lemma 1, i.e., the system conforms to Definition 7. □

Lemma 6. *The maximum value $CW_{max}(N, n)$ of the consistency measure $CW(\mathcal{S})$ for a system $\mathcal{S}$ with characteristics N, n is*

$$\begin{aligned} CW_{max}(N,n) &= \frac{(N-H)}{n}\left(\frac{n}{N}\cdot\frac{n-1}{n-1}\right) + 1\cdot\left(\frac{H}{N}\cdot\frac{H-1}{n-1}\right) \\ &= \frac{H^2 + N(n-1) - Hn}{N(n-1)} \,. \end{aligned} \tag{14}$$

Proof. Substitute to Eq.(5) the feature frequencies specified in Definition 7. □

3.5 Explicit Formula for Relative Weighted Consistency

Collecting the results from Sect. 3.3 and 3.4 we obtain the following proposition.

Proposition 1. *The relative weighted consistency measure of system $\mathcal{S}$ characterized by N, n and for given Y becomes*

$$CW_{rel}(\mathcal{S}, \mathrm{Y}) = \frac{|\mathrm{Y}|\left(N - D + \sum_{f\in \mathrm{Y}} F_f(F_f - 1)\right) - N^2 + D^2}{|\mathrm{Y}|\left(H^2 + n(N-H) - D\right) - N^2 + D^2} \,. \tag{15}$$

Proof. Substitute (5), (11) and (14) using (9) and (13) to Eq.(6). □

4 Practical Differences between the Discussed Measures

Assuming n is the number of subsets in $\mathcal{S}$, the GK time complexity is $O(n^2)$ as each pair of subsets is evaluated, while the complexity of C, CW and CW_{rel} is $O(n)$ as each subset is processed only once to collect feature occurrence statistics. The weighted consistency CW was defined to overcome the deficiency of consistency C which underrates systems of the type illustrated in Fig. 3. The measures C, CW and CW_{rel} all differ from GK in principle; GK evaluates pairwise similarities between subsets in system while measures C, CW and CW_{rel}

$\mathcal{S}_1$: { {1}, {1}, {1}, {1}, {1}, {1,2}, {1}, {1}, {1}, {1}, {1}, {1}, {3}, {1}, {1} }

$C(\mathcal{S}_1)=0.31$ $CW(\mathcal{S}_1)=0.813$ $CW_{rel}(\mathcal{S}_1)=0.806$ $GK(\mathcal{S}_1)=0.805$

Fig. 3. Illustrating the deficiency of the C measure on a system that clearly should not be rated as severely inconsistent

$\mathcal{S}_2$: {1, 2, 3, 4, 5, 6, 7}
{1, 2, 3, 4, 5, 6}
{1, 2, 3, 4, 5}
{1, 2, 3, 4}
{1, 2, 3}
{1, 2}
{1}

$GK(\mathcal{S}_2)=0.5$
$C(\mathcal{S}_2)=C(\mathcal{S}_3)=0.5$
$CW(\mathcal{S}_2)=CW(\mathcal{S}_3)=0.\overline{66}$
$CW_{rel}(\mathcal{S}_2)= CW_{rel}(\mathcal{S}_3)=0.\overline{33}$

$\mathcal{S}_3$: {1, 2, 3, 4}
{1, 2, 3, 4}
{1, 2, 3, 4}
{1, 2, 3, 4}
{1, 2, 3, 5}
{1, 2, 6, 5}
{1, 7, 6, 5}

$GK(\mathcal{S}_3)=0.564$

Fig. 4. Comparing the behavior of the considered measures on synthetic example

evaluate the overall occurrence of features in the system as a whole. The difference between C, CW, CW_{rel} and GK is illustrated in Fig. 4. The measures C, CW and GK tend to yield the higher values the closer the sizes of subsets in system are to the size of Y. This property seriously hinders the usability of these measures for comparison of various FS methods, should the compared methods yield differently sized subsets. The measure CW_{rel} overcomes this deficiency, yielding values unaffected by feature subset size issues.

Note that the measure CW_{rel} cannot be interpreted simply as a measure evaluating how much the selected subsets overlap. Instead, it shows the relative amount of randomness inherent in the concrete FS process. For a given total number of features in evaluated system and given size of Y it yields values on a scale $[0, 1]$ where 0 represents the outcome of completely random occurrence of features in the selected subsets and 1 indicates the most stable FS outcome possible. Note that even completely random FS process will lead to positive CW and GK values in most cases. The CW_{rel} helps to indicate cases where seemingly consistent results (that may be evaluated as highly consistent by CW or GK) are not the result of consistent FS performance, but follow from the inherent characteristics of certain systems of subsets.

5 Experimental Evaluation

In order to illustrate the considered measures we have conducted a series of FS experiments on real data from the UCI Repository [9]: *wine* data (13-dim., 3 classes of 59, 71, 48 samples), *wdbc* data (30-dim., 2 classes of 357 and 212 samples) and *cloud* data (10-dim., 2 classes of 1024 and 1024 samples).

We focused on comparing the stability of two standard FS methods: SFS and SFFS in the Wrapper [10] setting that allows the methods to be used as both the optimizers of feature subset and of subset size (subsets of all sizes are selected, then the one with the highest classification accuracy is chosen; in case of ties the one with lower cardinality is preferred).

Table 1. Consistency of FS Wrappers evaluated on Wine data, 13-dim, 3-class

	FS	Classif. rate		Subset size		C	CW	CW	GK	FS time	*CW*	*CW*
Wrap.	Meth.	Mean	*St.Dv.*	Mean	*St.Dv.*			rel		h:m:s	*min*	*max*
Gauss.	SFS	0.590	*0.023*	3.73	*1.70*	0.310	0.519	0.353	0.379	00:02:20	*0.286*	*0.947*
	SFFS	0.625	*0.023*	3.58	*1.23*	0.298	0.514	0.365	0.389	00:14:30	*0.275*	*0.932*
3-NN	SFS	0.982	*0.004*	7.12	*1.47*	0.547	0.752	0.467	0.615	00:07:45	*0.547*	*0.985*
	SFFS	0.987	*0.003*	6.91	*1.60*	0.531	0.763	0.508	0.637	00:33:20	*0.531*	*0.988*
SVM	SFS	0.980	*0.005*	9.09	*1.92*	0.699	0.758	0.203	0.611	00:16:49	*0.699*	*0.991*
	SFFS	0.989	*0.003*	8.46	*1.36*	0.650	0.816	0.516	0.697	01:08:41	*0.650*	*0.971*

Table 2. Consistency of FS Wrappers evaluated on WDBC data, 30-dim, 2-class

	FS	Classif. rate		Subset size		C	CW	CW	GK	FS time	*CW*	*CW*
Wrap.	Meth.	Mean	*St.Dv.*	Mean	*St.Dv.*			rel		h:m:s	*min*	*max*
Gauss.	SFS	0.963	*0.003*	11.95	*5.30*	0.398	0.506	0.181	0.332	01:02:04	*0.397*	*0.996*
	SFFS	0.969	*0.003*	12.17	*4.66*	0.405	0.556	0.259	0.387	09:13:03	*0.405*	*0.988*
3-NN	SFS	0.976	*0.002*	15.45	*5.74*	0.514	0.584	0.148	0.401	07:27:39	*0.514*	*0.984*
	SFFS	0.979	*0.002*	17.96	*5.67*	0.598	0.658	0.149	0.481	33:53:55	*0.598*	*0.998*
SVM	SFS	0.982	*0.002*	9.32	*4.12*	0.310	0.433	0.185	0.283	07:13:02	*0.310*	*0.977*
	SFFS	0.983	*0.002*	10.82	*4.58*	0.360	0.472	0.179	0.310	30:28:02	*0.360*	*0.987*

Table 3. Consistency of FS Wrappers evaluated on Cloud data, 10-dim, 2-class

	FS	Classif. rate		Subset size		C	CW	CW	GK	FS time	*CW*	*CW*
Wrap.	Meth.	Mean	*St.Dv.*	Mean	*St.Dv.*			rel		h:m:s	*min*	*max*
Gauss.	SFS	0.998	*4e-4*	4.80	*1.09*	0.480	0.794	0.644	0.671	00:03:54	*0.480*	*0.967*
	SFFS	0.999	*3e-4*	5.02	*0.87*	0.501	0.839	0.682	0.737	00:17:42	*0.501*	*0.997*
3-NN	SFS	1	*0*	1	*0*	1	1	1	1	05:25:24	*0.099*	*1.0*
	SFFS	1	*0*	1	*0*	1	1	1	1	11:04:39	*0.099*	*1.0*
SVM	SFS	1	*0*	1	*0*	1	1	1	1	02:41:40	*0.099*	*1.0*
	SFFS	1	*0*	1	*0*	1	1	1	1	04:13:19	*0.099*	*1.0*

We used the classification accuracy of three conceptually different classifiers as FS criteria: *Gaussian classifier*, *3-Nearest Neighbor* (majority voting) and *Support Vector Machine* (with Radial Basis Function kernel [11]).

In each setup FS was repeated 1000× on randomly sampled 80% of the data (class size ratios preserved). In each FS run the criterion was evaluated using 10-fold cross-validation, with 2/3 of available data randomly sampled for training and the remaining 1/3 used for testing.

5.1 Results

The results of our experiments are collected in Tables 1 to 3. Note that *CW* and *GK* exhibit similar behavior (except the slightly higher *CW* values' level), while

C is to be considered less reliable (cf. Fig. 3). The measure CW_{rel}, however, reveals different properties of FS process – note in Table 2 that with the *wdbc* data both FS methods yield too random results (note the low CW_{rel} values and also the high deviations in subset size). This may indicate some pitfall in the FS process – either there are no clearly preferable features in the set, or the methods overfit, etc. Another notable result is the consistent tendency of SFFS to yield higher CW, CW_{rel} and GK values than SFS in most of the experiments.

6 Conclusions

We propose several new consistency measures especially suitable for evaluating the stability of FS methods that yield subsets of varying sizes (although they can be used in fixed subset size problems). The key new measures CW and CW_{rel} are compared to the generalized Kalousis measure GK. Both CW and CW_{rel} are computationally less demanding than GK. Each of the considered measures evaluates the FS process from a different perspective – consequently they complement each other well. GK evaluates pairwise similarities between selected feature subsets. CW and CW_{rel} evaluate the overall occurrence of individual features in selected feature subsets. Unlike CW, the CW_{rel} shows the relative amount of "randomness" of the FS process, independently on subset size. The measures have been used to compare two standard FS methods on a set of examples. The SFFS has been shown as more stable than the SFS.

Acknowledgements. The work has been supported by GAAV CR project 1ET400750407 and AV0Z1075050506, GAČR 102/08/0593, 102/07/1594 and CR MŠMT grants 2C06019 ZIMOLEZ and 1M0572 DAR.

References

1. Dunne, K., Cunningham, P., Azuaje, F.: Solutions to instability problems with sequential wrapper-based approaches to feature selection. Technical Report TCD-CD-2002-28, Dept. of Computer Science, Trinity College, Dublin, Ireland (2002)
2. Kalousis, A., Prados, J., Hilario, M.: Stability of feature selection algorithms: a study on high-dimensional spaces. Knowledge and Inf. Syst. 12(1), 95–116 (2007)
3. Kuncheva, L.I.: A stability index for feature selection. In: Proc. 25th IASTED Int. Multi-Conf. Artificial Intelligence and Applications, pp. 421–427 (2007)
4. Křížek, P., Kittler, J., Hlaváč, V.: Improving stability of feature selection methods. In: Kropatsch, W.G., Kampel, M., Hanbury, A. (eds.) CAIP 2007. LNCS, vol. 4673, pp. 929–936. Springer, Heidelberg (2007)
5. Saeys, Y., Abeel, T., de Peer, Y.V.: Towards robust feature selection techniques. In: Proceedings of Benelearn, pp. 45–46 (2008)
6. Raudys, Š.: Feature over-selection. In: Yeung, D.-Y., Kwok, J.T., Fred, A., Roli, F., de Ridder, D. (eds.) SSPR 2006 and SPR 2006. LNCS, vol. 4109, pp. 622–631. Springer, Heidelberg (2006)
7. Devijver, P.A., Kittler, J.: Pattern Recognition: A Statistical Approach. Prentice-Hall International, London (1982)

8. Pudil, P., Novovičová, J., Kittler, J.: Floating search methods in feature selection. Pattern Recognition Letters 15, 1119–1125 (1994)
9. Asuncion, A., Newman, D.: UCI machine learning repository (2007), `http://www.ics.uci.edu/~mlearn/MLRepository.html`
10. Kohavi, R., John, G.: Wrappers for feature subset selection. Artificial Intelligence 97, 273–324 (1997)
11. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001), `http://www.csie.ntu.edu.tw/~cjlin/libsvm`