Evaluating Stability and Comparing Output of Feature Selectors that Optimize Feature Subset Cardinality

Petr Somol and Jana Novovičová

Abstract—Stability (robustness) of feature selection methods is a topic of recent interest, yet often neglected importance, with direct impact on the reliability of machine learning systems. We investigate the problem of evaluating the stability of feature selection processes yielding subsets of varying size. We introduce several novel feature selection stability measures and adjust some existing measures in a unifying framework that offers broad insight into the stability problem. We study in detail the properties of considered measures and demonstrate on various examples what information about the feature selection process can be gained. We also introduce an alternative approach to feature selection evaluation in the form of measures that enable comparing the similarity of two feature selection processes. These measures enable comparing, e.g., the output of two feature selection methods or two runs of one method with different parameters. The information obtained using the considered stability and similarity measures is shown to be usable for assessing feature selection methods (or criteria) as such.

Index Terms—Feature selection, feature stability, stability measures, similarity measures, sequential search, individual ranking, feature subset-size optimization, high dimensionality, small sample size.

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. Depending on the outcome of an FS algorithm, the result can be either a set of weighting-scoring, a ranking, or a subset of features. It has been pointed out recently that not only the model performance but also the stability (robustness) of the FS process is important [1], [2], [3], [4]. Domain experts prefer FS algorithms that perform stably when only small changes are made to the data set. Although low stability does not necessarily imply low classification rate (e.g., in presence of redundant, equally relevant features), it is often desirable to prefer unambiguous FS results. However, in many cases low stability follows from (and may help to indicate) fundamental problems in FS process. Nevertheless, relatively little attention has been devoted to the stability of FS methods so far.

In order to measure stability of FS algorithm, we need a measure of similarity for each of the above mentioned representations. Some recent works in the area of FS methods' stability focus on Pearson's correlation coefficient in order to measure similarity between two weighting-scorings produced by a given FS algorithm and Spearman's rank correlation coefficient to measure similarity between two rankings [2], [5]. Mainly, the attention is devoted to the stability of FS methods that produce a subset of features. Measuring the stability is based on various stability indexes, including measures based on the Hamming distance to measure similarity between two subsets of features, [1], on the adaptation of the Tanimoto distance [5], stability index, [3], Shannon entropy, [4], and the consistency-based measures [6]. A similarity measure for two sets of FS results based on weighted bipartite graph modeling is considered in [7]. Stability measures proposed in [3] and [4] assume constant subset size in each FS trial. Most of these recent works focus on the stability of single FS methods, while in [8] an ensemble of feature selectors is constructed and studied. The stability of FS procedures depends on sample size, criteria utilized to perform FS, and complexity of FS procedure [7], [9].

In this paper, we review and extend the framework of stability measures capable of evaluating feature selectors that yield subsets of varying size, i.e., where subset size may differ in each FS trial. The significant advantage of subset size-optimizing feature selectors (among others, the family of genetic algorithms [10], [11] or recent algorithms like Dynamic Oscillating Search [12]) is the fact that they exempt users from the necessity to choose the desired subset size—the choice that is often made based on insufficiently founded grounds, potentially degrading FS outcome.

The paper is organized as follows: A review of recent stability measures is given in Section 2. The framework of measures permitting varying subset size is devised in Section 2.1. In Section 3, the notion of intermeasures is introduced, allowing evaluation of the similarity of multiple FS processes' output. Section 4 puts measures into a taxonomy and discusses their properties. Section 5 presents examples based on real data. Section 6 summarizes the paper and suggests topics for further research.

• *The authors are with the Department of Pattern Recognition, Institute of Information Theory and Automation of the Czech Academy of Sciences, Pod vodarenskou vezi 4, 182 08 Prague 8, Czech Republic. E-mail: {somol, novovic}@utia.cas.cz.*

Manuscript received 18 Dec. 2008; revised 28 July 2009; accepted 13 Sept. 2009; published online 21 Jan. 2010.
Recommended for acceptance by M. Figueiridо.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-2008-12-0867.
Digital Object Identifier no. 10.1109/TPAMI.2010.34.

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 lead to degraded performance of the final classifier due to failure to identify the most relevant features.

Following [5], 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. The stability of the FS process for a given data set is the stability of the appearance of certain features after resampling the original data set. Let $\mathrm{Y} = \{f_1, \ldots, f_{|\mathrm{Y}|}\}$ be the set of all features of size (cardinality) $|\mathrm{Y}|$. In the following, we assume FS algorithms to express the feature preferences in the form of a subset of features $\mathrm{S} \subset \mathrm{Y}$.

Let $\mathcal{S} = \{\mathrm{S}_1, \ldots, \mathrm{S}_n\}$ be a system of n feature subsets,

$$\mathrm{S}_j = \big\{ f_{k_i} |\, i = 1, \ldots, d_j, f_{k_i} \in \mathrm{Y}, d_j \in \{1, \ldots, |\mathrm{Y}|\} \big\}, \quad j = 1, \ldots, n, n > 1, n \in \mathbb{N},$$

obtained from n runs of the evaluated FS algorithm on different samplings of a given data set. Let S_{id} and S_{jd} be subsets of features, $\mathrm{S}_{id}, \mathrm{S}_{jd} \subset \mathrm{Y}$, of the same size, $1 \le d \le |\mathrm{Y}|$. Let the measures evaluating stability of an FS process represented by system $\mathcal{S}$ be denoted as *intrameasures* (i.e., evaluating single system properties).

Dunne et al. [1] suggest measuring the stability of an FS method by the Average Normalized Hamming Distance (ANHD). Let m_j be the binary vector with $|\mathrm{Y}|$ dimensions corresponding to the subset S_j defined as

$$\mathrm{m}_j = (m_{j1}, \ldots, m_{j|\mathrm{Y}|}), \tag{1}$$

where $m_{jk} \in \{0, 1\}$, for all $j = 1, \ldots, n$, $k = 1, \ldots, |\mathrm{Y}|$, $m_{jk} = 1$ if feature $f_k \in \mathrm{Y}$ occurs in subset S_j and $m_{jk} = 0$ if feature $f_k \in \mathrm{Y}$ does not occur in subset S_j.

The *ANHD* is defined in [1] as follows:

$$ANHD(\mathcal{S}) = \frac{2}{|\mathrm{Y}| n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} HD(\mathrm{m}_i, \mathrm{m}_j). \tag{2}$$

Here,

$$HD(\mathrm{m}_i, \mathrm{m}_j) = \sum_{k=1}^{|\mathrm{Y}|} |m_{ik} - m_{jk}| \tag{3}$$

is the *Hamming distance* between the given pair of binary vectors m_i and m_j corresponding to the two subsets S_i and S_j. This measure determines how much variation there is in the distribution of features present in the subsets selected in different runs of the FS algorithm, with 0 indicating no variation and 1 indicating maximum variation. The ANHD is in the range $[0, 1]$.

Kalousis et al. [2], [5] proposed measuring similarity between two feature subsets S_i and S_j from system $\mathcal{S}$ using the *Tanimoto index (coefficient)* defined as the size of the intersection divided by the size of union of subsets S_i and S_j, [13]:

$$\begin{aligned} S_K(\mathrm{S}_i, \mathrm{S}_j) &= \frac{|\mathrm{S}_i \cap \mathrm{S}_j|}{|\mathrm{S}_i \cup \mathrm{S}_j|} = 1 - TD(\mathrm{S}_i, \mathrm{S}_j) \\ &= 1 - \frac{|\mathrm{S}_i| + |\mathrm{S}_j| - 2|\mathrm{S}_i \cap \mathrm{S}_j|}{|\mathrm{S}_i| + |\mathrm{S}_j| - |\mathrm{S}_i \cap \mathrm{S}_j|}, \end{aligned} \tag{4}$$

where $TD(\mathrm{S}_i, \mathrm{S}_j)$ is the *Tanimoto distance*, which measures the dissimilarity between two subsets S_i and S_j. The similarity index $S_K(\mathrm{S}_i, \mathrm{S}_j)$ takes values from $[0, 1]$, with 0 indicating empty intersection between two subsets $\mathrm{S}_i, \mathrm{S}_j$ of arbitrary size and 1 indicating that the two subsets are identical.

Kuncheva [3] introduced the stability index for a system $\mathcal{S} = \{\mathrm{S}_{1d}, \ldots, \mathrm{S}_{nd}\}$ for a fixed subset size, d,

$$\mathcal{I}_S(\mathcal{S}) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} I_C(\mathrm{S}_{id}, \mathrm{S}_{jd}), \tag{5}$$

where $I_C(\mathrm{S}_{id}, \mathrm{S}_{jd})$ is the consistency index for two subsets S_{id} and S_{jd} defined in [3] as

$$I_C(\mathrm{S}_{id}, \mathrm{S}_{jd}) = \frac{|\mathrm{S}_{id} \cap \mathrm{S}_{jd}| \cdot |\mathrm{Y}| - d^2}{d(|\mathrm{Y}| - d)}. \tag{6}$$

The maximum value of the index, $I_C(\mathrm{S}_{id}, \mathrm{S}_{jd}) = 1$, is achieved when $|\mathrm{S}_{id} \cap \mathrm{S}_{jd}| = d$. The minimum value of the index is bounded from below by -1. The index $I_C(\mathrm{S}_{id}, \mathrm{S}_{jd})$ is not defined for $d = 0$ or $d = |\mathrm{Y}|$.

In [4], the following stability measure based on Shannon entropy is proposed:

$$\gamma_d = -\sum_{j=1}^{K(|\mathrm{Y}|,d)} \hat{p}_{jd} \log_2 \hat{p}_{jd}, \tag{7}$$

where the convention that $0 \cdot \log_2 0 = 0$ is used. Here, $K(|\mathrm{Y}|, d)$ is the number of all possible subsets of size d from Y, i.e., $K(|\mathrm{Y}|, d) = \binom{|\mathrm{Y}|}{d}$, while s_{jd} is the number of occurrences of the set S_{jd} in the sequence of n subsets of size d; $\hat{p}_{jd} = \frac{s_{jd}}{n}$ is the relative frequency of the feature subset S_{jd} in the system $\mathcal{S}$,

$$\sum_{j=1}^{K(|\mathrm{Y}|,d)} \hat{p}_{jd} = \sum_{j=1}^{K(|\mathrm{Y}|,d)} \frac{s_{jd}}{n} = 1.$$

The stability measure (7) takes values in the range $[0, \log(\min\{n, K(|\mathrm{Y}|, d)\})]$.

Note that the stability measures proposed in [3] and [4] can be used only for fixed size of the subsets of features in the system $\mathcal{S}$.

2.1 Stability Measures for Use with Varying Feature Subset Sizes

The framework of currently available measures suffers two drawbacks: 1) Values yielded by various measures for the same system are differently bounded and thus hardly comparable, 2) most of available measures are considered only for FS problems with prespecified subset size d (to be denoted d-parametrized in the following) although many important FS methods allow the subset size to be optimized in the course of search (to be denoted d-optimizing).

In this section, we provide a framework of modified and newly defined measures to tackle the above problems. The desirable properties of considered stability measures $StabMeasure(\mathcal{S})$ of the system $\mathcal{S}$ are given below.

1. $0 \leq StabMeasure(\mathcal{S}) \leq 1$.
2. $StabMeasure(\mathcal{S})$ value close to 1 implies high level of FS algorithm stability and a value close to 0 implies low level of FS algorithm stability.
3. $\mathcal{S}$ may consist of subsets of varying size.

First, we introduce the concept of evaluating FS stability based on feature occurrence statistics. Let X be the subset of Y representing all features that appear anywhere in $\mathcal{S}$:

$$\mathrm{X} = \{f | f \in \mathrm{Y}, F_f > 0\} = \bigcup_{i=1}^{n} \mathrm{S}_i, \quad \mathrm{X} \neq \emptyset, \tag{8}$$

where F_f is the number of occurrences (frequency) of feature $f \in \mathrm{Y}$ in system $\mathcal{S}$. Let N denote the total number of occurrences of any feature in system $\mathcal{S}$, i.e.,

$$N = \sum_{g \in \mathrm{X}} F_g = \sum_{i=1}^{n} |\mathrm{S}_i|, \quad N \in \mathbb{N}, \quad N \geq n. \tag{9}$$

The *consistency* measure for measuring the stability of the system is developed in two steps.

STEP 1—define the measure of occurrence stability of the feature $f \in \mathrm{X}$ in the system $\mathcal{S}$. The minimum value F_{min} of F_f for all features $f \in \mathrm{X}$ in the system $\mathcal{S}$ is 1 and the maximum value F_{max} equals n. We require that the measure of stability of the feature $f \in X$ in the system $\mathcal{S}$ takes value from $[0, 1]$ with 0 meaning that f occurs only in one of the n subsets of the $\mathcal{S}$ and 1 that f occurs in each subset of the system $\mathcal{S}$.

Definition 1. *We define the* consistency $C(f)$ *of the feature* $f \in \mathrm{X}$ *in the system* $\mathcal{S}$ *as*

$$C(f) = \frac{F_f - F_{min}}{F_{max} - F_{min}}. \tag{10}$$

The consistency $C(f)$ of the feature $f \in X$ has the following two properties:

1. $C(f) = 0$ if the frequency of $f \in \mathrm{X}$ is $F_f = 1$.
2. $C(f) = 1$ if the frequency of $f \in \mathrm{X}$ is $F_f = n$.

STEP 2—extend the definition of consistency to evaluate whole system:

Definition 2. *The* consistency $C(\mathcal{S})$ *of system* $\mathcal{S}$ *of feature subsets is defined as the average of consistencies over all features in the set* X*:*

$$C(\mathcal{S}) = \frac{1}{|\mathrm{X}|} \sum_{f \in \mathrm{X}} C(f) = \frac{1}{|\mathrm{X}|} \sum_{f \in \mathrm{X}} \frac{F_f - F_{min}}{F_{max} - F_{min}}. \tag{11}$$

This measure, however, overemphasizes the presence of low frequency features (see Section 4.2 for discussion).

Therefore, we define a measure in which the more frequent features are expected to contribute proportionately more to the overall stability of the system $\mathcal{S}$. The value $\frac{F_f}{N}$ denotes the relative frequency of the feature $f \in \mathrm{X}$ in the system $\mathcal{S}$. The weighted sum of the consistencies of a single feature with weights equal to $\frac{F_f}{N}$ provides the more reliable stability measure:

Definition 3. *The* weighted consistency $CW(\mathcal{S})$ *of the system* $\mathcal{S}$ *is defined as*

$$CW(\mathcal{S}) = \sum_{f \in \mathrm{X}} w_f \frac{F_f - F_{min}}{F_{max} - F_{min}}, \tag{12}$$

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

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

$$CW(\mathcal{S}) = \sum_{f \in \mathrm{X}} \frac{F_f}{N} \cdot \frac{F_f - F_{min}}{F_{max} - F_{min}} = \sum_{f \in \mathrm{Y}} \frac{F_f}{N} \cdot \frac{F_f - 1}{n - 1}. \tag{13}$$

It is obvious that $CW(\mathcal{S}) = 0$ if and only if (iff) $N = |\mathrm{X}|$, i.e., iff $F_f = 1$ for all $f \in \mathrm{X}$. 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 cannot 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 (note the case when $N \bmod n \neq 0$).

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 a system, the subsets necessarily get more similar to each other. Consequently, using measure (11) or (12) for comparison of the stability of various FS methods may lead to misleading results if the methods tend to yield systems of differently sized subsets. We will refer to this problem as "the problem of subset-size bias." Note that most of the available stability measures are affected by the same problem. 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 4. *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*

$$CW_{rel}(\mathcal{S}, \mathrm{Y}) = \frac{CW(\mathcal{S}) - CW_{min}(N, n, \mathrm{Y})}{CW_{max}(N, n) - CW_{min}(N, n, \mathrm{Y})}, \tag{14}$$

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

Denoting $D = N \bmod |\mathrm{Y}|$ and $H = N \bmod n$ for simplicity, it has been shown in [6] that

$$CW_{min}(N, n, \mathrm{Y}) = \frac{N^2 - |\mathrm{Y}|(N - D) - D^2}{|\mathrm{Y}| N (n - 1)}, \tag{15}$$

and

$$CW_{max}(N, n) = \frac{H^2 + N(n - 1) - Hn}{N(n - 1)}. \tag{16}$$

The relative weighted consistency then becomes:

$$CW_{rel}(\mathcal{S}, \mathrm{Y}) = \frac{|\mathrm{Y}|\Big(N - D + \sum_{f \in \mathrm{Y}} F_f(F_f - 1)\Big) - N^2 + D^2}{|\mathrm{Y}|(H^2 + n(N - H) - D) - N^2 + D^2}. \tag{17}$$

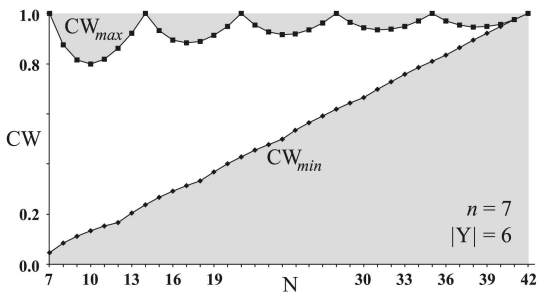

Fig. 1. Illustration of CW measure bounds.

The weighted consistency bounds $CW_{max}(N,n)$ and $CW_{min}(N,n,\mathrm{Y})$ are illustrated in Fig. 1. Note that CW_{rel} may be sensitive to small system changes if N approaches maximum (for given $|\mathrm{Y}|$ and n).

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 $CW_{rel}(\mathcal{S}_{max}) = 1$.

Measure (14) does not exhibit the unwanted behavior of yielding higher values for systems with subset sizes closer to $|\mathrm{Y}|$, i.e., it is independent of 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 (12).

Next, following the idea of Kalousis et al. [2], we define a conceptually different measure. It is derived from the similarity measure, $S_K(\mathrm{S}_i,\mathrm{S}_j)$, between two subsets of features S_i and S_j defined in (4).

Definition 5. *The* Average Tanimoto Index *of system $\mathcal{S}$ is defined as follows:*

$$ATI(\mathcal{S}) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} S_K(\mathrm{S}_i,\mathrm{S}_j). \tag{18}$$

$ATI(\mathcal{S})$ is the average similarity measure over all pairs of feature subsets in $\mathcal{S}$. It 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 of the system $\mathcal{S}$ are identical.

Next, we consider a Hamming Distance-based measure. It can be shown that the ANHD proposed in [1] and defined here in (2) can be rewritten using the frequency F_f of features $f \in \mathrm{Y}$ in a simpler form.

Lemma 1. *The Average Normalized Hamming Distance can be expressed in the form*

$$ANHD(\mathcal{S}) = \frac{2}{n(n-1)|\mathrm{Y}|} \sum_{f\in\mathrm{Y}} F_f(n-F_f). \tag{19}$$

Proof. It holds that

$$ANHD(\mathcal{S}) = \frac{2}{n(n-1)|\mathrm{Y}|} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sum_{k=1}^{|\mathrm{Y}|} |m_{ik} - m_{jk}|.$$

Let F_k be the frequency of the feature $f_k \in \mathrm{Y}$ in the system $\mathcal{S}$. Without loss of generality, we can suppose that $m_{1k} = m_{2k} = \cdots = m_{F_k,k} = 1$ and $m_{F_k+1,k} = m_{F_k+2,k} = \cdots = m_{n,k} = 0$. It means that $|m_{ik} - m_{F_k+l,k}| = 1$ for all $i = 1,\ldots,F_k$ and for all $l = 1,\ldots,n-F_k$; otherwise the absolute differences equal to zero. Therefore,

$$\sum_{k=1}^{|\mathrm{Y}|} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} |m_{ik} - m_{jk}| = \sum_{k=1}^{|\mathrm{Y}|} F_k(n-F_k) = \sum_{f\in\mathrm{Y}} F_f(n-F_f).$$

□

Next, following the ideas in [1] and [3], let us denote the *Normalized Hamming Index* (NHI) between two binary vectors (1) corresponding to subsets S_i and S_j to be

$$NHI(\mathrm{S}_i,\mathrm{S}_j) = 1 - \frac{1}{|\mathrm{Y}|} HD(\mathrm{S}_i,\mathrm{S}_j) = 1 - \frac{|\mathrm{S}_i \setminus \mathrm{S}_j| + |\mathrm{S}_j \setminus \mathrm{S}_i|}{|\mathrm{Y}|}, \tag{20}$$

where

$$HD(\mathrm{S}_i,\mathrm{S}_j) = |\mathrm{S}_i \setminus \mathrm{S}_j| + |\mathrm{S}_j \setminus \mathrm{S}_i| \tag{21}$$

is the Hamming Distance defined in (3) between two binary vectors (1) corresponding to the sets S_i and S_j in set notation. The NHI can be directly used in our context.

Definition 6. *The* Average Normalized Hamming Index *over all $n(n-1)/2$ pairs of binary vectors (1) corresponding to all pairs of subsets S_i and S_j is defined as:*

$$ANHI(\mathcal{S}) = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} NHI(\mathrm{S}_i,\mathrm{S}_j). \tag{22}$$

It follows from Lemma 1 that the ANHI can be rewritten by using the frequency F_f of the feature $f \in \mathrm{Y}$ in the following simpler form:

$$ANHI(\mathcal{S}) = 1 - \frac{2}{n(n-1)|\mathrm{Y}|} \sum_{f\in\mathrm{Y}} F_f(n-F_f). \tag{23}$$

Eventually, we introduce a frequency-based measure expressing the confidence of the feature selector about either selection or exclusion of each feature.

Definition 7. *The* Pseudo-Hamming *index $PH(\mathcal{S})$ of system $\mathcal{S}$ is defined as*

$$PH(\mathcal{S}) = \frac{2}{|\mathrm{Y}|} \sum_{f\in\mathrm{Y}} \frac{\max(F_f,\ n-F_f)}{n} - 1. \tag{24}$$

It takes values from $[0,1]$, with 1 indicating that all features are selected either always or never and 0 indicating that each feature appears in exactly half of all FS trials, i.e., features are selected/excluded with most uncertainty.

Note that all measures discussed in this section except CW_{rel} suffer the "subset-size-bias problem." The properties of all introduced measures are discussed further in Section 4.

3 Intermeasures

The measures discussed so far (*intrameasures*, see Section 2) are usable for evaluating the internal stability of one FS process. However, it may be valuable to compare two FS

processes. This would enable, e.g., evaluating the impact of various parameters of FS methods on the result of the selection process, comparing the behavior of two different FS methods on the same data, or evaluating how, for the given data set, different criteria differ in preferences of particular features, or comparing two FS processes on two data sets in an identical FS setting. This information cannot be obtained using *intrameasures*; two FS processes that yield results with similar or equal stability (according to any one *intrameasure*) may well differ in their preference of particular features. Therefore, we propose several *intermeasures* to enable comparison of multiple FS methods' outputs. The *intermeasures* should provide complementary information to *intrameasures*. Therefore, each of the following *intermeasures* is defined as an analogy to some *intrameasure*, based on the same or related principle.

Let $\mathcal{S}^l = \{\mathrm{S}_1^l, \ldots, \mathrm{S}_{n_l}^l\}$, be a system of $n_l > 1$ ($n_l \in \mathbb{N}$) feature subsets $\mathrm{S}_j^l = \{f_{k_i} |\ k_i = 1, \ldots, d_j^l, f_{k_i} \in \mathrm{Y}, d_j^l \in \{1, \ldots, |\mathrm{Y}|\}\}$, $j = 1, \ldots, n_l$, obtained from n_l runs of the evaluated FS algorithm on different samplings of a given data set with $l = 1, 2$ denoting the indices of the two compared systems. Let X_l be the subset of Y representing all features that appear anywhere in $\mathcal{S}^l$:

$$\mathrm{X}_l = \{f | f \in \mathrm{Y}, F_f^l > 0\} = \bigcup_{i=1}^{n_l} \mathrm{S}_i^l, \quad \mathrm{X}_l \neq \emptyset, \tag{25}$$

where F_f^l be the number of occurrences (frequency) of feature f in system $\mathcal{S}^l$. The desirable properties of each newly defined intermeasure $InterMeasure(\mathcal{S}^1, \mathcal{S}^2)$ are:

1. $0 \leq InterMeasure(\mathcal{S}^1, \mathcal{S}^2) \leq 1$.
2. $InterMeasure(\mathcal{S}^1, \mathcal{S}^2)$ value close to 1 implies high similarity and a value close to 0 implies low similarity of the two systems $\mathcal{S}^1$ and $\mathcal{S}^2$.
3. $\mathcal{S}^1$ and $\mathcal{S}^2$ may consist of subsets of varying size.
4. $\mathcal{S}^1$ and $\mathcal{S}^2$ may be systems of varying size (n_1 and n_2 need not be the same).

First, we define measures comparing two systems by means of average difference between relative feature frequencies.

Definition 8. *The* intersystem consistency $IC(\mathcal{S}_1, \mathcal{S}_2)$ *between two systems* $\mathcal{S}^1$ *and* $\mathcal{S}^2$ *is defined as*

$$IC(\mathcal{S}^1, \mathcal{S}^2) = 1 - \frac{1}{|\mathrm{X}_1 \cup \mathrm{X}_2|} \sum_{f \in \mathrm{X}_1 \cup \mathrm{X}_2} \left| \frac{F_f^1}{n_1} - \frac{F_f^2}{n_2} \right|. \tag{26}$$

Analogously to C, the measure IC is oversensitive to low-frequency features (see Sections 4.2 and 4.3). Therefore, we define its more reliable weighted counterpart:

Definition 9. *The* intersystem weighted consistency $ICW(\mathcal{S}^1, \mathcal{S}^2)$ *between two systems* $\mathcal{S}^1$ *and* $\mathcal{S}^2$ *is defined as*

$$ICW(\mathcal{S}^1, \mathcal{S}^2) = 1 - \sum_{f \in \mathrm{Y}} w_f \left| \frac{F_f^1}{n_1} - \frac{F_f^2}{n_2} \right|, \tag{27}$$

where

$$w_f = \frac{\max\left(\frac{F_f^1}{n_1}, \frac{F_f^2}{n_2}\right)}{\sum_{g \in \mathrm{Y}} \max\left(\frac{F_g^1}{n_1}, \frac{F_g^2}{n_2}\right)}.$$

Remark. The weighing in (27) assigns the most importance to features that are most frequent in 1) only one or 2) both of the systems. Both of these cases are to be considered equally important as they represent the cases of 1) minimum similarity or 2) maximum similarity of the two systems with respect to the evaluated feature.

Both IC and ICW take values from $[0, 1]$, with 0 indicating that no feature appears in more than one system and 1 indicating that the relative frequencies are equal for each feature in both systems, i.e., feature selector confidence regarding each feature is equal among the two compared systems.

Next, we define straightforward analogies to the ATI and ANHI measures:

Definition 10. *The* intersystem Average Tanimoto Index *(IATI) between two systems* $\mathcal{S}^1$ *and* $\mathcal{S}^2$ *is defined as*

$$IATI(\mathcal{S}^1, \mathcal{S}^2) = \frac{1}{n_1 \cdot n_2 \cdot |\mathrm{Y}|} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \frac{|\mathrm{S}_i^1 \cap \mathrm{S}_j^2|}{|\mathrm{S}_i^1 \cup \mathrm{S}_j^2|}. \tag{28}$$

$IATI(\mathcal{S}^1, \mathcal{S}^2)$ takes values from $[0, 1]$ with 0 indicating empty intersection between any pair of subsets, with one from $\mathcal{S}^1$ and the other from $\mathcal{S}^2$, and 1 indicating that all subsets in both systems $\mathcal{S}^1$ and $\mathcal{S}^2$ are identical.

Definition 11. *The* intersystem Average Normalized Hamming Index *between two systems* $\mathcal{S}^1$ *and* $\mathcal{S}^2$ *is defined as*

$$IANHI(\mathcal{S}^1, \mathcal{S}^2) = 1 - \frac{1}{n_1 \cdot n_2 \cdot |\mathrm{Y}|} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} HD\big(\mathrm{S}_i^1, \mathrm{S}_j^2\big), \tag{29}$$

where $HD(\cdot, \cdot)$ *is defined in (21).*

The IANHI can be expressed in the simpler form by using the frequencies F_f^1 and F_f^2.

Lemma 2. *The intersystem Average Normalized Hamming Index can be expressed in the form:*

$$\begin{aligned} IANHI(\mathcal{S}^1, \mathcal{S}^2) = 1 - \frac{1}{n_1 \cdot n_2 \cdot |\mathrm{Y}|} \sum_{f \in \mathrm{Y}} \big[& F_f^1 \big(n_2 - F_f^2\big) \\ & + F_f^2 \big(n_1 - F_f^1\big)\big]. \end{aligned} \tag{30}$$

Proof. It holds

$$\begin{aligned} IANHD(\mathcal{S}^1, \mathcal{S}^2) &= 1 - IANHI(\mathcal{S}^1, \mathcal{S}^2) \\ &= \frac{1}{n_1 \cdot n_2 \cdot |\mathrm{Y}|} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{|\mathrm{Y}|} \big| m_{ik}^1 - m_{jk}^2 \big|, \end{aligned}$$

where $\mathrm{m}_j^l = (m_{j1}^l, \ldots, m_{j|\mathrm{Y}|}^l)$, $l = 1, 2$, $j = 1, \ldots, n_l$, is the binary vector with $|\mathrm{Y}|$ dimensions corresponding to the subset S_j^l. Let F_k^l denote the frequency of the feature $f_k \in \mathrm{Y}$ in the system $\mathcal{S}^l$, $l = 1, 2$. Without loss of generality, we can suppose that

$$m^1_{1k} = m^1_{2k} = \cdots = m^1_{F^1_k,k} = 1,$$
$$m^1_{F^1_k+1,k} = m^1_{F^1_k+2,k} = \cdots = m^1_{n_1,k} = 0,$$
$$m^2_{1k} = m^2_{2k} = \cdots = m^2_{F^2_k,k} = 1,$$
$$m^2_{F^2_k+1,k} = m^2_{F^2_k+2,k} = \cdots = m^1_{n_2,k} = 0.$$

It means that $|m^1_{ik} - m^2_{F^2_k+r,k}| = 1$ for all $i = 1, \ldots, F^1_k$ and for all $r = 1, \ldots, n_2 - F^2_k$, and $|m^1_{F^1_k+s,k} - m^2_{j,k}| = 1$ for all $j = 1, \ldots, F^2_k$ and for all $s = 1, \ldots, n_1 - F^1_k$; otherwise the absolute differences equal to zero. Therefore,

$$\sum_{k=1}^{|\mathrm{Y}|}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}|m^1_{ik} - m^2_{jk}| = \sum_{k=1}^{|\mathrm{Y}|}\left[F^1_k(n_2 - F^2_k) + F^2_k(n_1 - F^1_k)\right].$$

□

The information that can be gained using the new intermeasures is discussed in Section 4 and illustrated in Section 5 (see Tables 9 and 10).

Remark. The intermeasures can be computed for $L(L-1)/2$ pairs of systems, $\mathcal{S}^l$ and $\mathcal{S}^m$, $l, m = 1, \ldots, L$, and the final intermeasure is the average intermeasure over all pairs.

4 Properties of the Considered Measures

In this section, we discuss the properties of the considered measures viewed from various perspectives. First, we assign each measure to a taxonomy, then we focus on properties that may have practical implications. We will investigate measures' behavior on synthetic examples simulating changing incidence of either the most relevant (constantly present in each subset) or the least relevant (randomly occurring) features.

Note. For the sake of explanation clarity, in all illustrations in this section, we resort to FS processes yielding subsets of constant size. In all randomized tests, values are drawn randomly from a uniform distribution.

4.1 Taxonomical View

The notion of FS stability as such is difficult to formalize unanimously. As shown above, a number of FS stability measures can be defined, with each measure expressing a slightly different aspect of the problem. Nevertheless, some common properties of certain measures can be identified. To make further discussion clearer, we introduce several basic differentiation approaches. First, the considered measures can be divided according to evaluation scope:

- *Feature-focused* measures—evaluate overall feature occurrence frequency over the system as a whole (regardless of concrete feature presence in concrete subsets).
- *Subset-focused* measures—evaluate features with respect to their occurrence in each particular subset in the system.

The *feature-focused* measures include: C (11), CW (12), CW_{rel} (14), $ANHI$ due to the existence of its form (23), PH (24), IC (26), ICW (27), and $IANHI$ due to the existence of its form (30). The information given by feature-focused measures is useful for assessing the confidence of feature selector

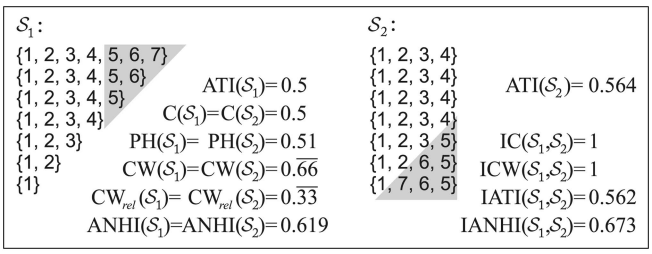

Fig. 2. Comparing the behavior of the considered measures on a synthetic example.

regarding particular feature preference. The *subset-focused* measures include: ATI (18), $\mathcal{I}_S$ (5), γ_d (7), and $IATI$ (28). They assign the most importance to concrete feature configurations in each subset. Therefore, they are more sensitive even to slight fluctuations in the investigated FS process. It should be noted that *feature-focused* measures give a coarser overview. As illustrated in Fig. 2, different systems with equal overall feature statistics (and, accordingly, equal *feature-focused* measure values) may consist of notably different subsets and, accordingly, yield different *subset-focused* measure values.

Next, the considered measures can be divided according to the importance assigned to feature exclusion:

- *Selection-registering* measures—ignore the information on the stability of feature exclusion ($|\mathrm{Y}|$ not taken into account).
- *Selection-exclusion-registering* measures—take into account both the stability of presence and the absence of features in subsets (knowledge of Y required).

The *selection-registering* measures include: γ_d (7), C (11), CW (12), ATI (18), IC (26), ICW (27), and $IATI$ (28). The *selection-exclusion-registering* measures include: $\mathcal{I}_S$ (5), $ANHI$ (22), PH (24), and $IANHI$ (29). The *selection-exclusion-registering* measures may give a fuller view of feature selector behavior. However, the information they give may become biased if $d \ll |\mathrm{Y}|$ like in many high-dimensional problems where the large number of consistently excluded features may misleadingly indicate high FS stability. Note: CW_{rel} requires $|\mathrm{Y}|$ for computing CW bounds but is not defined to evaluate the exclusion stability of features.

Next, the considered measures can be divided according to their behavior with respect to the "subset-size-bias problem":

- *Subset-size-biased* measures—yield values bounded more tightly than by $[0, 1]$ depending on the size of subsets in a system.
- *Subset-size-unbiased* measures—for each system containing subsets of arbitrary sizes, there exists "minimal" (resp. "maximal") configuration of features in the respective subsets for which the measure yields 0 (resp. 1).

The *subset-size-biased* measures include all considered measures except CW_{rel}. They may be misleading when used for comparing the stability of various FS processes that tend to yield subsets of different prevailing size. Due to existence of subset-size-dependent bounds, the *subset-size-biased* measures may yield considerably different values for feature selectors that select features with similar (un)certainty (e.g., in the presence of a large number of redundant features) but

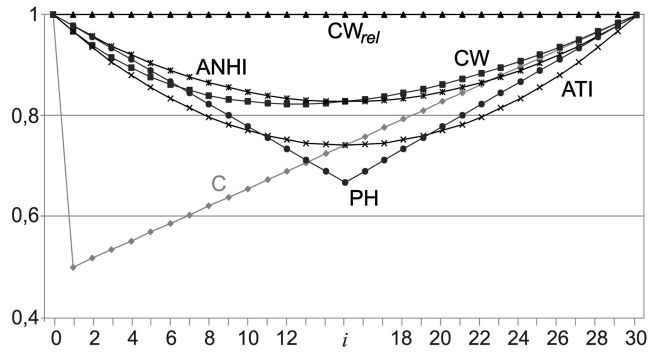

Fig. 3. Intrameasure sensitivity to single feature frequency change.

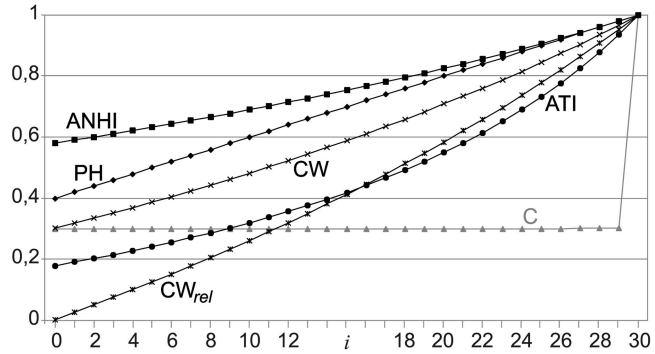

Fig. 4. Intrameasure sensitivity to changing the proportion between stably and unstably selected features ($|\mathrm{Y}| = 100$, $d = 30$, from which i features are fixed across the trials and $30 - i$ features are randomly drawn).

different subset-size preference. The only *subset-size-unbiased* measure in the presented framework is CW_{rel}.

Finally, as discussed in Sections 2 and 3, we distinguish FS process stability measures from FS process similarity measures:

- *Intrameasures*—evaluate the stability of one FS process.
- *Intermeasures*—compare output of multiple FS processes.

Basic comparison of the behavior of various measures is given in Fig. 2. The two example systems $\mathcal{S}_1$ and $\mathcal{S}_2$ have equal feature frequency characteristics although the subsets in them are composed differently. Note that *feature-focused intrameasures* yield, for both systems, the same values. Accordingly, *feature-focused intermeasures* yield 1 indicating maximum similarity.

Those measures, normalized to $[0, 1]$ and capable of evaluating systems with varying subset size, will be investigated in more detail in the following.

4.2 Properties of Intrameasures

As suggested in the previous section, each FS stability measure yields slightly different type of information. Nevertheless, all of them should be expected to rate higher such systems that reflect high feature selector confidence. Depending on the particular measure definition, this may mean that features get consistently selected or excluded or that the selected subsets do not differ much among each other.

In Fig. 3, we illustrate how the considered intrameasures respond to changing occurrence of one feature in a system. For $i = 0, \ldots, 30$, we evaluate systems $\mathcal{S}_{(i)}$ consisting of $n = 30$ subsets selected from $\mathrm{Y} = \{1, 2, 3\}$. In system $\mathcal{S}_{(i)}$, feature 1 is always selected, feature 2 is selected i times, and feature 3 is never selected. In this example, the measures CW, $ANHI$, ATI, and PH clearly reflect the notion that in a consistent system each feature is either consistently selected or consistently excluded. Accordingly, these measures indicate deteriorating stability in system $\mathcal{S}_{(i)}$ for i approaching $\frac{n}{2} = 15$.

Two of the measures exhibit notably different behavior. Note that CW_{rel} cannot be interpreted in the same sense as the other considered measures. Instead, it indicates the relative amount of randomness inherent in the system with respect to system size n, total number of feature occurrences N and given $|\mathrm{Y}|$. In Fig. 3, CW_{rel} correctly indicates that $\mathcal{S}_{(i)}$ is, for each $i = 0, \ldots, 30$, the system with the least random feature occurrence possible. The graph also illustrates a fundamental flaw in the behavior of C, which tends to indicate exaggeratedly unstable system in presence of isolated features with very low frequency. Remark: The weighted consistency CW was defined to overcome this problem.

Fig. 4 illustrates intrameasure response in the case of FS output consisting of both stably selected and unstably selected features. For $i = 0, \ldots, 30$, we evaluate systems $\mathcal{S}_{(i)}$ consisting of subsets $\mathrm{S}_k^{(i)} \in \mathcal{S}_{(i)}$, $\mathrm{S}_k^{(i)} \subset \mathrm{Y}$, $|\mathrm{Y}| = 100$, $|\mathrm{S}_k^{(i)}| = 30$, $k = 1, \ldots, 1{,}000$ having the form $\mathrm{S}_k^{(i)} = \mathrm{C}^{(i)} \cup \mathrm{R}_k^{(i)}$ where $\mathrm{C}^{(i)} \subset \mathrm{Y}$, $|\mathrm{C}^{(i)}| = i$, is a constant subset for each $k = 1, \ldots, 1{,}000$ and subset $\mathrm{R}_k^{(i)} \subset \mathrm{Y}$, $|\mathrm{R}_k^{(i)}| = 30 - i$, is drawn randomly from $\mathrm{Y} \setminus \mathrm{C}^{(i)}$.

In Fig. 4, most measures respond correctly to the changing proportion of stably and unstably selected features. The differences in measures' behavior are emphasized with low i values. Note that the *selection-exclusion-registering* measures yield higher values than the *selection-registering* ones. The graph gives another example of potentially misleading C performance. Note that only the *subset-size-unbiased* method CW_{rel} yields 0 for $i = 0$.

4.2.1 Complementarity of Information Gained from Evaluating Various Intrameasures

Assessing the stability of an FS process based on any single measure only may lead to misleading conclusions. For instance, very low ATI value may not necessarily indicate failure to identify important features—ATI is likely to be low in the presence of highly relevant but redundant features that may appear in various combinations in selected subsets. On the other hand, selection-exclusion-registering measures may evaluate such a system as highly stable, provided the remaining features are of low importance and, as such, consistently excluded. However, high $ANHI$ or PH value may lead to misleading conclusions about high stability in cases of high problem dimensionality where most of features remain excluded and high instability among the selected features gets neglected. Similarly, low C value suggests that, on average, features get selected only rarely. Yet that does not necessarily indicate a severely unstable system; if, at the same time, CW value is high, then part of the features get selected consistently with high confidence. To conclude, no single measure is capable of expressing all the information that can be useful to assess the stability of an FS process. It is recommended to consider evaluating a set of measures of

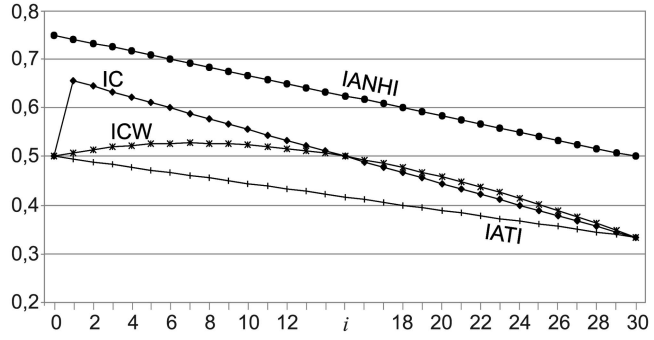

Fig. 5. Intermeasure sensitivity to single feature frequency change.

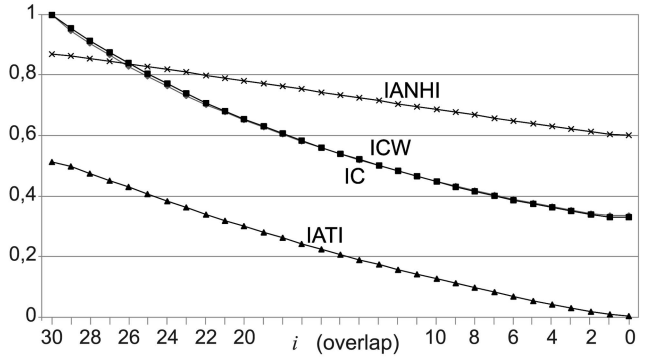

Fig. 6. Intermeasure sensitivity to changing overlap between random subsets.

different type (both selection-registering and selection-exclusion-registering and both feature-focused and subset-focused as well as subset-size-unbiased one) to get reasonable information about the evaluated FS process.

4.3 Properties of Intermeasures

The purpose of intermeasures is to compare the output of multiple FS processes and to assess their "similarity." Comparing multiple subset systems may reveal differences in feature preference among various feature selectors or among differently parameterized runs of the same feature selector. Thus, intermeasures provide complementary information about FS processes that cannot be gained using intrameasures.

In analogy to intrameasures, there is no unanimous definition of the term "similarity" of multiple FS processes' output. Again, the available intermeasures give various types of information that is not interchangeable.

In Fig. 5, we illustrate how the considered intermeasures respond to changing occurrence of one feature in system. For $i = 0, \ldots, 30$, we compare the output of systems pairs $\mathcal{S}^1_{(i)}$ and $\mathcal{S}^2$ with common $\mathrm{Y} = \{1, 2, 3, 4\}$. System $\mathcal{S}^1_{(i)}$ contains 30 subsets, where feature 1 is always selected, feature 2 is selected i times, and features 3 and 4 are never selected. System $\mathcal{S}^2$ contains 30 subsets, where features 1 and 3 are always selected and features 2 and 4 are never selected.

In this example, the intermeasures should indicate growing dissimilarity between $\mathcal{S}^1_{(i)}$ and $\mathcal{S}^2$ for increasing i. This is clearly the case with the *selection-registering* $IATI$ and *selection-exclusion-registering* $IANHI$, which yields values on a higher level as it takes the constant exclusion of feature 4 into account. Measure IC exhibits a problem similar to C (see Figs. 3 and 4), where the occurrence of a low frequent feature inadequately increases its value when i changes from 0 to 1. Measure ICW suppresses the negative impact of isolated features by means of weighing, but at a cost of counterintuitive behavior in this example (ICW does not decrease monotonically with increasing i). Nevertheless, both IC and ICW correctly evaluate the pair $\mathcal{S}^1_{(0)}, \mathcal{S}^2$ as more similar than the pair $\mathcal{S}^1_{(30)}, \mathcal{S}^2$.

Remark. Although the example in Fig. 5 seems to indicate inferior usability of IC and ICW, in a different context they prove more reliable than $IATI$ and $IANHI$. In Section 4.4, it will be shown that both IC and ICW respond better in the presence of random features. The principal difference between intrameasures and intermeasures can be illustrated by evaluating the systems in Fig. 5 using intrameasures. Note that systems $\mathcal{S}^1_{(i)}$ for $i = 0, \ldots, 30$ would yield *selection-registering* measure values equal to those in Fig. 3. Note that all intrameasures would yield 1 for systems $\mathcal{S}^1_{(0)}$, $\mathcal{S}^1_{(30)}$, and $\mathcal{S}^2$.

Fig. 6 illustrates intermeasure response on a pair of increasingly distinct, slightly unstable systems. Assuming $|\mathrm{Y}| = 100$, we compare systems $\mathcal{S}^1_{(i)}$ and $\mathcal{S}^2_{(i)}$, each containing 1,000 subsets of $d = 20$ features. Subsets in $\mathcal{S}^1_{(i)}$ are drawn randomly from $\mathrm{Z}^1_{(i)} \subset \mathrm{Y}$, $|\mathrm{Z}^1_{(i)}| = 30$, and subsets in $\mathcal{S}^2_{(i)}$ are drawn randomly from $\mathrm{Z}^2_{(i)} \subset \mathrm{Y}$, $|\mathrm{Z}^2_{(i)}| = 30$, where, for $i = 0, \ldots, 30$, the overlap (simulating the increasing similarity of the two systems) is $|\mathrm{Z}^1_{(i)} \cap \mathrm{Z}^2_{(i)}| = i$. Drawing features from $\mathrm{Z}^1_{(i)}$ and $\mathrm{Z}^2_{(i)}$ simulates the situation when there are 30 relevant but redundant and indistinguishable features from which only 20 get selected in each FS trial.

This example gives another view of the differences in intermeasure behavior. Both IC and ICW indicate well the similarity of the compared systems for $i = 30$. However, $IATI$ is the only measure to reflect that, for $i = 0$, there is no feature occurring in both of the systems $\mathcal{S}^1_{(0)}$ and $\mathcal{S}^2_{(0)}$.

4.3.1 Complementarity of Information Gained from Evaluating Various Intermeasures

Analogously to intrameasures, no single intermeasure can be considered sufficient to evaluate the similarity of two systems in entirety. The available intermeasures have been defined with the intention to provide complementary information to particular intrameasures. In analogy to C, the measure IC evaluates the similarity of feature frequencies in the two compared systems. In analogy to CW, the weighted measure ICW puts emphasis on comparing frequencies of the more frequent features. $IATI$, resp. $IAHNI$, has been defined to yield analogical information to ATI, resp. $AHNI$. Evaluating both $IATI$ and $IAHNI$ may give complementary information about the proportion of features being selected and those excluded consistently in both systems (see Fig. 6 where, for $i = 0$, the $IANHI = 0.6$ indicates that 60 percent of features have been excluded in both systems, while $IATI = 0$ indicates that the rest appear in the systems with completely different preferences).

4.4 Ability to Identify Randomness in Feature Selection Process

An important property of an FS process stability measure is the ability to indicate randomness (or feature preference

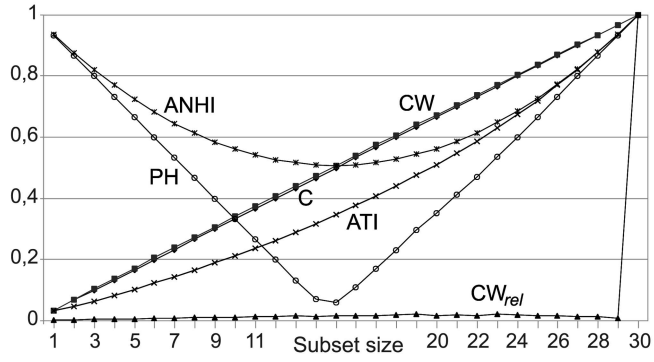

Fig. 7. The response of intrameasures to randomness in FS process.

uncertainty). Fig. 7 shows that most of the considered intrameasures do not indicate randomness clearly. In the experiment, 1,000 subsets were drawn randomly from Y, $|\mathrm{Y}| = 30$, for each subset size $d = 1, \ldots, 30$. It can be clearly seen that the *selection-registering* measures yield increasing values with increasing subset size, while the *selection-exclusion-registering* measures yield increasing values with subset size getting farther from $\frac{1}{2}|\mathrm{Y}|$ (the only exception being CW_{rel}). This behavior follows from the simple facts that with increasing subset size it is more likely that there will be more overlap among subsets, even when features are selected randomly, while with decreasing subset size there will be more overlap between *excluded* features. This effect ("the subset-size-bias problem") makes it difficult to compare the stability of multiple FS methods yielding differently sized subsets using any *subset-size-biased* measure.

The only measure capable of identifying randomness regardless subset size is CW_{rel}, which has been defined for this purpose. Its performance in this respect is well visible in Figs. 4 and 7. For randomly selected subsets, it yields values close to 0.

Note. If the size of selected subsets is close to $\frac{1}{2}|\mathrm{Y}|$, then randomness in the FS process results in $ANHI$ value close to 0.5 (see [1] for reasoning) and PH value close to 0 with a sufficient number of trials.

Intermeasures can be used to compare an FS process against a knowingly random FS process (assuming equal Y). Identifying randomness in this way is, however, possible only in cases when the average size of subsets in both compared systems is similar. (All considered intermeasures are *subset-size-biased*.) In such a case, the measures IC and ICW yield values close to 1 if both systems consist of randomly selected subsets (i.e., feature frequencies are roughly similar in both systems).

Fig. 8 illustrates the behavior of intermeasures when comparing two systems of random subsets of differing sizes. In both systems, 1,000 subsets are drawn randomly from Y, $|\mathrm{Y}| = 30$, with different subset sizes in each system, as indicated in Fig. 8. It can be seen that with increasing difference between the systems' subset sizes, all measures indicate lower systems' similarity, despite the fact that both systems always consist of random subsets.

Fig. 9 illustrates the behavior of intermeasures when comparing two systems of random subsets of equal sizes. In both systems, 1,000 subsets are drawn randomly from Y, $|\mathrm{Y}| = 30$, with equal subset sizes in both systems. It can be

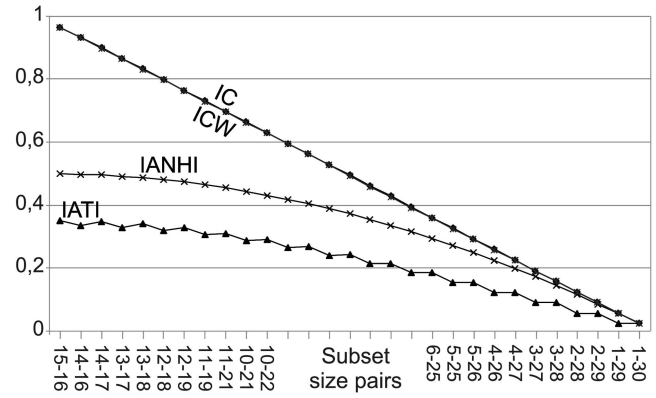

Fig. 8. Comparing two systems of random subsets of different sizes.

seen that the *feature-focused*, *selection-registering* IC and ICW clearly identify the closeness of feature frequencies, while the *subset-focused* $IATI$ and *selection-exclusion-registering* $IANHI$ evaluate system differences in more detail, which leads to emphasis of differences between the contents of the randomly selected subsets and eventually to lower measure values.

Remark. The shapes of graphs in Figs. 7 and 9 illustrate well the principal closeness between $ANHI$ and $IANHI$ and between ATI and $IATI$.

4.5 The Impact of Very High Problem Dimensionality

Many FS tasks today involve data sets of very high dimensionality, e.g., in genetics, image analysis, or text categorization. It is known that very high problem dimensionality causes serious problems in machine learning. Among others, the effects of the "curse of dimensionality" [14] seriously degrade the ability of learning algorithms to devise robust models what leads to degraded generalization ability. In FS context, very high dimensionality prevents many well-known sophisticated methods from being used at all due to search time complexity. Moreover, the effects of overselection [9] are emphasized.

High dimensionality also affects the information that can be gained using stability measures. Although the principle of the considered stability measures is dimensionality independent, higher dimensionality may shift values closer to bounds and consequently make them more difficult to interpret.

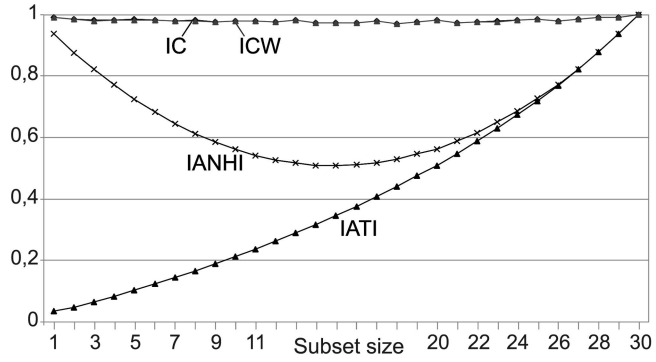

Fig. 9. Using intermeasures to compare two systems of random subsets of equal sizes.

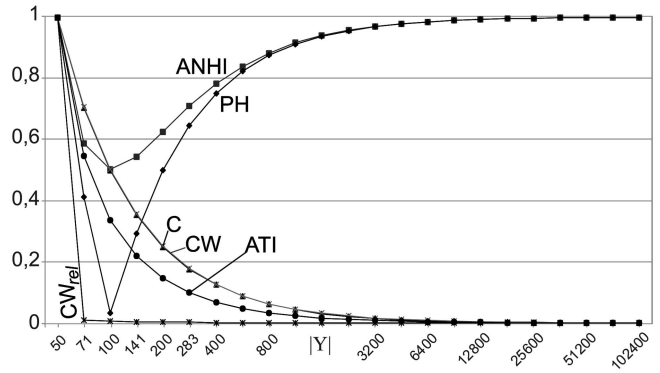

Fig. 10. Intrameasure behavior with increasing dimensionality ($d = 50$).

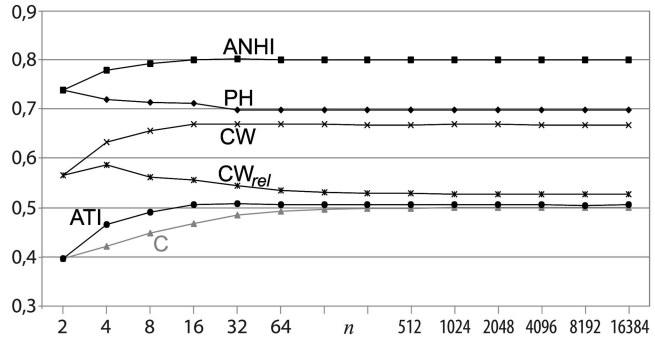

Fig. 11. Illustrating the stability of stability measures with respect to number of FS trials.

It is the increasing differences between selected subset size and problem dimensionality and accordingly changed sampling properties that lead to stability measure output shift. Fig. 10 illustrates this behavior. In the experiment, 1,000 subsets are drawn randomly for each $|\mathrm{Y}|$ with constant subset size $d = 50$.

4.6 Stability of Stability Measures

The value yielded by various stability measures depends on the size of the investigated system (number of FS trials). In Fig. 11, it can be seen that, in order to get reliable stability measure response, the number of evaluated FS trials should be reasonably high, preferably not lower than problem dimensionality. The experiment in Fig. 11 is repeated for various numbers of subsets "selected" from Y, where $|\mathrm{Y}| = 100$. Assuming fixed $\mathrm{Z}_1 \subset \mathrm{Y}$, $|\mathrm{Z}_1| = 15$ and fixed $\mathrm{Z}_2 \subset \mathrm{Y}$, $|\mathrm{Z}_2| = 60$, $\mathrm{Z}_1 \cap \mathrm{Z}_2 = \emptyset$, each subset $\mathrm{X} \subset \mathrm{Y}$ is "selected" so as to contain 15 consistently occurring and 15 less consistently occurring features, i.e., each subset $\mathrm{X} = \mathrm{Z}_1 \cup \mathrm{X}_2$, where features in X_2 are drawn randomly from Z_2 so that $|\mathrm{X}_2| = 15$.

5 Experimental Evaluation on Real Data

In this section, we investigate the behavior of all considered measures on real FS tasks using low-to-mid and high-dimensional data. We will also investigate the impact of modifying FS method parameters and the impact of improving estimator properties.

In order to illustrate the performance of the considered measures, we have conducted a series of FS experiments on standard data from the UCI Repository [15]: *wine* data (13-dim., 3 classes of 59, 71, and 48 samples), *wdbc* data (30-dim., 2 classes of 357 and 212 samples), *sonar* data (60-dim., 2 classes of 103 and 105 samples), *spectf* data (44-dim., 2 classes of 212 and 55 samples), *mammo* data (65-dim., 2 classes of 57 and 29 samples) and *cloud* data (10-dim., 2 classes of 1,024 and 1,024 samples). Note that the UCI data represent the type of real-world problems characterized by low to moderate problem dimensionality and limited (even insufficient) amount of training samples. This type of problem often appears in medicine (or economics), where data gathering is costly and access to patient (or company-internal) data is often restricted.

To illustrate another common type of classification problem, we evaluate all considered measures on high-dimensional text categorization task [16] using the *Reuters-21578* data[1] (10,105-dim., 33 classes, 2 classes dominant with 3,924 and 2,292 samples, others with less than 300 samples, total of 8,941 samples). The data have been preprocessed by means of removing all nonalphabetic characters, words containing nonalphanumeric characters, words with less than three occurrences, stopwords, and by means of Porter's stemming.

5.1 Experimental Setup: Search Methods

In our experiments, we use several FS methods of various properties and optimization performance. Apart from best individual features (BIF [17], [18]) and random selection, we investigate the family of sequential FS methods, covering methods of various properties and optimization strength, including Sequential Forward Selection [19], Sequential Forward Floating Selection [20], as well as the recent Dynamic Oscillating Search [12].

To simplify further discussion, let us overview the principle of sequential FS methods. Most of them share the same "core mechanism" of adding and removing features to/from a current subset. The respective algorithm steps can be described simply as follows (in nongeneralized form [21]):

Definition 12. *For a given current feature set X_d, let f^+ be the feature such that*

$$f^+ = \arg \max_{f \in \mathrm{Y} \setminus \mathrm{X}_d} J(\mathrm{X}_d \cup \{f\}), \tag{31}$$

where $J(\cdot)$ denotes the criterion function used to evaluate candidate feature subsets. Then, we shall say that $ADD(\mathrm{X}_d)$ is an operation of adding feature f^+ to the current set X_d to obtain set X_{d+1} if

$$ADD(\mathrm{X}_d) \equiv \mathrm{X}_d \cup \{f^+\} = \mathrm{X}_{d+1}, \quad \mathrm{X}_d, \mathrm{X}_{d+1} \subset \mathrm{Y}.$$

Definition 13. *For a given current feature set X_d, let f^- be the feature such that*

$$f^- = \arg \max_{f \in \mathrm{X}_d} J(\mathrm{X}_d \setminus \{f\}), \tag{32}$$

where $J(\cdot)$ denotes the criterion function used to evaluate candidate feature subsets. Then, we shall say that $RMV(\mathrm{X}_d)$ is

1. http://www.daviddlewis.com/resources/testcollections/reuters21578.

an operation of removing feature f^- from the current set X_d to obtain set X_{d-1} if

$$RMV(\mathrm{X}_d) \equiv \mathrm{X}_d \setminus \{f^-\} = \mathrm{X}_{d-1}, \quad \mathrm{X}_d, \mathrm{X}_{d-1} \subset \mathrm{Y}.$$

In order to simplify the notation for a repeated application of FS operations, we introduce notation

$$\mathrm{X}_{d+2} = ADD(\mathrm{X}_{d+1}) = ADD(ADD(\mathrm{X}_d)) = ADD^2(\mathrm{X}_d),$$
$$\mathrm{X}_{d-2} = RMV(\mathrm{X}_{d-1}) = RMV(RMV(\mathrm{X}_d)) = RMV^2(\mathrm{X}_d),$$

and, more generally,

$$\mathrm{X}_{d+\delta} = ADD^\delta(\mathrm{X}_d), \quad \mathrm{X}_{d-\delta} = RMV^\delta(\mathrm{X}_d).$$

Now the considered sequential FS methods can be described as follows:

Sequential Forward Selection (**SFS** *[19]*) yielding a subset of d features:
1) $\mathrm{X}_d = ADD^d(\emptyset)$.

Sequential Forward Floating Selection (**SFFS** *[20]*) yielding a subset of d features, with optional search-restricting parameter $\Delta \in \{0, 1, \ldots, |\mathrm{Y}| - d\}$:
1) Start with $\mathrm{X}_0 = \emptyset$, $k = 0$.
2) $\mathrm{X}_{k+1} = ADD(\mathrm{X}_k)$, $k = k + 1$.
3) Repeat $\mathrm{X}_{k-1} = RMV(\mathrm{X}_k)$, $k = k - 1$ as long solutions already known for the lower k improve.
4) If $k < d + \Delta$ go to 2.

Dynamic Oscillating Search (**DOS** *[12]*) yielding a subset of optimized size k, with search-restricting parameter $\Delta \geq 1$; default $\Delta = |\mathrm{Y}|$):
1) Start with $\mathrm{X}_k = ADD^3(\emptyset)$, k = 3. Set "oscillation cycle depth" to $\delta = 1$.
2) Compute $ADD^\delta(RMV^\delta(\mathrm{X}_t))$; if any intermediate subset X_i, $i \in \{k - \delta, \ldots, k\}$ is found better than X_k, let it become the new X_k with $k = i$, let $\delta = 1$ and restart step 2.
3) Compute $RMV^\delta(ADD^\delta(\mathrm{X}_t))$; if any intermediate subset X_j, $j \in \{k, \ldots, k + \delta\}$ is found better than X_k, let it become the new X_k with $k = j$, let $\delta = 1$ and go to 2.
4) If $\delta < \Delta$ let $\delta = \delta + 1$ and go to 2.

Specifically for the purpose of high-dimensional FS, we also include the simplest form of Oscillating Search [22] that is, unlike the methods above, extremely time efficient at the cost of reduced search effectivity.

Oscillating Search (**OS** *[22]*) sequentially improves given initial solution X_d. Here, in simplified form to enable high-dimensional FS:
1) $\mathrm{X}_d^- = ADD(RMV(\mathrm{X}_d))$; if X_d^- is better than X_d let it become the new X_d and restart step 1.
2) $\mathrm{X}_d^+ = RMV(ADD(\mathrm{X}_d))$; if X_d^+ is better than X_d let it become the new X_d and go to step 1.

The considered methods BIF, SFS, SFFS, and OS are d-parameterized. Because we primarily focus on stability measures that enable evaluating systems of subsets of varying size, we define a d-optimizing extension of d-parameterized methods. The respective d-parameterized method is applied repeatedly for each subset size $d = 1, \ldots, |\mathrm{Y}|$, then, among the $|\mathrm{Y}|$ results, the one with highest criterion value (and lowest subset size in case of ties) is eventually selected. We will refer to d-optimizing forms of BIF, SFS, and SFFS as to BIF*, SFS*, and SFFS*.

For comparison, we also include random selection, where both the subset-size choice and feature selection are performed randomly according to uniform value distribution without respect to any criteria or data.

5.2 Experimental Setup: Selection Criteria

We conducted two series of experiments. With UCI data, we tested the FS methods in the *wrapper* [23] setting, i.e., we used classifier accuracy as the FS criterion. We included three conceptually different classifiers (see, e.g., [13]) in our tests: *gaussian classifier* or bayesian classifier assuming normal distribution, *3-nearest neighbor* with majority voting (3NN), and *support vector machine* with radial basis function kernel (SVM) [24].

With the high-dimensional Reuters data, neither the *wrapper* setting nor the complex search algorithms are applicable due to computational complexity. Therefore, in this case, we resorted to *filter* [23] setting using only BIF and OS as search methods, considering three distance functions as criteria. We consider the multinomial model for bag of words representation for text documents [25]. To evaluate individual features, we employed the average mutual information between class of document and the word in the document known as the Information Gain (IG) [26], [27], [25] and the multiclass Individual Bhattacharyya distance (IB) for one feature corresponding to one word in the given vocabulary of different words that occur in the collection of documents [22]. To evaluate feature subsets within the OS course of search, we employed multiclass Bhattacharyya distance [22].

If not stated otherwise, the following setup was used in all experiments. The number of FS trials (size of evaluated system of subsets) was set to $n = 100$. From each data set, 25 percent of data in each class was reserved for testing and as such excluded from FS process. In each FS trial, 90 percent of the remaining data was randomly sampled to form a trial-local data set. In the *wrapper* FS setting, the criterion value has been obtained as the average over 10 classification rates obtained using 10-fold holdout, where, in each loop, the trial-local data had been randomly scattered to 60 percent training, 30 percent validation, and 10 percent unused data. In the *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.

5.3 Experiments: Evaluating Stability of Wrapper-Based Feature Selection

Tables 1, 2, and 3 and Figs. 12, 13, 14, 15, 16, and 17 collect the results obtained for each UCI data setup. Graphs show stability values, and tables report the classification rate and subset size as optimized by each respective FS process.

Note that the sequence BIF*, SFS*, SFFS*, and DOS roughly orders the considered FS methods according to growing complexity and optimization performance (and presumably growing risk of feature overselection [9]). The ordering is well visible in all Tables 1, 2, and 3 on the achieved criterion values.

TABLE 1
Feature Selection Results on Wine Data, 13-dim., 3-Class, and WDBC Data, 30-dim., 2-Class

		a) WINE data							b) WDBC data						
	FS	Crit. value		Classif. rate		Subset size		Time	Crit. value		Classif. rate		Subset size		Time
Wrap.	Meth.	Mean	S.Dv.	Mean	S.Dv.	Mean	S.Dv.	(h:m)	Mean	S.Dv.	Mean	S.Dv.	Mean	S.Dv.	(h:m)
Gauss.	rand	.442	*.072*	.429	*.106*	5.97	*3.50*	00:00	.900	*.061*	.912	*.073*	14.5	*8.09*	00:00
	BIF⋆	.601	*.023*	.532	*.036*	2.41	*.814*	00:00	.940	*.003*	.940	*.007*	22.7	*7.64*	00:00
	SFS⋆	.645	*.025*	.515	*.065*	3.27	*.772*	00:00	.962	*.004*	.947	*.014*	8.37	*3.51*	00:07
	SFFS⋆	.672	*.013*	.579	*.068*	3.61	*1.18*	00:01	.966	*.003*	.954	*.016*	9.18	*3.24*	00:57
	DOS	.676	*.013*	.557	*.066*	3.56	*1.04*	00:01	.971	*.002*	.961	*.013*	8.13	*2.45*	01:44
3NN	rand	.856	*.116*	.878	*.107*	6.72	*3.50*	00:00	.939	*.043*	.938	*.044*	14.4	*7.31*	00:00
	BIF⋆	.969	*.005*	.959	*.018*	9.14	*1.89*	00:00	.969	*.002*	.963	*.006*	23.7	*3.71*	00:02
	SFS⋆	.978	*.006*	.972	*.021*	7.71	*1.58*	00:00	.978	*.002*	.956	*.015*	11.5	*5.37*	00:30
	SFFS⋆	.985	*.003*	.968	*.022*	7.57	*1.81*	00:02	.981	*.002*	.958	*.012*	13.2	*5.16*	02:10
	DOS	.987	*.003*	.966	*.022*	6.51	*1.56*	00:04	.983	*.002*	.958	*.012*	10.8	*4.24*	05:56
SVM	rand	.861	*.116*	.892	*.125*	6.54	*3.17*	00:00	.945	*.049*	.953	*.054*	15.0	*8.56*	00:00
	BIF⋆	.983	*.005*	.932	*.023*	9.16	*1.43*	00:00	.977	*.003*	.979	*.001*	21.6	*3.28*	00:03
	SFS⋆	.986	*.004*	.951	*.026*	7.43	*1.70*	00:01	.983	*.002*	.967	*.012*	10.7	*4.25*	00:34
	SFFS⋆	.991	*.002*	.948	*.025*	8.28	*1.67*	00:06	.985	*.002*	.968	*.012*	12.5	*4.40*	02:31
	DOS	.993	*.002*	.946	*.021*	7.3	*1.64*	00:11	.987	*.000*	.966	*.011*	9.45	*3.11*	07:33

TABLE 2
Feature Selection Results on SONAR Data, 60-dim., 2-Class and SPECTF Data, 44-dim., 2-Class

		a) SONAR data							b) SPECTF data						
	FS	Crit. value		Classif. rate		Subset size		Time	Crit. value		Classif. rate		Subset size		Time
Wrap.	Meth.	Mean	S.Dv.	Mean	S.Dv.	Mean	S.Dv.	(h:m)	Mean	S.Dv.	Mean	S.Dv.	Mean	S.Dv.	(h:m)
Gauss.	rand	.593	*.070*	.531	*.096*	30.8	*16.6*	00:00	.776	*.035*	.759	*.046*	20.2	*11.8*	00:00
	BIF⋆	.705	*.016*	.507	*.045*	10.1	*4.51*	00:05	.800	*.001*	.783	*.020*	4.18	*7.59*	00:01
	SFS⋆	.795	*.015*	.551	*.091*	14.3	*4.97*	01:39	.806	*.004*	.758	*.034*	14.5	*5.16*	00:22
	SFFS⋆	.819	*.015*	.548	*.085*	16.7	*5.10*	10:53	.814	*.008*	.746	*.036*	12.3	*4.94*	01:33
	DOS	.835	*.012*	.581	*.102*	13.4	*4.08*	42:56	.821	*.007*	.750	*.034*	10.8	*4.66*	07:40
3NN	rand	.761	*.061*	.437	*.096*	29.6	*17.3*	00:00	.745	*.021*	.722	*.038*	21.8	*11.9*	00:00
	BIF⋆	.855	*.010*	.649	*.084*	20.7	*7.16*	00:01	.804	*.011*	.762	*.037*	6.12	*7.20*	00:01
	SFS⋆	.885	*.013*	.516	*.086*	20.7	*8.35*	00:30	.846	*.011*	.746	*.041*	8.31	*4.26*	00:19
	SFFS⋆	.906	*.012*	.496	*.076*	22.9	*8.31*	02:10	.859	*.012*	.752	*.039*	10.1	*5.28*	01:33
	DOS	.922	*.009*	.535	*.076*	15.8	*5.29*	07:15	.870	*.009*	.758	*.031*	7.63	*2.87*	03:31
SVM	rand	.737	*.050*	.606	*.110*	28.2	*17.3*	00:00	.789	*.018*	.774	*.029*	20.88	*12.9*	00:00
	BIF⋆	.823	*.013*	.608	*.057*	24.4	*14.4*	00:04	.815	*.007*	.783	*.014*	38.44	*6.34*	00:05
	SFS⋆	.875	*.014*	.614	*.046*	18.1	*9.03*	02:39	.850	*.010*	.766	*.032*	17.47	*7.80*	01:46
	SFFS⋆	.895	*.012*	.624	*.041*	17.9	*8.16*	13:41	.875	*.009*	.77	*.029*	12.31	*6.02*	06:32
	DOS	.913	*.008*	.620	*.039*	14.3	*4.62*	52:11	.894	*.007*	.776	*.029*	11.5	*2.36*	24:11

TABLE 3
Feature Selection Results on MAMMO Data, 65-dim., 2-Class and CLOUD Data, 10-dim., 2-Class

		a) MAMMO data							b) CLOUD data						
	FS	Crit. value		Classif. rate		Subset size		Time	Crit. value		Classif. rate		Subset size		Time
Wrap.	Meth.	Mean	S.Dv.	Mean	S.Dv.	Mean	S.Dv.	(h:m)	Mean	S.Dv.	Mean	S.Dv.	Mean	S.Dv.	(h:m)
Gauss.	rand	.645	*.016*	.666	*.037*	31.5	*17.6*	00:00	.759	*.238*	.748	*.226*	5.27	*2.38*	00:00
	BIF⋆	.701	*.019*	.685	*.057*	3.13	*1.97*	00:07	.997	*.001*	.994	*.004*	3.05	*2.15*	00:00
	SFS⋆	.746	*.020*	.671	*.081*	5.33	*2.38*	02:17	.998	*.001*	.993	*.013*	4.52	*1.37*	00:01
	SFFS⋆	.754	*.022*	.648	*.079*	5.56	*2.52*	09:21	.999	*.000*	.992	*.013*	4.41	*1.15*	00:03
	DOS	.788	*.019*	.631	*.072*	6.01	*1.69*	65:28	.999	*.000*	.981	*.017*	3.92	*1.10*	00:03
3NN	rand	.633	*.044*	.661	*.070*	30.9	*16.9*	00:00	.976	*.070*	.938	*.128*	4.84	*2.39*	00:00
	BIF⋆	.792	*.026*	.757	*.068*	10.9	*7.51*	00:00	1	*0*	1	*0*	1	*0*	00:07
	SFS⋆	.872	*.037*	.810	*.117*	10.1	*5.71*	00:07	1	*0*	1	*0*	1	*0*	00:23
	SFFS⋆	.909	*.035*	.886	*.102*	7.28	*4.58*	00:40	1	*0*	1	*0*	1	*0*	00:46
	DOS	.936	*.010*	.936	*.064*	5.43	*1.25*	02:33	1	*0*	1	*0*	1	*0*	01:41
SVM	rand	.687	*.048*	.690	*.041*	33.4	*18.0*	00:00	.929	*.134*	.881	*.151*	4.93	*2.65*	00:02
	BIF⋆	.794	*.022*	.699	*.040*	54.6	*14.9*	00:01	1	*0*	1	*0*	1	*0*	00:24
	SFS⋆	.890	*.015*	.777	*.071*	13.6	*8.21*	00:23	1	*0*	1	*0*	1	*0*	01:12
	SFFS⋆	.897	*.013*	.790	*.074*	15.8	*9.53*	02:16	1	*0*	1	*0*	1	*0*	02:09
	DOS	.919	*.008*	.779	*.076*	10.4	*3.96*	09:27	1	*0*	1	*0*	1	*0*	04:32

Let us point out some of the most notable phenomena that can be observed regarding the presented stability results. First, it can be seen that in most cases there is visible agreement among the considered measures in terms of trends—better (or worse) FS stability is reflected in higher (or lower) value of most of the stability measures. Notable correspondence in behavior can be seen especially among the measures within the *selection-registering* group and

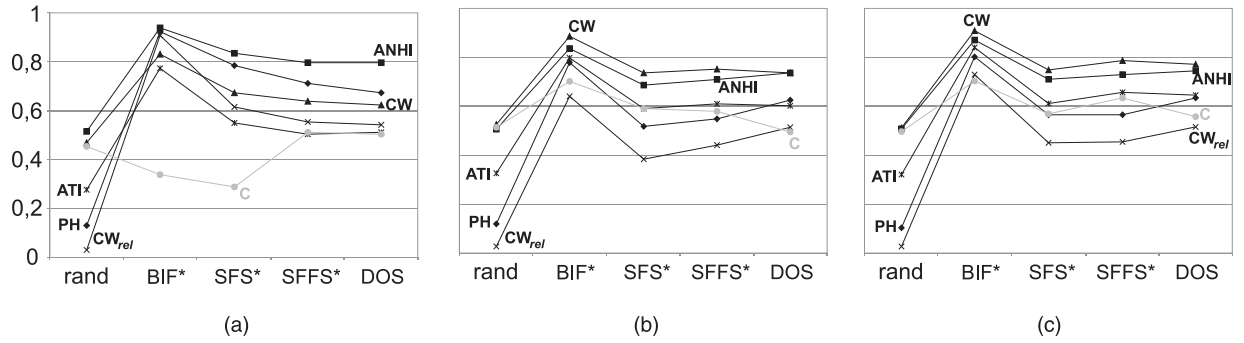

Fig. 12. Comparing feature selector stability on WINE data. (a) Wrapper: Gauss. (b) Wrapper: 3NN. (c) Wrapper: SVM.

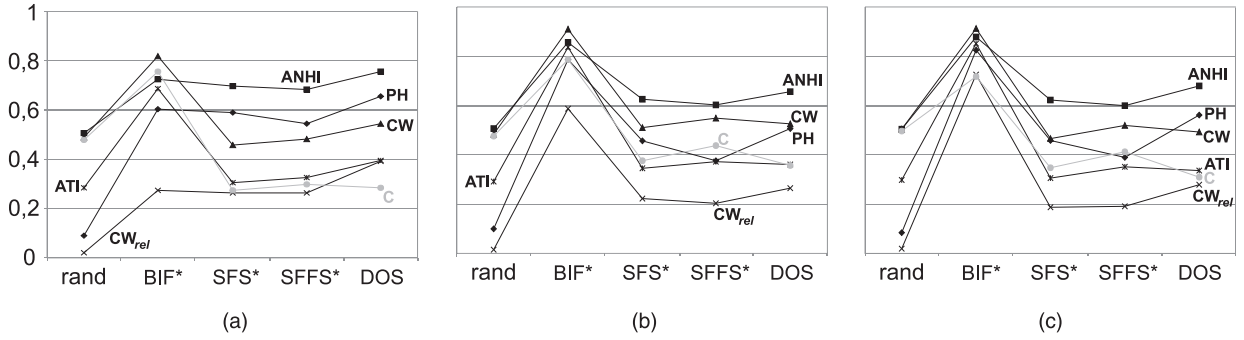

Fig. 13. Comparing feature selector stability on WDBC data. (a) Wrapper: Gauss. (b) Wrapper: 3NN. (c) Wrapper: SVM.

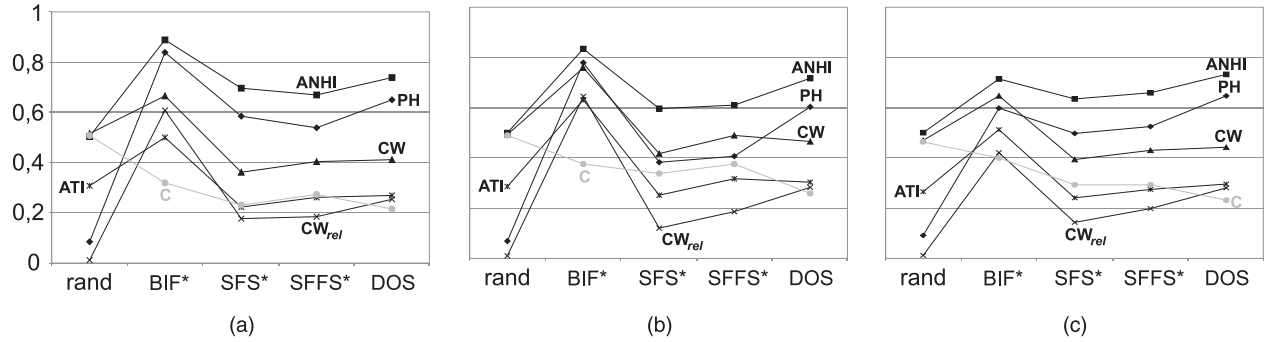

Fig. 14. Comparing feature selector stability on SONAR data. (a) Wrapper: Gauss. (b) Wrapper: 3NN. (c) Wrapper: SVM.

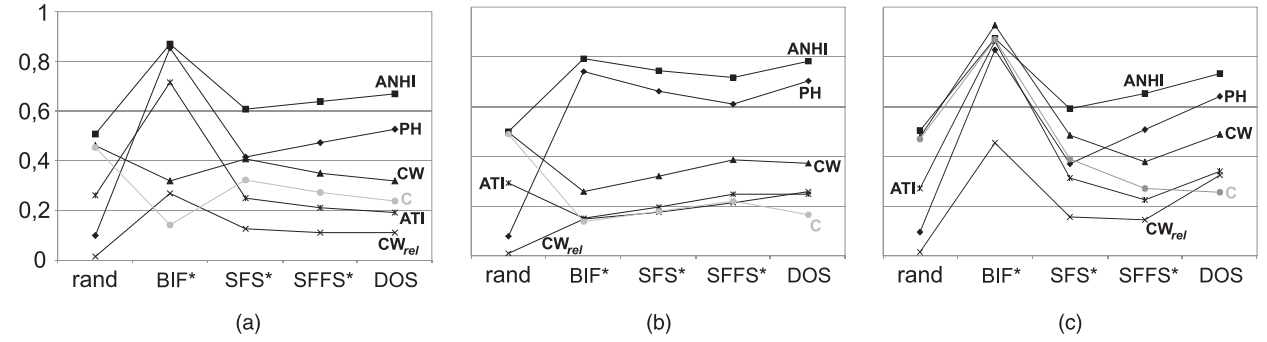

Fig. 15. Comparing feature selector stability on SPECTF data. (a) Wrapper: Gauss. (b) Wrapper: 3NN. (c) Wrapper: SVM.

within the *selection-exclusion-registering* group. (The information given by C is to be considered supplemental only due to its flaws, as reported in Section 4.) Second, the value level differs considerably among the considered measures, clearly showing the differences in their meaning. Third, the overall stability level in most experiments only rarely approaches 1, showing that FS tasks would be better approached with caution to prevent unaccounted failure of the devised decision rules.

According to [3], BIF is recommendable for cases when other FS methods fail to produce stable output. In our experiments, the most notable difference between stability measure values of BIF* and the other methods appears in Figs. 13b, 13c, 14a, 14b, 15a, and 15c. In accordance with [3], in these cases (with the exception of Guass. wrapper in

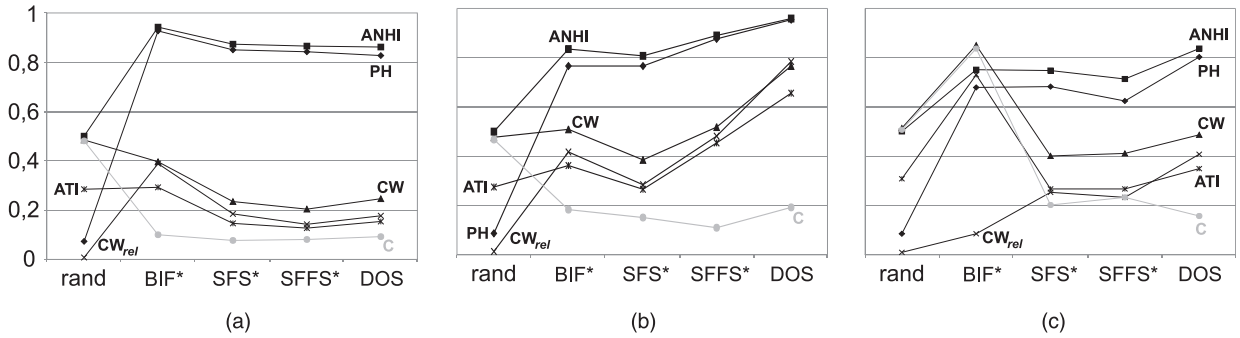

Fig. 16. Comparing feature selector stability on MAMMO data. (a) Wrapper: Gauss. (b) Wrapper: 3NN. (c) Wrapper: SVM.

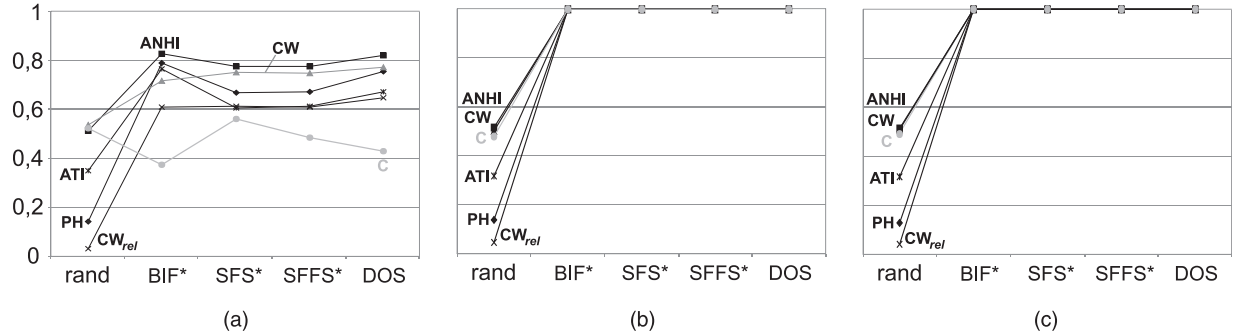

Fig. 17. Comparing feature selector stability on CLOUD data. (a) Wrapper: Gauss. (b) Wrapper: 3NN. (c) Wrapper: SVM.

Table 2a), BIF* proves to be the best performing FS method in terms of classification accuracy on independent data as confirmed by statistical significance t-test at significance level 0.05.

Less difference in terms of method stability can be observed among SFS*, SFFS*, and DOS, with DOS being most different from the other two methods. Despite its stronger optimization performance, DOS often yields more stable results than SFS* and SFFS* (see especially Figs. 13, 14, and 16). This can be explained by the fact that DOS is defined to guide the course of search toward smaller subsets.

The feature overselection [9] problem can be observed in Tables 1, 2, and 3 whenever BIF* overperforms other methods in terms of classification accuracy on independent data as well as in cases when the classification accuracy differs considerably from reported criterion value. Nevertheless, the examples also suggest that stronger feature selectors do not always overfit more than weaker selectors. This is especially the case if training data are sufficiently large with respect to dimensionality or the criterion to be optimized has sufficient generalization ability (e.g., Gaussian classifier in Fig. 13 and Table 1b, SVM in Fig. 14 and Table 2a, or 3NN in Fig. 16 and Table 3a, all cases confirmed by statistical significance t-test at significance level 0.05). Note also that DOS often yields the lowest variance in subset size among all considered methods.

The information given by various stability measures can complement each other to reveal more details of the evaluated FS process. Let us comment on several observations.

1. Note in Fig. 16a the consistently high difference between the values yielded by *selection-exclusion-registering* and *selection-registering* measures. This suggests that a large number of features are consistently excluded while the rest appear in the selected subsets with low stability. It suggests high redundancy among the limited number of features that get selected.
2. Note in Fig. 16 and Table 3a that CW_{rel} was the only stability measure to suggest a problem with BIF* on *mammo* data. In this case, BIF* produced wrong large feature subsets, resulting in poor classification performance (compare to random selection). Here, the *subset-size-biased* measures fail to identify wrong BIF* stability. Similarly as with *wdbc* data, compare the CW_{rel} stability reported for BIF* in Figs. 13a and 13b. Note in Table 1b that, with Gaussian wrapper, BIF* is overperformed by stronger FS methods while CW_{rel} is low, but, with 3NN, the opposite is true and CW_{rel} is high.
3. In Fig. 17a, note that the value of C is close to 0.5 for both DOS and random selection. The value of PH is high for DOS but very low for random selection. This may suggest that most of the features get selected either with very high or very low overall frequency, with neither group being prevalent. If this was not the case and higher or lower frequencies prevailed, then the C value would be more distant from 0.5. If a significant number of relative frequencies was distant from both 0 and 1, then the PH value for random selection would be higher and the PH value for DOS would be lower.

5.4 Experiments: Evaluating Stability of High-Dimensional Feature Selection

Figs. 18 and 19 and Table 4 collect the results obtained for the 10,105-dimensional Reuters data. The high dimensionality effectively prohibits the use of *wrappers* as well as search-based subset-size optimization. Thus, we followed the standard approach of selecting features by means of BIF based on *filter* criteria. Fig. 18a shows stability values for BIF

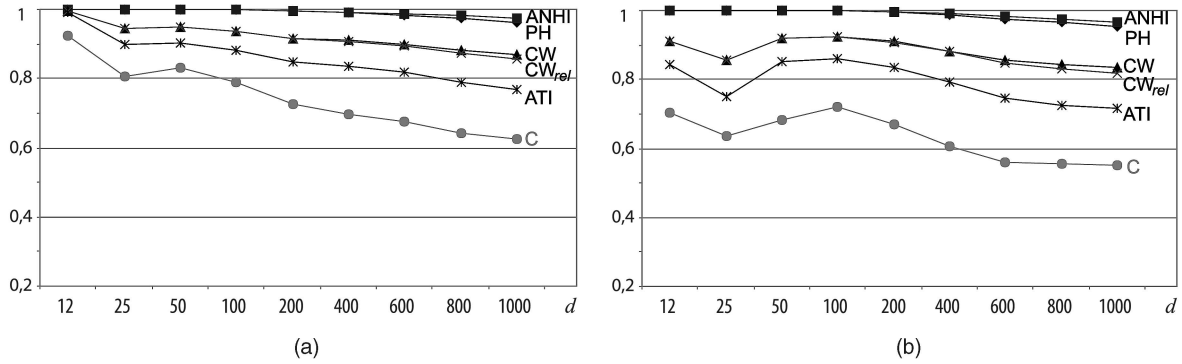

Fig. 18. Comparing BIF-IB and OS(BIF-IB) filter stability on high-dimensional REUTERS text data. (a) Filter: BIF-IB. (b) Filter: OS(BIF-IB, $\Delta = 1$).

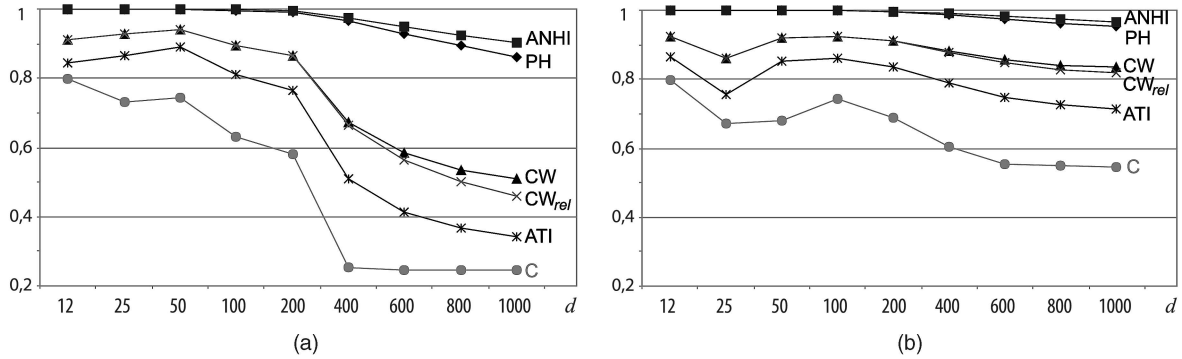

Fig. 19. Comparing BIF-IG and OS(BIF-IG) filter stability on high-dimensional REUTERS text data. (a) Filter: BIF-IG. (b) Filter: OS(BIF-IG, $\Delta = 1$).

TABLE 4
Classification Accuracy as Result of High-Dimensional Filter FS on Reuters Data

	a) BIF – IB					b) OS (BIF – IB)				
	Crit.value		Classif.rate		Time	Crit.value		Classif.rate		Time
d	Mean	*S.Dv.*	Mean	*S.Dv.*		Mean	*S.Dv.*	Mean	*S.Dv.*	
12	0.76	*.016*	.624	*.002*	7m	1.34	*.018*	.664	*.005*	25m
25	1.34	*.045*	.723	*.007*	7m	1.99	*.027*	.771	*.007*	1h
50	2.15	*.039*	.780	*.003*	7m	2.71	*.028*	.842	*.005*	2h
100	3.08	*.041*	.840	*.004*	7m	3.58	*.037*	.889	*.003*	5h
200	4.11	*.039*	.889	*.003*	7m	4.53	*.039*	.915	*.003*	25h
400	5.11	*.044*	.907	*.003*	7m	5.42	*.047*	.925	*.003*	55h
600	5.60	*.059*	.918	*.002*	7m	5.85	*.062*	.929	*.002*	133h
800	5.92	*.056*	.924	*.002*	7m	6.13	*.057*	.932	*.003*	223h
1000	6.15	*.056*	.927	*.003*	7m	6.34	*.057*	.933	*.002*	429h
	c) BIF – IG					d) OS (BIF – IG)				
	Crit.value		Classif.rate		Time	Crit.value		Classif.rate		Time
d	Mean	*S.Dv.*	Mean	*S.Dv.*		Mean	*S.Dv.*	Mean	*S.Dv.*	
12	0.77	*.080*	.596	*.017*	5m	1.34	*.021*	.662	*.003*	23m
25	1.32	*.026*	.701	*.006*	5m	2.00	*.026*	.770	*.007*	1h
50	2.08	*.039*	.788	*.005*	5m	2.71	*.030*	.841	*.005*	3h
100	2.77	*.047*	.831	*.005*	5m	3.60	*.038*	.889	*.003*	8h
200	3.60	*.041*	.871	*.003*	5m	4.54	*.041*	.916	*.003*	27h
400	4.02	*.057*	.881	*.003*	5m	5.41	*.053*	.925	*.003*	146h
600	4.13	*.061*	.879	*.003*	5m	5.85	*.058*	.929	*.003*	336h
800	4.15	*.059*	.877	*.003*	5m	6.14	*.057*	.931	*.002*	523h
1000	4.16	*.054*	.875	*.003*	5m	6.34	*.065*	.932	*.002*	920h

with Individual Bhattacharyya [22], and Fig. 19a with Information Gain [26], [25]. The respective classification accuracies using Multinomial Bayes classifier [25] are reported in Table 4. In addition to the two simple BIF setups, we include experiments aimed at improving the BIF solutions by means of subsequent OS-based search optimizing the Bhattacharyya distance [22]. In Table 4, OS is confirmed to be capable of improving both BIF-based solutions.

Table 4 shows slight superiority of IB over IG in BIF-based search. For subset sizes roughly above 200, the difference becomes more notable (better classification accuracy on independent data has been confirmed here by statistical significance t-test at significance level 0.05). Both BIF solutions are nevertheless overperformed by the OS solutions, most notably with subset sizes roughly up to 400 features (confirmed by statistical significance t-test at

TABLE 5
Comparing d-Optimizing and d-Parametrized Methods on MAMMO Data

Wrap.	FS Meth.	Crit. value Mean	Crit. value S.Dv.	Classif. rate Mean	Classif. rate S.Dv.	Subset size d Mean	Subset size d S.Dv.	PH	$ANHI$	CW	CW_{rel}	ATI	C	Time (h:m)
3NN	BIF*	.792	*.026*	.757	*.068*	10.89	*7.51*	.769	.835	.509	.420	.361	*.186*	00:00
	BIF	.735	*.034*	.775	*.081*	11	*0*	.816	.869	.614	.539	.450	*.261*	00:00
	SFS*	.872	*.037*	.810	*.117*	10.14	*5.71*	.768	.808	.386	.284	.268	*.152*	00:07
	SFS	.845	*.038*	.784	*.108*	10	*0*	.775	.819	.413	.313	.272	*.158*	00:01
	SFFS*	.909	*.035*	.886	*.102*	7.28	*4.58*	.879	.893	.521	.481	.454	*.112*	00:40
	SFFS	.875	*.044*	.829	*.118*	7	*0*	.866	.888	.482	.425	.344	*.118*	00:21

TABLE 6
Evaluating the Impact of More Thorough Estimation on SPECTF Data

Wrap.	Est. Meth.	Crit. value Mean	Crit. value S.Dv.	Classif. rate Mean	Classif. rate S.Dv.	Subset size Mean	Subset size S.Dv.	PH	$ANHI$	CW	CW_{rel}	ATI	C	Time (h:m)
3NN	ho3	.888	*.009*	.752	*.042*	8.47	*4.42*	.642	.723	.281	.123	.169	*.184*	01:47
+SFS*	ho10	.871	*.010*	.753	*.034*	8.01	*3.00*	.698	.775	.382	.253	.256	*.178*	05:34
	ho50	.869	*.010*	.757	*.022*	7.53	*1.84*	.810	.859	.587	.528	.496	*.201*	18:17
	ho100	.870	*.009*	.760	*.014*	7.77	*1.34*	.883	.902	.722	.684	.647	*.261*	33:20
	ho200	.869	*.009*	.758	*.017*	8.11	*1.11*	.942	.946	.854	.835	.804	*.346*	64:48
	ho300	.870	*.007*	.760	*.009*	8.22	*0.98*	.949	.952	.871	.865	.849	*.286*	97:46

significance level 0.05). Note that both OS initialized by BIF-IB and OS initialized by BIF-IG tend to converge to very similar solutions. This is confirmed by the notably similar graphs in Figs. 18b and 19b. Let us comment on several observations.

The overall stability of FS on Reuters data appears to be very high (compare to results in Sections 5.3 and 4.5) despite the high problem dimensionality. With increasing subset size it slightly declines, but for all considered d values it is clear that the importance of individual features is evaluated with high confidence throughout the FS process (note the very high $ANHI$ and PH values).

A sharp decline in *selection-registering* measure values can be observed for IG-based solutions of subset size 400 and larger. Unlike IB, IG ability to distinguish among less-important features apparently declines after about 200-400 most important features have been selected as confirmed in Table 4.

A slight local increase of *selection-registering* measure values can be observed in both Figs. 18 and 19 for subsets of roughly 50 features. Apparently, a group of preferable features roughly of this size can be well distinguished from the rest. However, redundancy is likely to be present within this group as indicated by lower stability measure values obtained when selecting only 25 features.

5.5 Experiments: Fixed versus Varying Subset Size

Table 5 illustrates the difference between d-optimizing and d-parameterized forms of several feature selectors. Fixed d value has been chosen for each d-parameterized experiment to be as closest as possible to the subset-size preference of the method's d-optimizing form.

The d-optimizing forms yield higher criterion values in accordance with the fact that their search space is larger. In terms of classification accuracy on independent data as well as FS stability, it appears that in this experiment stronger d-optimizing methods benefit more from the extended search scope. Note that SFFS* yields better classification accuracy on independent data as well as better stability than SFFS (better SFFS* classification accuracy confirmed by statistical significance t-test at significance level 0.05). With BIF* and BIF, the opposite is true (better BIF classification accuracy confirmed by statistical significance t-test at significance level 0.05).

5.6 Experiments: Evaluating the Impact of More Thorough Estimation

An important question is the reliability of classification accuracy estimation in the course of FS process. Tables 6 and 7 show experiments with increasing number of holdout loops. Table 6 collects experiments for 3NN-SFS* where more thorough estimation in *wrapper* criterion value computation has been tested with the aim to decrease feature preference fluctuations and consequently to improve poor FS stability on *spectf* data. The tables clearly show that more thorough estimation leads to better stability (indicated unanimously by all measures) and more stable subset-size preference at the cost of prolonged computation.

Table 7 compares the output of FS processes for 10-fold and 100-fold holdout estimation. Higher values indicate lower output change. Note that BIF* is indicated here by most measures in most cases (especially by the *selection-exclusion-registering* measures) as the method being the least affected by changing estimator performance. This confirms its suitability for problems where estimation may be difficult, i.e., in small sample problems.

TABLE 7
Comparing FS Output Based on 10-Fold and 100-Fold Holdout Estimation on WDBC Data

Wrap.	FS Meth.	IC	ICW	$IATI$	$IANHI$
Gauss.	BIF*	.814	.809	.775	.786
	SFS*	.910	.882	.320	.683
	SFFS*	.910	.887	.372	.720
3NN	BIF*	.979	.987	.838	.857
	SFS*	.883	.877	.377	.658
	SFFS*	.858	.861	.401	.637
SVM	BIF*	.941	.972	.878	.902
	SFS*	.894	.892	.334	.635
	SFFS*	.905	.920	.369	.627

TABLE 8
Experiment: Evaluating the Impact of FS Method Search Scope Extension on MAMMO Data

Wrap.	FS Meth.	Crit. value Mean	Crit. value S.Dv.	Classif. rate Mean	Classif. rate S.Dv.	Subset size Mean	Subset size S.Dv.	PH	AN HI	CW	CW^{rel}	ATI	C	Time (h:m)
gauss.	DOS(3)	.755	*.028*	.66	*.086*	3.76	*1.66*	.892	.911	.230	.201	.158	*.058*	00:02
	DOS(7)	.767	*.025*	.66	*.078*	4.69	*1.79*	.871	.890	.235	.193	.151	*.072*	00:05
	DOS(15)	.776	*.022*	.647	*.087*	5.57	*2.10*	.842	.863	.201	.142	.125	*.082*	00:24
	DOS(25)	.776	*.022*	.657	*.058*	5.63	*1.68*	.839	.865	.222	.164	.136	*.083*	01:57
3NN	DOS(3)	.887	*.063*	.863	*.106*	4.09	*1.27*	.935	.945	.564	.551	.426	*.102*	00:05
	DOS(7)	.922	*.038*	.911	*.097*	4.9	*1.11*	.948	.953	.685	.676	.562	*.140*	00:10
	DOS(15)	.931	*.018*	.929	*.066*	5.41	*1.43*	.952	.958	.747	.764	.645	*.146*	00:17
	DOS(25)	.933	*.011*	.945	*.049*	5.38	*1.51*	.956	.961	.766	.785	.660	*.171*	00:34
SVM	DOS(3)	.890	*.024*	.801	*.079*	5.93	*1.68*	.891	.911	.511	.473	.359	*.101*	00:14
	DOS(7)	.899	*.021*	.807	*.084*	6.41	*1.93*	.890	.907	.531	.506	.377	*.112*	00:30
	DOS(15)	.911	*.013*	.793	*.064*	7.98	*2.69*	.856	.883	.524	.466	.376	*.122*	01:09
	DOS(25)	.915	*.011*	.774	*.073*	9.16	*3.10*	.828	.858	.498	.428	.353	*.137*	02:25

5.7 Experiments: Evaluating the Impact of FS Method Parameter Change

The outcome of some FS methods depends on parameters to be set by user. Parameters usually affect the scope of search, allowing either faster or more thorough FS process. Similarly to the choice between simple and complex FS methods, setting parameters of a method may affect its susceptibility to feature overselection.

In Table 8, we collect results for various values of DOS parameter Δ on MAMMO data. It can be seen that higher Δ leads in all cases to higher achieved criterion value at the cost of higher computational time. The impact on classification accuracy on independent data, however, differs on each of the three tested wrappers. With the Gaussian wrapper, all *selection-registering* measures report very low values without clear connection to the Δ value, which coincides with the Gaussian wrapper's unsatisfactory performance in this case. With 3NN, all stability measures report considerable stability improvement with increasing Δ. High stability is accompanied here by high classification accuracy on independent data and low subset-size variance, confirming 3NN as the best wrapper choice in this experiment (3NN wrapper classification accuracy superiority confirmed here by statistical significance t-test at significance level 0.05). With SVM, the increase of Δ leads to a slight decrease of stability, as indicated by all measures. This undesirable behavior copies here the degradation of classifier generalization ability.

5.8 Experiments: Comparing the Output of Two Feature Selection Processes

Tables 9 and 10 collect the intermeasure results obtained for each considered *wrapper* setup. The information given by various intermeasures can reveal interesting details of the evaluated FS process. Let us comment on several observations.

1. In Table 9, a prevailing ordering can be recognized among FS method pairs (lowest output similarity first): BIF*-DOS, BIF*-SFFS*, BIF*-SFS*, SFS*-DOS, SFS*-SFFS*, SFFS*-DOS. This suggests that the newer (more complex) the method is, the less difference there is in its output with respect to its predecessor.
2. Apart from the trivial *cloud* data case in Table 9 where both 3NN and SVM identify correctly the single sufficient feature (see also Table 3), the highest agreement can be seen between all FS methods on *wine* data with 3NN and SVM. Accordingly, in Table 1, the difference between all FS methods remains only about 1 percent both in terms of criterion value and classification accuracy. The difference in subset-size preferences is low as well.
3. Note that in Table 9 for *wdbc* data, BIF* produces output considerably different from all other FS methods. Fig. 13 confirms that indeed even the BIF* stability differs considerably from the other methods. This observation puts the stability performance of all methods except BIF* on *wdbc* data in question.
4. In Table 10, for *wine* data, it can be seen that there is high similarity between feature subsets produced by 3NN and SVM but the output of Gaussian wrapper differs considerably from both 3NN and SVM (observable with all FS methods). This may suggest a problem with Gaussian wrapper, i.e., its ability to model *wine* data well enough. Accordingly, Table 1 confirms the poor performance of Gaussian classifier on *wine* data.
5. In Table 10, for *mammo* data and BIF*, it can be seen that there is high similarity between feature subsets produced by 3NN and gaussian wrapper but the output of SVM wrapper differs considerably from the other two. Table 3 confirms poor BIF* performance with SVM. Note that this is not the case with stronger feature selectors (see Table 10).

6 Conclusions

The primary purpose of evaluating FS stability is to reveal possible overtraining and other issues in machine learning process and consequently to prevent degraded performance of devised decision rules. FS stability measures can be additionally used to evaluate and compare properties of various FS methods and criteria as some tools may show to be inherently more stable than others. Consequently, the right tools for specific tasks can be chosen.

The notion of FS stability is difficult to formalize unanimously. Various measures can be defined, with each measure expressing a slightly different view of the problem, while none can give the full picture. Moreover, measure behavior is affected by factors like problem dimensionality or whether or not the FS process yields subsets of constant size.

We focused primarily on the problem of evaluating FS processes that optimize subset size, i.e., where subset sizes may vary across FS trials, as the battery of tools

TABLE 9
Comparing Outputs of Various Feature Selectors Using Intermeasures (a—IC, b—ICW, c—$IATI$, and d—$IANHI$)

		BIF*–SFS*	BIF*–SFFS*	BIF*–DOS	SFS*–SFFS*	SFS*–DOS	SFFS*–DOS	BIF*–SFS*	BIF*–SFFS*	BIF*–DOS	SFS*–SFFS*	SFS*–DOS	SFFS*–DOS	BIF*–SFS*	BIF*–SFFS*	BIF*–DOS	SFS*–SFFS*	SFS*–DOS	SFFS*–DOS
		WINE data						WDBC data						MAMMO data					
Gauss.	a	.887	.718	.75	.876	.903	.933	.505	.511	.458	.927	.893	.925	.945	.941	.921	.967	.947	.962
	b	.795	.616	.665	.811	.865	.933	.473	.484	.437	.896	.846	.916	.922	.896	.845	.948	.894	.934
	c	.514	.380	.419	.472	.509	.500	.271	.277	.248	.298	.306	.346	.176	.157	.127	.133	.122	.131
	d	.847	.773	.790	.786	.801	.794	.406	.404	.381	.679	.703	.710	.904	.897	.888	.868	.862	.860
3NN	a	.784	.745	.728	.914	.865	.917	.572	.598	.534	.920	.940	.907	.883	.868	.847	.94	.903	.956
	b	.782	.758	.718	.920	.864	.929	.513	.557	.481	.922	.931	.911	.795	.734	.677	.912	.855	.934
	c	.601	.582	.561	.589	.564	.598	.390	.428	.364	.352	.347	.359	.236	.256	.275	.326	.367	.542
	d	.674	.655	.652	.687	.684	.713	.491	.512	.466	.608	.637	.620	.791	.819	.840	.846	.871	.925
SVM	a	.812	.878	.838	.899	.918	.918	.578	.619	.560	.919	.911	.896	.367	.400	.314	.946	.926	.906
	b	.778	.860	.805	.903	.924	.931	.483	.551	.455	.914	.911	.905	.365	.396	.311	.937	.902	.900
	c	.621	.702	.665	.615	.618	.643	.354	.415	.342	.323	.310	.331	.209	.236	.164	.265	.289	.288
	d	.696	.762	.739	.702	.719	.729	.491	.533	.491	.605	.642	.629	.320	.341	.284	.729	.786	.766
		SONAR data						SPECTF data						CLOUD data					
Gauss.	a	.793	.75	.778	.931	.924	.929	.764	.802	.813	.913	.880	.933	.701	.702	.688	.934	.816	.881
	b	.7	.663	.637	.919	.904	.925	.730	.752	.751	.886	.845	.926	.686	.664	.569	.915	.764	.855
	c	.193	.179	.174	.237	.234	.259	.130	.111	.095	.215	.190	.197	.426	.407	.348	.602	.557	.601
	d	.724	.687	.723	.678	.711	.698	.669	.702	.721	.612	.615	.649	.679	.679	.670	.770	.747	.776
3NN	a	.731	.712	.738	.913	.872	.874	.900	.875	.890	.938	.941	.941	1	1	1	1	1	1
	b	.655	.652	.585	.898	.862	.873	.865	.802	.816	.916	.921	.934	1	1	1	1	1	1
	c	.282	.295	.268	.275	.246	.291	.135	.131	.126	.215	.211	.247	1	1	1	1	1	1
	d	.620	.611	.650	.597	.637	.648	.755	.726	.764	.727	.760	.748	1	1	1	1	1	1
SVM	a	.788	.764	.746	.937	.897	.921	.663	.495	.558	.829	.853	.918	1	1	1	1	1	1
	b	.691	.666	.622	.928	.886	.921	.631	.495	.553	.832	.853	.918	1	1	1	1	1	1
	c	.253	.250	.235	.253	.245	.278	.428	.414	.379	.737	.676	.821	1	1	1	1	1	1
	d	.587	.582	.599	.644	.671	.692	.487	.448	.450	.747	.696	.823	1	1	1	1	1	1

TABLE 10
Comparing the Output of Various Feature Selection Criteria

Data	Wrap.	BIF*			SFS*			SFFS*			DOS		
	Inter-meas.	Gaus. 3NN	Gaus. SVM	3NN SVM	Gaus. 3NN	Gaus. SVM	3NN SVM	Gaus. 3NN	Gaus. SVM	3NN SVM	Gaus. 3NN	Gaus. SVM	3NN SVM
WINE	IC	.428	.464	.857	.474	.498	.846	.485	.468	.812	.487	.494	.835
	ICW	.288	.312	.888	.361	.387	.832	.377	.404	.818	.359	.406	.838
	$IATI$	.187	.213	.760	.156	.168	.543	.164	.188	.561	.127	.169	.562
	$IANHI$	.397	.421	.801	.384	.413	.651	.387	.375	.653	.402	.407	.691
WDBC	IC	.851	.831	.926	.795	.800	.914	.746	.765	.898	.728	.739	.885
	ICW	.863	.849	.948	.706	.694	.908	.677	.682	.895	.611	.600	.869
	$IATI$	.714	.706	.820	.215	.195	.314	.229	.221	.343	.199	.186	.320
	$IANHI$	.753	.754	.842	.581	.578	.614	.540	.545	.584	.582	.599	.646
SONAR	IC	.709	.663	.786	.801	.844	.861	.780	.851	.810	.797	.826	.813
	ICW	.487	.509	.723	.757	.783	.828	.720	.771	.764	.679	.723	.726
	$IATI$	.244	.187	.424	.166	.165	.209	.201	.196	.230	.158	.180	.198
	$IANHI$	.689	.607	.691	.589	.621	.588	.565	.617	.582	.647	.680	.665
SPECTF	IC	.864	.659	.678	.774	.560	.399	.789	.341	.292	.772	.378	.307
	ICW	.719	.663	.659	.702	.526	.355	.721	.337	.286	.706	.367	.307
	$IATI$	.017	.511	.059	.114	.301	.146	.117	.271	.222	.081	.227	.174
	$IANHI$	.787	.577	.540	.596	.430	.310	.608	.303	.258	.651	.303	.269
MAMMO	IC	.849	.209	.324	.880	.851	.915	.888	.813	.867	.850	.838	.910
	ICW	.629	.197	.316	.715	.702	.892	.629	.663	.854	.535	.607	.867
	$IATI$	.153	.056	.166	.102	.108	.244	.088	.084	.274	.071	.073	.381
	$IANHI$	.837	.196	.281	.802	.762	.767	.833	.721	.779	.845	.780	.882

usable for this purpose has been very limited. First, we reviewed the currently available FS stability measures. Then, we proposed several new measures, provided modified or simplified forms of existing ones (e.g., feature-frequency-based form of Average Normalized Hamming Index), identified their principal differences, and eventually organized them in a unifying framework. We showed that the diverse measures may complement each other in evaluating the FS process.

We pointed out the "subset-size-bias problem." Most of the discussed measures (i.e., (11), (12), (18), (23), (24), and (26)-(29)) have been defined to yield values from $[0, 1]$, but their actual bounds depend on the size of subsets in the evaluated system and may be tighter than $[0, 1]$. These bounds make it difficult to compare measure values for systems of differently sized subsets. The *relative weighted consistency* measure has been devised to overcome the

problem and to allow more reliable comparison of the stability of various feature selectors.

Next, we introduced the family of *intermeasures*. Note that two processes that yield results with similar or equal stability (according to any one stability measure) may well differ in their preference of particular features. Intermeasures can be used for revealing this difference.

The considered measures have been evaluated on a series of experiments. In the experiments, we investigated the properties of various feature selectors, the impact of very high dimensionality as well as changing estimator properties. It has been confirmed that in cases of severely unstable FS performance, it is recommended to resort to the simple *best individual features* FS method. The feature overselection [9] problem that may affect stronger FS methods often hinders FS results and leads to degraded classification performance on independent data. Nevertheless, strong selectors (as the *dynamic oscillating search*) have been found best performing and/or the most stable ones in several of our examples. Thus, it is recommended to select the right tool for each task with caution, possibly with assistance of some of the discussed measures.

ACKNOWLEDGMENTS

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

REFERENCES

[1] K. Dunne, P. Cunningham, and F. Azuaje, "Solutions to Instability Problems with Sequential Wrapper-Based Approaches to Feature Selection," Technical Report TCD-CD-2002-28, Dept. of Computer Science, Trinity College, 2002.

[2] A. Kalousis, J. Prados, and M. Hilario, "Stability of Feature Selection Algorithms," *Proc. Fifth IEEE Int'l Conf. Data Mining,* pp. 218-225, 2005.

[3] L.I. Kuncheva, "A Stability Index for Feature Selection," *Proc. 25th IASTED Int'l Multi-Conf. Artificial Intelligence and Applications,* pp. 421-427, 2007.

[4] P. Křížek, J. Kittler, and V. Hlaváč, "Improving Stability of Feature Selection Methods," *Proc. 12th Int'l Conf. Computer Analysis of Images and Patterns,* pp. 929-936, 2007.

[5] A. Kalousis, J. Prados, and M. Hilario, "Stability of Feature Selection Algorithms: A Study on High-Dimensional Spaces," *Knowledge and Information Systems,* vol. 12, no. 1, pp. 95-116, 2007.

[6] P. Somol and J. Novovičová, "Evaluating the Stability of Feature Selectors that Optimize Feature Subset Cardinality," *Proc. Joint IAPR Int'l Workshop Structural, Syntactic, and Statistical Pattern Recognition,* pp. 956-966, 2008.

[7] S. Loscalzo, L. Yu, and C.H.Q. Ding, "Consensus Group Stable Feature Selection," *Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining,* J.F. Elder, IV, F. Fogelman-Soulié, P.A. Flach, and M.J. Zaki, eds., pp. 567-576, http://doi.acm.org/10.1145/1557019.1557084, 2009.

[8] Y. Saeys, T. Abeel, and Y.V. de Peer, "Towards Robust Feature Selection Techniques," *Proc. Belgian-Dutch Conf. Machine Learning,* pp. 45-46, 2008.

[9] S. Raudys, "Feature Over-Selection," *Lecture Notes in Computer Science,* vol. 4109, pp. 622-631, Springer, 2006.

[10] H. Vafaie and K.D. Jong, "Genetic Algorithms as a Tool for Feature Selection in Machine Learning," *Proc. 1992 IEEE Int'l Conf. Tools with AI,* pp. 200-204, 1992.

[11] F. Hussein, R. Ward, and N. Kharma, "Genetic Algorithms for Feature Selection and Weighting, A Review and Study," *Proc. Int'l Conf. Document Analysis and Recognition,* p. 1240, 2001.

[12] P. Somol, J. Novovičová, P. Pudil, and J. Grim, "Dynamic Oscillating Search Algorithm for Feature Selection," *Proc. 19th Int'l Conf. Pattern Recognition,* Dec. 2008.

[13] R. Duda, P. Hart, and D. Stork, *Pattern Classification and Scene Analysis.* J. Wiley, 2001.

[14] R.E. Bellman, *Adaptive Control Processes.* Princeton Univ. Press, 1961.

[15] A. Asuncion and D. Newman, "UCI Machine Learning Repository," http://www.ics.uci.edu/~mlearn/MLRepository.html, 2007.

[16] F. Sebastiani, "Machine Learning in Automated Text Categorization," *ACM Computing Surveys,* vol. 34, no. 1, pp. 1-47, Mar. 2002.

[17] T. Cover, "The Best Two Independent Measurements Are Not the Two Best," *IEEE Trans. Systems, Man, and Cybernetics,* vol. 4, no. 1, pp. 116-117, Jan. 1974.

[18] A.K. Jain, R.P.W. Duin, and J. Mao, "Statistical Pattern Recognition: A Review," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 1, pp. 4-37, Jan. 2000.

[19] A.W. Whitney, "A Direct Method of Nonparametric Measurement Selection," *IEEE Trans. Computers,* vol. 20, no. 9, pp. 1100-1103, Sept. 1971.

[20] P. Pudil, J. Novovičová, and J. Kittler, "Floating Search Methods in Feature Selection," *Pattern Recognition Letters,* vol. 15, pp. 1119-1125, 1994.

[21] P.A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach.* Prentice-Hall Int'l, 1982.

[22] J. Novovičová, P. Somol, and P. Pudil, "Oscillating Feature Subset Search Algorithm for Text Categorization," *Lecture Notes in Computer Science,* vol. 4225, pp. 572-587, Springer, 2006.

[23] R. Kohavi and G. John, "Wrappers for Feature Subset Selection," *Artificial Intelligence,* vol. 97, pp. 273-324, 1997.

[24] C.-C. Chang and C.-J. Lin, *LIBSVM: A Library for Support Vector Machines,* http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[25] A. McCallum and K. Nigam, "A Comparison of Event Models for Naive Bayes Text Classification," *Proc. AAAI-98 Workshop Learning for Text Categorization,* pp. 41-48, 1998.

[26] G. Forman, "An Experimental Study of Feature Selection Metrics for Text Categorization." *J. Machine Learning Research,* vol. 3, pp. 1289-1305, 2003.

[27] Y. Yang and J.O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," *Proc. 14th Int'l Conf. Machine Learning,* pp. 412-420, citeseer.ist.psu.edu/yang97comparative.html, 1997.

Petr Somol received the MSc and the Rerum Naturalium Doctoris degrees in informatics and the PhD degree in computer science from the Faculty of Mathematics and Physics, Charles University, Prague. He is currently with the Department of Pattern Recognition at the Institute of Information Theory, Academy of Sciences of the Czech Republic. He worked in the Medical Informatics Unit at Cambridge University from 2000 to 2002. He worked for the RealReflect EC project (www.realreflect.org) from 2002 to 2005. His current interests include statistical approach to pattern recognition (feature selection, modeling, classification), combinatorial algorithms, data mining, modern programming, information presentation, graphics, and typography.

Jana Novovičová received the MSc degree in mathematics and the Rerum Naturalium Doctoris degree in mathematical statistics, both from the Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic, and the PhD degree in theoretical cybernetics from the Institute of Information Theory and Automation, Czech Academy of Sciences (UTIA CAS), Prague. She is with the Department of Pattern Recognition, UTIA CAS, and is an associate professor of informatics with the Faculty of Transportation Sciences, Czech Technical University. Her research interests include statistical approach to pattern recognition (feature selection, classification, and mixture models) and text document classification.