Akademie věd České republiky
Ústav teorie informace a automatizace

Academy of Sciences of the Czech Republic
Institute of Information Theory and Automation

RESEARCH REPORT

Somol P., Grim J.
Institute of Information Theory and Automation
Academy of Sciences of the Czech Republic

Fast Dependency-Aware Feature Selection in Very-High-Dimensional Pattern Recognition Problems

No. 2295 February 2011

ÚTIA AV ČR, v.v.i., P.O. Box 18, 182 08 Prague, Czech Republic
E-mail: {somol,grim}@utia.cas.cz
Tel: (+420)266052215
Fax: (+420)284683031

Fast Dependency-Aware Feature Selection in Very-High-Dimensional Pattern Recognition Problems

Petr Somol and Jiří Grim
Dept. of Pattern Recognition, Inst. of Information
Theory and Automation, Czech Academy of Sciences
Pod vodárenskou věží 4, 182 08 Prague 8, Czech Republic
{somol, grim}@utia.cas.cz

***Abstract*—The paper addresses the problem of making dependency-aware feature selection feasible in pattern recognition problems of very high dimensionality. The idea of individually best ranking is generalized to evaluate the contextual quality of each feature in a series of randomly generated feature subsets. Each random subset is evaluated by a criterion function of arbitrary choice (permitting functions of high complexity). Eventually, the novel dependency-aware feature rank is computed, expressing the average benefit of including a feature into feature subsets. The method is efficient and generalizes well especially in very-high-dimensional problems, where traditional context-aware feature selection methods fail due to prohibitive computational complexity or to over-fitting. The method is shown well capable of over-performing the commonly applied individual ranking which ignores important contextual information contained in data.**

***Index Terms*—feature selection, high dimensionality, ranking, generalization, over-fitting, stability, classification, machine learning, pattern recognition**

I. Introduction

Dimensionality reduction (DR) is a vital step in many machine learning and pattern recognition tasks. It is capable of improving model interpretability, recognition performance, accuracy, as well as economy of the designed system. In the most general DR form – *feature extraction* (FE) [1] – a set of new features is generated using a suitable transformation from all original measurements. FE may reveal information hardly accessible in original data. In the special simplified case of FE – in *feature selection* (FS) [2]–[6] – also known as variable or attribute selection, and in the more general *feature weighing* (FW) [7] the original measurement meaning is preserved; discarding irrelevant and redundant information thus can reduce measurement acquisition cost as well as speed up the learning process. FS is widely applied in various fields; in supervised and unsupervised learning [1], [6], [8], [9], remote sensing [10], [11], document categorization [12], image retrieval [13], bio-informatics [14], [15], medical diagnostics [16], etc.

An extensive battery of various types of FS tools is now available (for overview see, e.g., [1], [3], [6], [8], [9], [17], [18]). The common classification [1], [3] is to *wrappers* (direct maximization of classification accuracy), *filters* (assessing the merit of features directly from data), *embedded* (where FS is part of model optimization) and *hybrids* (combining the merits of several approaches). Nevertheless, no single method can be claimed generally best for all types of problems. Properties like search speed, ability to find close-to-optimal results and robustness against over-fitting are often contradicting each other. In Figure 1 we illustrate briefly the applicability of representative tools from current FS frameworks.

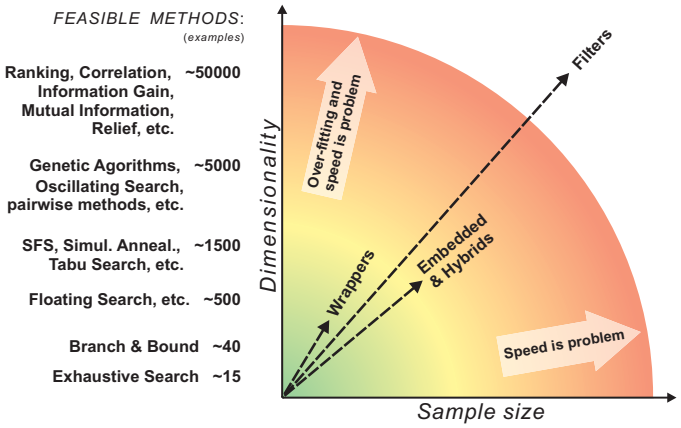

Fig. 1. Particular Feature Selection methods are usually well suited only for particular problem settings. With increasing dimensionality and sample size the battery of applicable FS tools diminishes quickly.

In recent years the focus of FS research is moving from the relatively well covered area of low-to-mid-dimensional recognition problems towards high- and very-high-dimensional problems [19]. Very high dimensionality is common, e.g., in bio-informatics and gene search [14], [15], text processing [12], [20]–[22], image analysis [23], [24], etc. Very high dimensionality is particularly challenging for two reasons: curse of dimensionality and computational complexity.

A. Curse of dimensionality

High-dimensional FS is more prone to problems following from insufficient sample size with respect to problem dimensionality, although such situation can appear even in low-to-mid-dimensional problems as is common, e.g., in economics or medicine [16], where the number of observed cases is

often too limited. In cases of insufficient data sample size it may be questioned what information about features is reliably obtainable from the data at all [25]–[27]. The commonly suggested work-around is to refrain from complex analysis of feature subsets in favor of simpler FS methods or even trivial feature ranking, also known as Best Individual Features (BIF) method [28]–[30]. It is commonly assumed that ignoring inter-feature dependencies is less harmful than obtaining misleading information through serious estimation errors due to over-fitting.

B. FS Computational Complexity

Computational complexity increases exponentially with the problem dimensionality due to a growing number of possible feature subset combinations (see Figure 1 for illustration). Optimal subset search methods are applicable roughly up to 40-dimensional problems, sub-optimal methods based on various forms of hill-climbing are applicable roughly with hundreds of features. Very-high-dimensional FS problems effectively prohibit the use of sophisticated subset optimization schemes. Parallelized search can extend this limitation only to minor extent [31]. This is another reason for the popularity of ranking methods, which often constitute the only computationally feasible option.

Wrapper-based FS [3] is considered not feasible in very-high-dimensional FS, the common approach is filter-based. Various correlation [32], [33] and information measures [17], [34] have gained popularity in recent years due to favorable trade-off between evaluation complexity and informational ability, in addition to the traditional probabilistic FS criteria [8].

C. Randomization

The exponential burden can be in some cases tackled by the use of randomized methods, where the search process can be user-restricted by a time limit. This is at the cost of optimality, yet many methods are capable of providing sufficiently good results. The Relief algorithm [35], [36] is based on a simple idea of repeated randomized sampling of one pattern followed by feature weights update according to the nearest-hit (same class) and the nearest-miss (different class) neighbours. Relief has proven to be very effective under various scenarios and as such has been studied and extended in many ways [7], [37].

Genetic algorithms [38], [39], Tabu Search [39], [40] and Simulated Annealing [39], [41] all implement a reasonably strong optimization mechanism and can be used with arbitrary FS criteria, but in very-high-dimensional setting may need considerable time to converge. Hybrid randomized-greedy algorithms have been shown better suited for high-dimensional problems in some cases [42], [43]. Pure Monte Carlo methods can easily miss good solutions in high-dimensional FS problems. The potential of randomized FS has been further exploited by performing the complete FS process repeatedly on various random subspaces to eventually combine the information about selected feature subsets into the final feature ranking [44]. This approach randomly restricts evaluation of inter-feature dependencies to reduce over-fitting.

Few methods of the above can cope with very high dimensionality unless refraining from taking complex dependencies among features into account.

D. This paper

The contribution of this paper consists in a novel approach to feature selection, particularly suitable for high and very-high dimensional pattern recognition problems. It utilizes randomization and an arbitrarily chosen FS criterion function to evaluate the average behavior of each feature over various contexts. The proposed method will be shown computationally very effective and capable of considerably over-performing the common methods of choice in very-high-dimensional domain – methods of BIF type. In this sense it notably extends the applicability of complex feature subset evaluation functions that are traditionally considered applicable with limited dimensionality only (e.g., wrapper criteria, see above).

In Sections II to IV the new method is introduced. In Section V we illustrate that the method generalizes well, i.e., is capable of improving recognition accuracy on independent data as well as with multiple decision rules.

II. Preliminaries

Assume a general pattern recognition problem (typically a classification or clustering problem) in N-dimensional feature space. In the particular case of classification, some objects described by means of features $f_1, f_2, \ldots, f_N$ (real valued or discrete) are to be classified into one of a finite number of mutually exclusive classes. The common initial step in classifier design is to choose a reasonably small subset of informative features by using a feature selection method.

Denoting F the set of all features

$$F = \{f_1, f_2, \ldots, f_N\} \tag{1}$$

we assume that for each subset of features $S \subset F$ a feature selection criterion $J(\cdot)$ can be used as a measure of quality of S (typically but not necessarily from the classification point of view). According to standard FS paradigm the resulting feature subset is obtained by maximizing $J(S)$ over the class of all subsets $S \subset F$.

Here we do not impose any restrictions on the function $J(\cdot)$ except that we expect it to be capable of reflecting feature behavior in context, i.e., it should provide more than just combined information on individual feature merit.

A specific feature selection problem arises in case of very high dimensionality, e.g., if $N \approx 10^3 \div 10^6$ or even more. Even the simplest sub-optimal optimization techniques are exceedingly time-consuming in such cases and, consequently, only very basic tools can be used to optimize features. The common approach is a simple ranking of features based on the individual feature quality (cf. BIF). By ordering the features according to the inequality

$$J(\{f_{i_{n-1}}\}) \leq J(\{f_{i_n}\}); \quad n = 2, 3, \ldots, N \tag{2}$$

we can easily identify a subset of d individually best features $f_{i_{N-d+1}}, f_{i_{N-d+2}}, \ldots, f_{i_N}$ but, in this way, we completely

ignore the potentionally crucial dependence among features and the resulting subset thus may be far from optimal.

In this paper we attempt to generalize the idea of individually best ranking by evaluating the quality of each feature repeatedly in the context of randomly chosen feature subsets. In other words, we evaluate the quality $J(\{f_n\} \cup S)$ of the subset $\{f_n\} \cup S$ for sufficiently many random subsets $S \subset F, (f_n \notin S)$ and compare the corresponding mean value with the analogous mean of $J(S)$ for subsets $S \subset F$ not containing the feature f_n, (i.e. $f_n \notin S$).

This idea is based on the intuitive assumption that "good" features exhibit reasonably consistent behavior in context with other features, that this information is obtainable easily enough and that it can improve upon the information about individual feature quality. An analogous mechanism has been shown to perform well in Fast Branch & Bound algorithm [18], where feature behavior is studied in variable context and the averaged information is utilized to predict $J(\cdot)$ values, enabling considerable acceleration of the search process.

III. Dependency-Aware Feature Rank

The starting point of the proposed dependency-aware feature ranking is a randomly generated sequence of feature subsets, to be denoted *probe* subsets

$$\mathbb{S} = \{S_1, S_2, \ldots, S_K\}, \quad S_j \subset F, \quad j = 1, 2, \ldots, K, \qquad (3)$$

where each subset is evaluated by a criterion function $J(\cdot)$. The cardinality of the subsets $S \in \mathbb{S}$ should vary and the resulting sequence $\mathbb{S}$ should be long enough to "approximate" the class of all possible subsets of F in a reasonably uniform way. For each feature $f \in F$ there should be enough subsets in $\mathbb{S}$ that do contain it as well as enough subsets that don't.

To generate a random probe subset we use the following simple procedure: first the subset size d is randomly chosen so that $d \in [1, \min\{N, \tau\}]$ where $\tau \in [1, N]$ is an optional user-specified upper limit. Next, indexes of features to be selected are randomly generated from $[1, N]$ as long as the number of unique feature indexes is lower than d.

Another possibility to generate a random probe subset $S \subset F$ is to decide the choice randomly for all possible features $f \in F$. In particular, assume that each feature $f \in F$ is included in the set S with probability p and it is not included with the complementary probability $(1-p)$. In this way the probability that the resulting set contains exactly d features is polynomial

$$P\{|S| = d\} = \binom{N}{d} p^d (1-p)^{(N-d)}, \quad 0 \le d \le N, \qquad (4)$$

the mean number of features in the sets S is pN and the resulting sets are bounded by the inequality $0 \le |S| \le N$. Note that by means of the probability p we can control the mean number of features in S without any strict limits while keeping the "natural" proportions of differently large subsets.

The required size of $\mathbb{S}$ as well as the role of parameter τ is discussed in Section V and V-C.

A. Dependency-Aware Rank Definition

Given a sufficiently large sequence of feature subsets $\mathbb{S}$, we are interested to utilize the information contained in the criterion values $J(S_1), J(S_2), \ldots, J(S_K)$ in depth. Instead of measuring the classification "power" of individual features $f \in F$, we compare the quality of probe subsets containing f with the quality of probe subsets not including f.

A straightforward idea is to compute the mean quality μ_f of subsets $S \in \mathbb{S}$ containing the considered feature $f \in F$

$$\mu_f = \frac{1}{|\mathbb{S}_f|} \sum_{S \in \mathbb{S}_f} J(S), \qquad \mathbb{S}_f = \{S \in \mathbb{S} : f \in S\} \qquad (5)$$

and the mean quality $\bar{\mu}_f$ of subsets $S \in \mathbb{S}$ not containing the considered feature f:

$$\bar{\mu}_f = \frac{1}{|\bar{\mathbb{S}}_f|} \sum_{S \in \bar{\mathbb{S}}_f} J(S), \quad \bar{\mathbb{S}}_f = \{S \in \mathbb{S} : f \notin S\} \qquad (6)$$

with the aim to use the difference of both values as a criterion for ranking the features:

$$DAF_0(f) = \mu_f - \bar{\mu}_f, \quad f \in F. \qquad (7)$$

Note that the "dependency aware" ranking criterion DAF_0 does not measure the individual quality of a feature $f \in F$ separately by means of the criterion value $J(\{f\})$, but takes into account the quality of feature f in the context of other features occurring in the sets $S \in \mathbb{S}$. The value $DAF_0(f)$ can be viewed as the average benefit of including the feature f into the feature subsets $S \in \mathbb{S}$.

An essential advantage of this approach is the computational "economy" since each probe subset $S \in \mathbb{S}$ can be used in the Eqs. (5) as many times as there are features in S and, analogously, $(N - |S|)$-times in the Eqs. (6).

A question arises if the criterion DAF_0 could be "refined" for the sake of ranking comparisons. One possibility is to compute the variances related to the mean values $\mu_f, \bar{\mu}_f$

$$\sigma_f^2 = \frac{1}{|\mathbb{S}_f|} \sum_{S \in \mathbb{S}_f} [J(S) - \mu_f]^2, \quad f \in F \qquad (8)$$

$$\bar{\sigma}_f^2 = \frac{1}{|\bar{\mathbb{S}}_f|} \sum_{S \in \bar{\mathbb{S}}_f} [J(S) - \bar{\mu}_f]^2, \quad f \in F \qquad (9)$$

and to use them to norm Eq. (7), resulting in an expression analogous to Mahalanobis univariate probabilistic inter-class distance [8] in normal case:

$$DAF_1(f) = \frac{(\mu_f - \bar{\mu}_f)|\mathbb{S}|}{|\mathbb{S}_f|\sigma_f + |\bar{\mathbb{S}}_f|\bar{\sigma}_f}; \quad f \in F. \qquad (10)$$

Another possibility is to norm directly the criterion values $J(S)$ because the variability of $J(S)$ could be undesirably influenced by the number of features $|S|$. Intuitively, the variance of the criterion $J(S)$ should change proportionately to the size of $\mathbb{S}$. In this sense, denoting $\mu(d), \sigma(d)$ the mean

and variance of the values $J(S)$ for all subsets $S \in \mathbb{S}$ of cardinality d,

$$\mu(d) = \frac{1}{|\mathbb{S}(d)|} \sum_{S \in \mathbb{S}(d)} J(S), \quad \mathbb{S}(d) = \{S \in \mathbb{S} : |S| = d\}, \tag{11}$$

$$\sigma(d)^2 = \frac{1}{|\mathbb{S}(d)|} \sum_{S \in \mathbb{S}(d) f} [J(S) - \mu(d)]^2, \tag{12}$$

we can norm the criterion values $J(S)$ directly in the Eqs. (5), (6):

$$\theta_f = \frac{1}{|\mathbb{S}_f|} \sum_{S \in \mathbb{S}_f} \frac{J(S)}{\sigma(|S|)}, \quad f \in F, \tag{13}$$

$$\bar{\theta}_f = \frac{1}{|\bar{\mathbb{S}}_f|} \sum_{S \in \bar{\mathbb{S}}_f} \frac{J(S)}{\sigma(|S|)}, \quad f \in F \tag{14}$$

with the resulting ranking criterion

$$DAF_2(f) = (\theta_f - \bar{\theta}_f); \quad f \in F. \tag{15}$$

Note that, in view of the randomly generated sequence of sets, one should be careful in computation of $\sigma(d)$, when the number of sets $|\mathbb{S}(d)|$ is small.

Remark: The simplest form of the ranking criterion $DAF_0(f)$ is also well justifiable because there is actually no strong reason to norm the variance of the criterion function $J(\cdot)$ except as pre-causion against possibly misleading emphasis put on features that appear important but behave unstably.

IV. Feature Selection Procedure

The proposed FS procedure based on DAF_0 or DAF_1 ranking requires a criterion $J(\cdot)$, an optional upper limit τ of probe subsets' cardinality, and the choice of stopping condition. The straightforward stopping condition options are a) time, or b) number of probe subsets to be generated, or c) a requirement that each feature will be assesed based on at least σ different subset evaluations:

$$\min_{f \in F} \{|\mathbb{S}_f|, |\bar{\mathbb{S}}_f|\} \geq \sigma. \tag{16}$$

Then the actual feature selection procedure is the same for all DAF_i rank criteria, $i = 0, 1, 2$:

1) generate sequence $\mathbb{S}$ of random probe subsets, where for each subset $S \in \mathbb{S}$ the value $J(S)$ is immediately evaluated, until stopping condition is met
2) compute the value $DAF_i(f)$ for each feature $f \in F$
3) select required number of features with highest DAF_i values

Note that we do not address here the question of determining the correct final subset size, which in general is a difficult problem out of scope of this paper. We suggest to follow the same practices as with other FS methods that provide a ranking of features.

Remark: From the computational complexity point of view only Step 1 is to be concerned. Once the sequence $\mathbb{S}$ is generated and all subsets $S \in \mathbb{S}$ evaluated by criterion $J(\cdot)$, the computation of all DAF_i rank criteria, $i = 0, 1, 2$, is of negligible computational complexity.

TABLE I
Data sets used in experiments

Data	Features	Classes	Samples	Description
Reuters-21578	10105	33	8941	real text
Gisette	5000	2	1000	handwritten digits
Madelon	500	2	2000	artificial, non-linear
Spambase	57	2	4601	e-mails

V. Experimetal Evaluation

To evaluate the proposed method we conduct a series of experiments on four data sets (see Table I). In each experiment the data is randomly split to 50% training and 50% testing part with stratification, i.e., preserving relative class sizes. All experiments are of wrapper type to show that the method is usable in computationally demanding context – the FS criterion $J(\cdot)$ has been the accuracy of k-Nearest Neighbor classifier or Support Vector Machine (SVM) [45] estimated by means of 3-fold cross-validation on training data part only. The testing data part has been used only once for final verification of classification accuracy on the selected subspace. If not stated otherwise, all experiments have been repreated $20\times$ with different random train-test data splits. Thus, the solid lines in Figures 2, 3, 4 and 5 show the achieved mean classification accuracy for various final subset sizes over the trials while the gray areas denote the respective standard deviations. Note that SVM performance depends on parameters. To save time we optimized SVM parameters only once per trial on the training data part with all features before the actual FS took place. In this sense there remains space for further improvement of the presented results.

Remark: All experiments have been conducted using the Feature Selection Toolbox 3 C++ framework's[1] concurrent FS implementation on a multi-core Opteron server under Linux.

A. Data Sets

The Reuters-21578 data set is commonly used in text categorization benchmarking.[2] Our text preprocessing included removing all non-alphabetic characters, ignoring all the words that contained digits or non alpha-numeric characters, removing words from a stop-word list. We replaced each word by its morphological root, removed all the words which had less than three occurrences. After pre-processing the data set contained 33 classes of document-representing vectors of dimensionality 10105. The largest class contained 3924, the smallest only 19 non-zero documents.

The Gisette and Madelon datasets have been used in NIST 2003 feature selection challenge and are now available through UCI Repository [46]. Gisette data represent two hardly distinguishable handwritten digits '4' and '9'. Our experiments are restricted to the subset of the original data intended originally for verification. From the 5000 features 2500 are known to have no predictive power.

Madelon is an artificial dataset known to contain 20 informative features and 480 features with no predictive power.

[1] http://fst.utia.cz

[2] http://www.daviddlewis.com/resources/testcollections/reuters21578

TABLE II
SUMMARY OF CONDUCTED DAF EXPERIMENTS

Data	Time limit	thr-eads	τ	$J(\cdot)$ evals.	average $\lvert\mathbb{S}_f\rvert$	average $\lvert\bar{\mathbb{S}}_f\rvert$
Reuters	200min	24	200	∼148000	∼1500	∼146500
Gisette	200min	16	200	∼46400	∼950	∼45450
Madelon	180min	16	150	∼25000	∼3600	∼21400
Spambase	200min	8	57	∼160000	∼79500	∼80500

The data is defined so that the class separation hyperplane is multivariate and highly non-linear.

The Spambase dataset, available through UCI Repository [46], is only 57-dimensional and as such is used here to enable comparison with Sequential Forward Floating Selection (SFFS) [47] procedure which is known to be capable of yielding near-optimal or optimal results. It is, nevertheless, not applicable with higher-dimensional problems due to computational complexity.

B. Results

Figures 2a, 3a, 4a and 5a show the achieved classification accuracy on independent test data using the same classifier that had been used in feature selection. Figures 2b, 3b, 4b and 5b test the same feature subsets using a different classifier. (In Figure 2 the SVM has been used with linear kernel, in all other cases with radial basis function kernel [45].) Figures 2c-e, 3c-e, 4c-e and 5c-e show the stability properties of considered FS procedures evaluated by various stability measures [48]. All stability measures yield values from $[0, 1]$, higher stability values are desirable indicate FS results less dependent on particular data sampling.

If not stated otherwise, the stopping condition in each DAF based experiment (each single trial of the 20 conducted) has been set in form of time limit to 200 minutes run time. This is very strict limitation especially in Reuters and Gisette cases in view of the high dimensionality involved, but illustrates well the suitability of DAF for such high-dimensional setting. In cases of Reuters and Gisette the cardinality of probe subsets generated in the course of DAF feature selection has been limited by $\tau = 200$, in case of Madelon by $\tau = 150$, in case of Spambase there was no limit (i.e., $\tau = N = 57$). See Table II for summary on the conducted search processes (per one trial).

BIF time complexity has proved to be negligible. In the slowest case of Reuters data it needed less than 5 minutes per trial. In contrary, SFFS slows down rapidly with increasing dimensionality. It needed 16 minutes per trial with 57-dimensional Spambase data, but could not finish a single trial in 5 days with 500-dimensional Madelon data.

In all experiments presented in Figures 2, 3, 4 and 5 the DAF based feature selection has performed considerably better than individual feature ranking (BIF). Let us point out the particular case of highly non-linear Madelon data, known to contain exactly 20 informative features. In this case BIF failed completely while the DAF methods succeeded in identifying the right features (see Figure 4).

In case of Spambase data (Figure 5) we compare DAF and BIF with SFFS that represents here a stronger but principally slower optimizer. Figure 5a illustrates well the limits of DAF ranking when compared to the strong optimizer SFFS, while in contrary Figure 5b illustrates its advantage – DAF generalizes better than SFFS in that its results are more usable in context of a principally different classifier.

The reported stability of DAF methods (Figures 2c-e, 3c-e, 4c-e and 5c-e) does not seem to be significantly different from that of BIF (BIF is considered one of the most stable FS methods [30]). In case of Madelon data DAF stability is better due to complete BIF failure. Note the considerably worse stability of SFFS in Figure 5c-e, confirming that stronger optimization ability is here outweighed by strong over-fitting.

Remark: further experiments suggest that increasing DAF search time from 200 to 400 minutes in case of Reuters data improves classification accuracy by $\sim 1\%$ when compared to results presented in Figure 2.

C. Discussion

To investigate how the choice of stopping condition and the optional τ probe subset cardinality limit affect DAF feature selection performance we have conducted two supplemental experiments on the 500-dimensional Madelon data.

Figure 6 illustrates the impact of changing DAF time limit on the resulting feature subset performance. It can be seen that roughly 2 hours suffice to make the most of the method (compare to SFFS complexity in this case).

Figure 7 illustrates the impact of changing τ cardinality limit. The importance of parameter τ appears less crucial, except that too low values make it impossible to reveal multi-feature dependencies to sufficient extent, while in contrary too high values may make it impossible to distinguish feature specific behavior in too wide a context. Specifically in case of Reuters data we have observed that setting $\tau \approx 10^3 \div 10^4$ led to poor DAF generalization performance. This is expectable because attempting to reaveal dependencies among too many features in case of insufficient sample size with respect to dimensionality may easily lead to over-fitting.

Remark: The factor of importance in DAF based FS is actually the achieved sizes of $\mathbb{S}_f$ and $\bar{\mathbb{S}}_f$ for all $f \in F$ (see Section III-A). Each feature $f \in F$ is needed to be evaluated in large enough number of contexts (random probe subsets that do and do not contain it), otherwise its general behavior can not be meaningfully estimated. Note that probe subsets $S \in \mathbb{S}$ of lower cardinality can provide information about lower number of features, yet this information is likely to be more reliably estimated than in case of larger S. With computationally complex $J(\cdot)$ criteria the setting of τ affects the sizes of $\mathbb{S}_f$ and $\bar{\mathbb{S}}_f$ also indirectly. Given a constant time limit, for lower τ values the criterion is evaluated on smaller subsets faster and, thus, more times, while for higher τ the opposite is true.

VI. CONCLUSION

A novel feature selection method denoted "dependency-aware feature ranking" has been introduced. The rank is

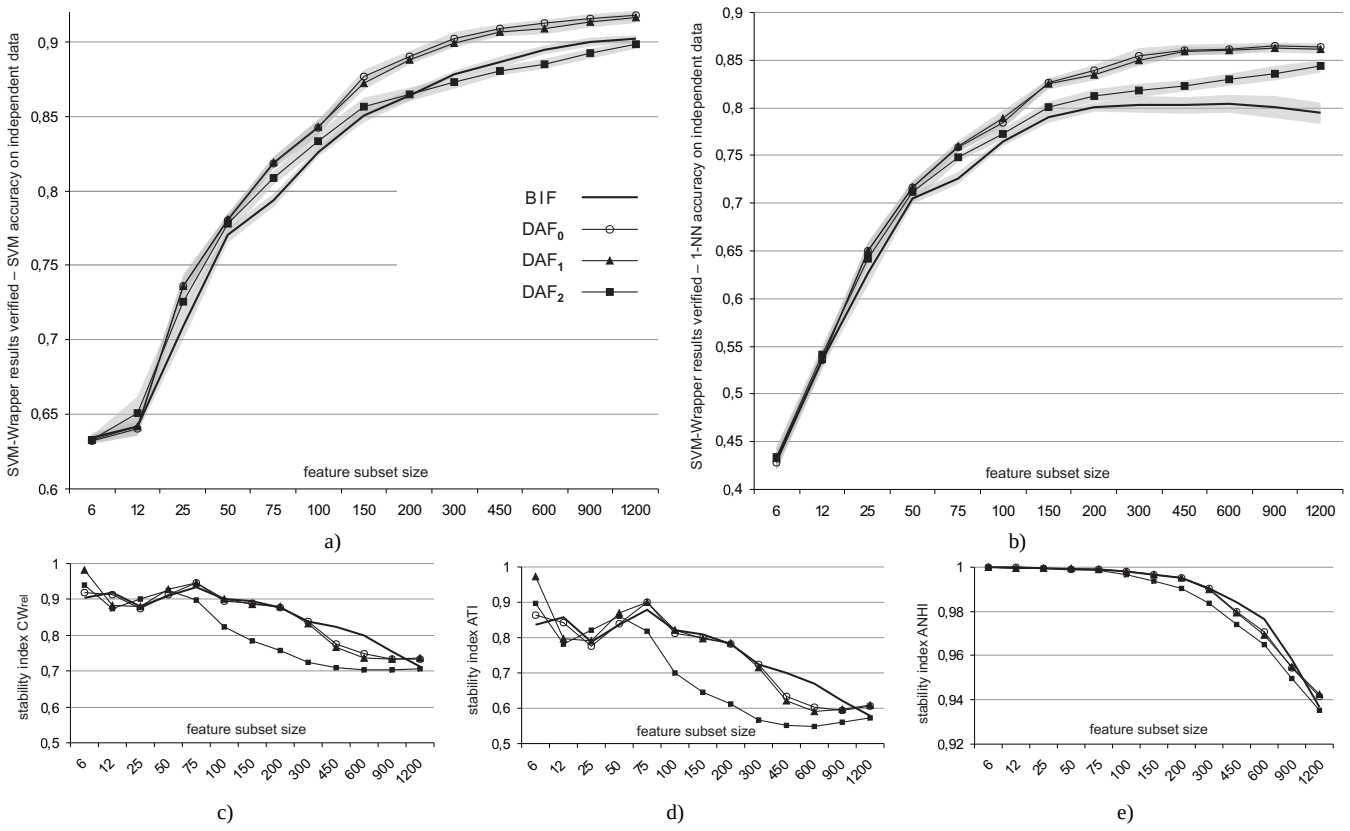

Fig. 2. 10105-dimensional Reuters data – SVM(lin)-Wrapper feature selection results over 20 trials with different random train-test data splits. Classification accuracy on independent data using a) SVM(lin), b) 1-NN. Stability evaluated using c) CW_{rel}, d) ATI, e) $ANHI$ measures. (Search time 200 min.)

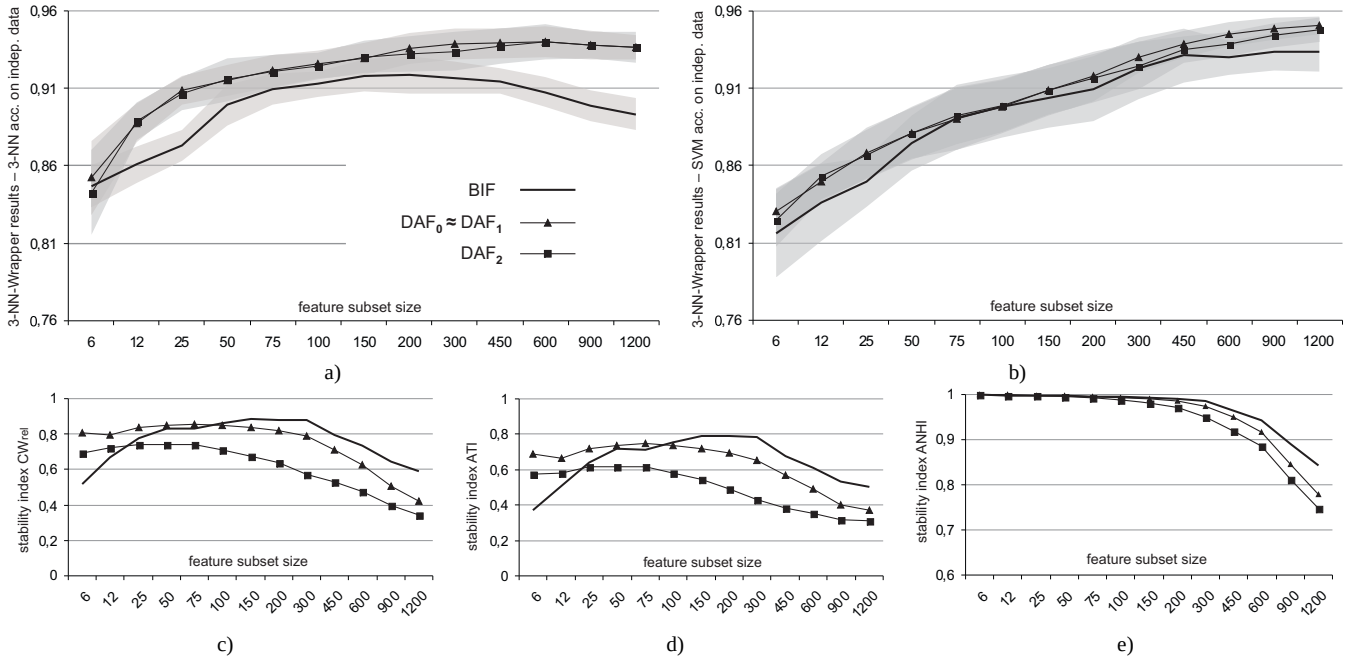

Fig. 3. 5000-dimensional Gisette Data – 3-NN-Wrapper feature selection results over 20 trials with different random train-test data splits. Classification accuracy on independent data using a) 3-NN, b) SVM(rbf). Stability evaluated using c) CW_{rel}, d) ATI, e) $ANHI$ measures. (Search time 200 min.)

computed as the average benefit of including a feature into a number of randomly generated feature subsets. The benefit is expressed as the difference of mean criterion values computed for subsets that do and do not contain the feature, based on arbitrarily chosen feature selection criterion. This simple idea has been shown well suitable for high and very-high-

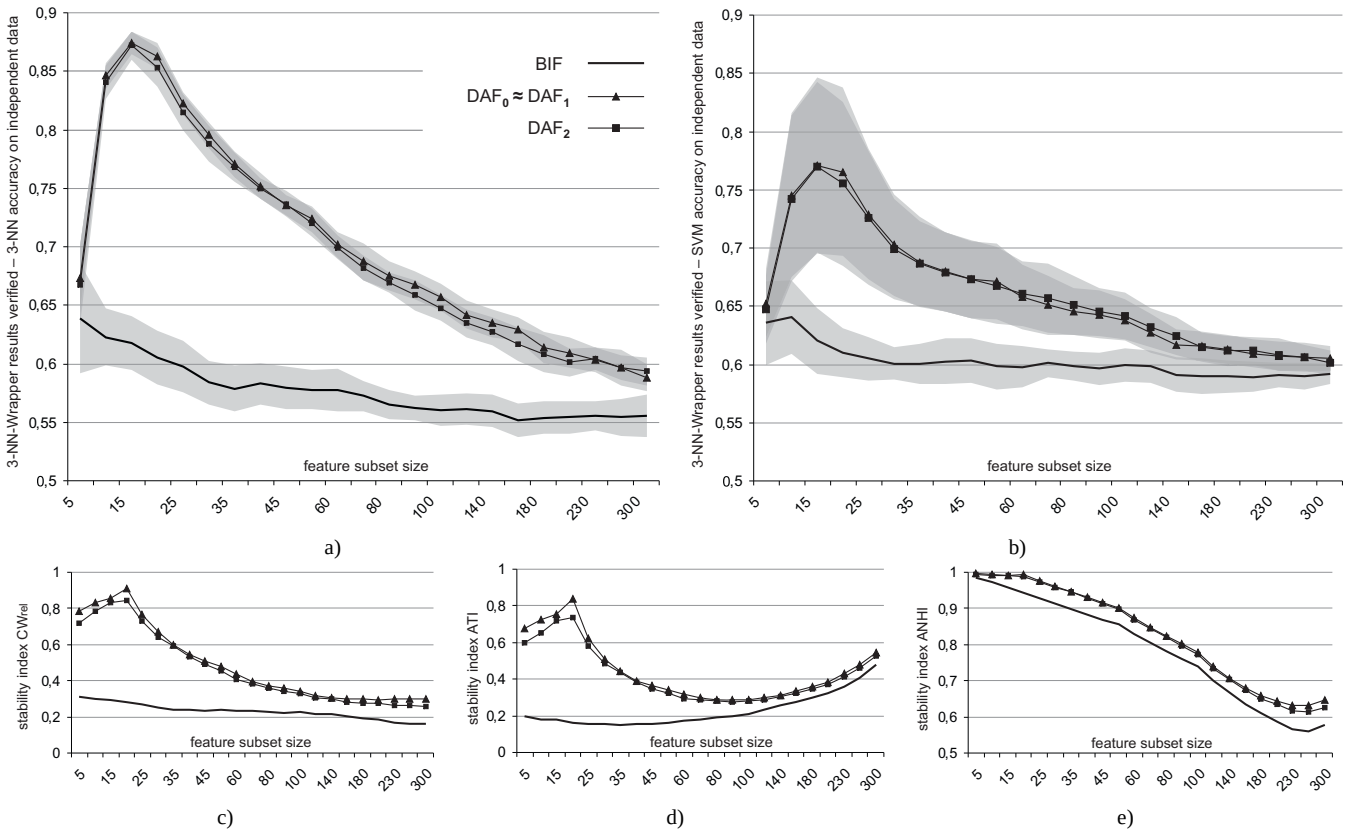

Fig. 4. 500-dimensional Madelon Data – 3-NN-Wrapper feature selection results over 20 trials with different random train-test data splits. Classification accuracy on independent data using a) 3-NN, b) SVM(rbf). Stability evaluated using c) CW_{rel}, d) ATI, e) $ANHI$ measures. (Search time 180 min.)

dimensional feature selection problems where it is capable of considerably over-performing the commonly used individual feature ranking approaches due to its favorable mix of properties: the ability to reveal contextual information, reasonable speed, and generalization ability.

ACKNOWLEDGMENT

The work has been supported by grants of the Czech Ministry of Education 2C06019 ZIMOLEZ and 1M0572 DAR.

REFERENCES

[1] I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, Eds., *Feature Extraction – Foundations and Applications*, ser. Studies in Fuzziness and Soft Computing. Physica-Verlag, Springer, 2006, vol. 207.

[2] F. J. Ferri, P. Pudil, M. Hatef, and J. Kittler, "Comparative study of techniques for large-scale feature selection," *Machine Intelligence and Pattern Recognition*, vol. 16, 1994.

[3] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1-2, pp. 273–324, 1997.

[4] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Norwell, MA, USA: Kluwer Academic Publishers, 1998.

[5] "Special issue on variable and feature selection," *J. of Machine Learning Research. http://www. jmlr.org/papers/special/feature.html*, 2003.

[6] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 4, pp. 491–502, 2005.

[7] Y. Sun, "Iterative relief for feature weighting: Algorithms, theories, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1035–1051, 2007.

[8] P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. Englewood Cliffs, London, UK: Prentice Hall, 1982.

[9] A. Salappa, M. Doumpos, and C. Zopounidis, "Feature selection algorithms in classification problems: An experimental evaluation," *Optimization Methods and Software*, vol. 22, no. 1, pp. 199–212, 2007.

[10] S. B. Serpico and L. Bruzzone, "A new search algorithm for feature selection in hyper-spectral remote sensing images," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 39, no. 7, pp. 1360–1367, 2001.

[11] S. Yu, "Feature selection and classifier ensembles: A study on hyper-spectral remote sensing data," Ph.D. dissertation, University of Antwerp, Antwerp, Netherlands, 2003.

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

[13] J. G. Dy and C. E. Brodley, "Feature selection for unsupervised learning," *J. Mach. Learn. Res.*, vol. 5, pp. 845–889, 2004.

[14] E. P. Xing, *Feature Selection in Microarray Analysis*. Springer, 2003, pp. 110–129.

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

[16] M. A. Tahir, A. Bouridane, F. Kurugollu, and A. Amira, "Feature selection using tabu search for improving the classification rate of prostate needle biopsies," in *Proc. ICPR '04*, vol. 2. Washington, DC, USA: IEEE Computer Society, 2004, pp. 335–338.

[17] G. Brown, "A New Perspective for Information Theoretic Feature Selection," in *Proc. AISTATS '09, vol. 5 of JMLR: W&CP 5*, 2009, pp. 49–56.

[18] P. Somol, P. Pudil, and J. Kittler, "Fast branch & bound algorithms for optimal feature selection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 7, pp. 900–912, 2004.

[19] J. Fan and R. Li, "Statistical challenges with high dimensionality: Feature selection in knowledge discovery," 2006. [Online]. Available: http://www.citebase.org/abstract?id=oai:arXiv.org:math/0602133

[20] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," in *ICML '97: Proc. 14th Int. Conf. on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997, pp. 412–420.

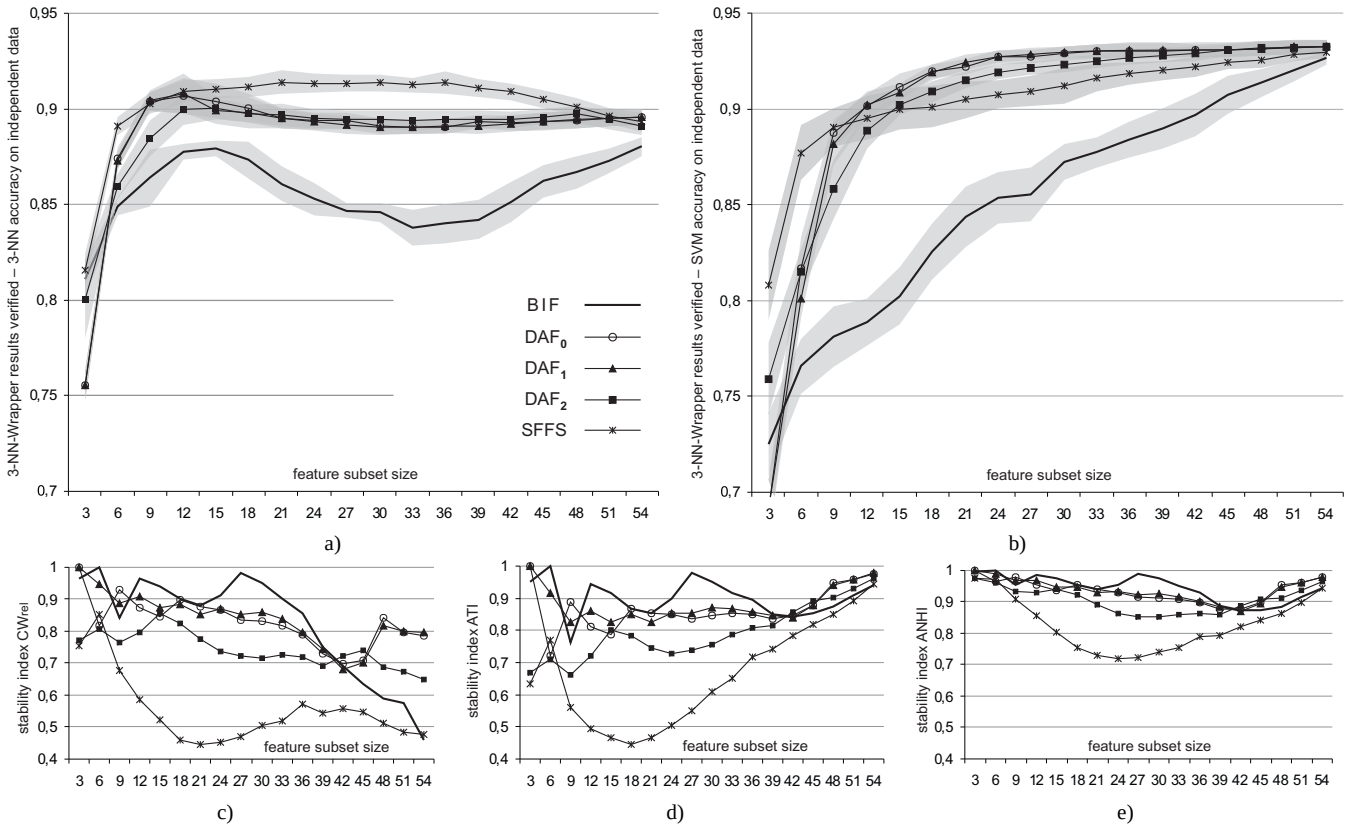

Fig. 5. 57-dimensional UCI Spambase Data – 3-NN-Wrapper feature selection results over 20 trials with different random train-test data splits. Classification accuracy on independent data using a) 3-NN, b) SVM(rbf). Stability evaluated using c) CW_{rel}, d) ATI, e) $ANHI$ measures.

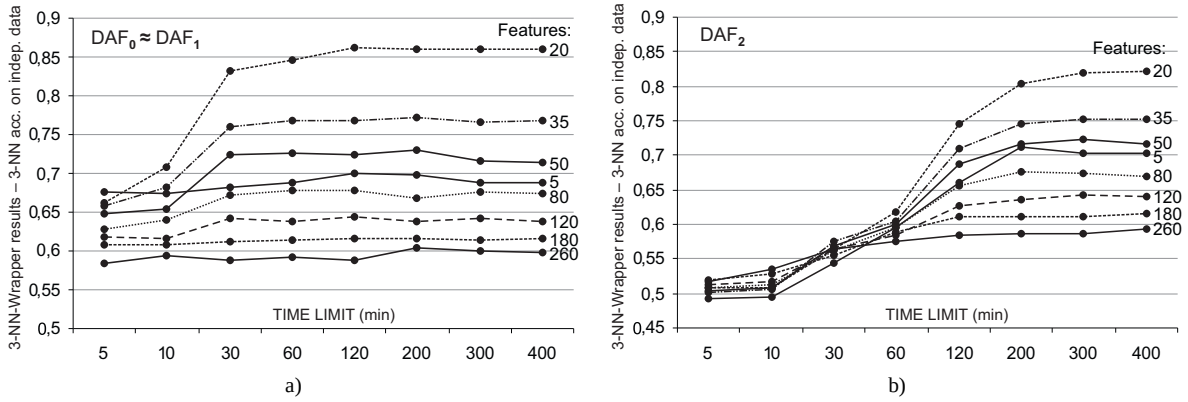

Fig. 6. Illustrating the impact of time limit in DAF feature selection on the resulting classification performance – 500-dimensional Madelon Data, cardinality limit $\tau = 150$.

[21] D. Mladenić and M. Grobelnik, "Feature selection for unbalanced class distribution and naive bayes," in *ICML '99: Proc. 16th Int. Conf. on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, pp. 258–267.

[22] T. Liu, S. Liu, Z. Chen, and W.-Y. Ma, "An evaluation on feature selection for text clustering," in *Proc. 20th Int. Conf. on Machine Learning (ICML-2003)*, Washington, DC, USA, 2003, pp. 488–495.

[23] N. Vasconcelos, "Feature selection by maximum marginal diversity: optimality and implications for visual recognition," in *Proc. IEEE Conf. On Computer Vision And Pattern Recognition*, vol. 1. Washington, DC, USA: IEEE Computer Society, June 2003, pp. 762–769.

[24] M. E. Farmer, S. Bapna, and A. K. Jain, "Large scale feature selection using modified random mutation hill climbing," in *ICPR '04: Proc. 17th Int. Conf. on Pattern Recognition*, vol. 2. Washington, DC, USA: IEEE Computer Society, 2004, pp. 287–290.

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

[26] ——, "A pitfall in determining the optimal feature subset size," in *Proc. 4th Int. Workshop on Pattern Recognition in Information Systems (PRIS 2004)*, Porto, Portugal, 2004, pp. 176–185.

[27] ——, "Less biased measurement of feature selection benefits," in *Statistical and Optimization Perspectives Workshop, SLSFS 2005, Bohinj, Slovenia, February 23-25, 2005, Revised Selected Papers*, C. Saunders, M. Grobelnik, S. Gunn, and J. Shawe-Taylor, Eds., vol. LNCS 3940. Springer Berlin / Heidelberg, 2006, pp. 198–208.

[28] Š. J. Raudys, R. Baumgartner, and R. Somorjai, "On understanding and assessing feature selection bias," in *Artificial Intelligence in Medicine*, vol. LNAI 3581. Springer-Verlag, 2005, pp. 468–472.

[29] Š. J. Raudys, "Feature over-selection," in *Structural, Syntactic, and Statistical Pattern Recognition*, vol. LNCS 4109. Berlin / Heidelberg,

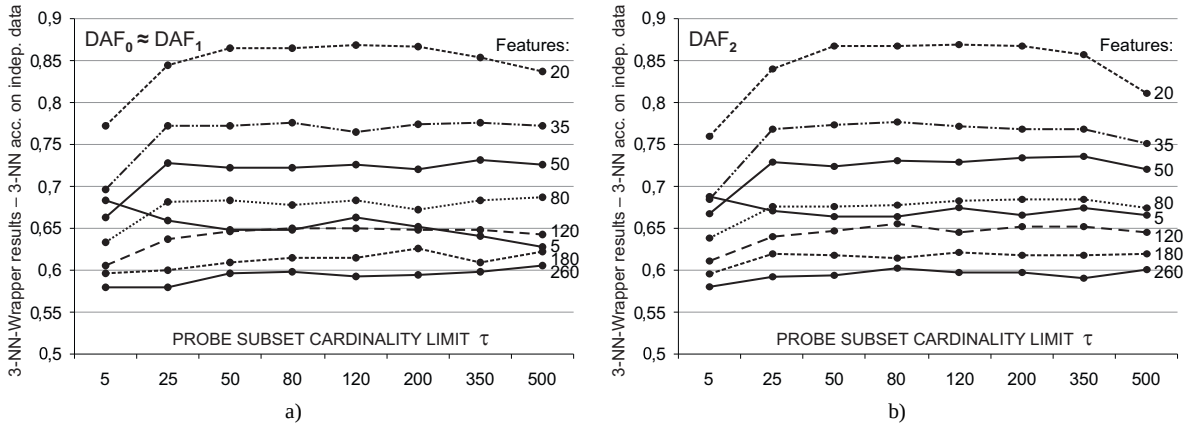

Fig. 7. Illustrating the impact of random probe subsets' cardinality upper limit τ in DAF feature selection on the resulting classification performance – 500-dimensional Madelon Data, search time limit set to 300 min.

Germany: Springer-Verlag, 2006, pp. 622–631.

[30] L. I. Kuncheva, "A stability index for feature selection," in *Proc. 25th IASTED International Multi-Conference AIAP'07*. Anaheim, CA, USA: ACTA Press, 2007, pp. 390–395.

[31] J. T. Souza, S. Matwin, and N. Japkowicz, "Parallelizing feature selection," *Algorithmica*, vol. 45, no. 3, pp. 433–456, 2006.

[32] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *ICML-03: Proc. 20th Int. Conf. on Machine Learning*, vol. 20. Washington DC, USA: Morgan Kaufmann Publishers Inc., 2003, pp. 856–863.

[33] F. Fleuret, "Fast binary feature selection with conditional mutual information," *J. Mach. Learn. Res.*, vol. 5, pp. 1531–1555, 2004.

[34] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, 2005.

[35] K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algoirthm," in *AAAI-92: Proc. 10th Nat. Conf. on Artifical Intelligence*, W. Swartout, Ed. AAAI Press/The MIT Press, August 1992, pp. 129–134.

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

[37] J. Yang and Y.-P. Li, "Orthogonal relief algorithm for feature selection," in *Intelligent Computing*, vol. LNCS 4113. Berlin / Heidelberg, Germany: Springer-Verlag, 2006, pp. 227–234.

[38] F. Hussein, R. Ward, and N. Kharma, "Genetic algorithms for feat. select. and weighting, a review and study," *icdar*, vol. 00, p. 1240, 2001.

[39] F. W. Glover and G. A. Kochenberger, Eds., *Handbook of Metaheuristics*, ser. Int. Series in Operations Research & Management Science. Springer, 2003, vol. 57.

[40] M. A. Tahir, A. Bouridane, and F. Kurugollu, "Simultaneous feature selection and feature weighting using hybrid tabu search/k-nearest neighbor classifier," *Pattern Recogn. Lett.*, vol. 28, no. 4, pp. 438–446, 2007.

[41] J. C. W. Debuse and V. J. Rayward-Smith, "Feature subset selection within a simulated annealing data mining algorithm," *J. Intell. Inf. Syst.*, vol. 9, no. 1, pp. 57–81, 1997.

[42] J. Novovičová, P. Somol, and P. Pudil, "Oscillating feature subset search algorithm for text categorization," in *Structural, Syntactic, and Statistical Pattern Recognition*, vol. LNCS 4109. Berlin / Heidelberg, Germany: Springer-Verlag, 2006, pp. 578–587.

[43] I. A. Gheyas and L. S. Smith, "Feature subset selection in large dimensionality domains," *Pattern Recognition*, vol. 43, no. 1, pp. 5 – 13, 2010.

[44] C. Lai, M. J. T. Reinders, and L. Wessels, "Random subspace method for multivariate feature selection," *Pattern Recogn. Lett.*, vol. 27, no. 10, pp. 1067–1076, 2006.

[45] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for SVM*, 2001, http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[46] A. Asuncion and D. Newman, "UCI machine learning repository, http://www.ics.uci.edu/ ∼mlearn/ mlrepository.html," 2007. [Online]. Available: http://www.ics.uci.edu/∼mlearn/MLRepository.html

[47] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Pattern Rec. Lett.*, vol. 15, no. 11, pp. 1119–1125, 1994.

[48] P. Somol and J. Novovičová, "Evaluating stability and comparing output of feature selectors that optimize feature subset cardinality," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 11, pp. 1921–1939, 2010.