On Stopping Rules in Dependency-Aware Feature Ranking

Petr Somol[1,2], Jiří Grim[2], Jiří Filip[2], and Pavel Pudil[1]

[1] Faculty of Management, Prague University of Economics, Czech Republic
[2] Institute of Information Theory and Automation of the AS CR, Czech Republic

Abstract. Feature Selection in very-high-dimensional or small sample problems is particularly prone to computational and robustness complications. It is common to resort to feature ranking approaches only or to randomization techniques. A recent novel approach to the randomization idea in form of Dependency-Aware Feature Ranking (DAF) has shown great potential in tackling these problems well. Its original definition, however, leaves several technical questions open. In this paper we address one of these questions: how to define stopping rules of the randomized computation that stands at the core of the DAF method. We define stopping rules that are easier to interpret and show that the number of randomly generated probes does not need to be extensive.

Keywords: dimensionality reduction, feature selection, randomization, stopping rule

1 Introduction

Feature selection (FS) is one of dimensionality reduction techniques, that preserves meaning of the selected original data features, while irrelevant features are discarded. 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. The first step in solving the FS problem involves choosing appropriate method based on the knowledge (or lack of therein) of available training data properties. The key decision to be made involves the choice of the criterion and the search algorithm capable of optimizing such a criterion. Note that feature subset search is potentially an expensive combinatorial problem as the number of candidate subsets is very high. The search is stopped according to chosen stopping rule; it can be defined in terms of achieved completeness of search, criterion convergence threshold, subset size limit, time, etc.

In recent years the focus of feature selection research is moving from the relatively well covered area of low-to-mid-dimensional recognition problems towards very-high-dimensional problems [1]. As the high-dimensional FS is susceptible

to problems arising from insufficient sample size and computational complexity, the FS methods often prefer simpler analysis ignoring inter-feature dependencies, e.g., based on feature ranking [2]. This simplifications is commonly assumed less harmful than obtaining misleading information through serious estimation errors due to over-fitting. The computational complexity can be reduced by resorting to randomized methods, however, this is counterbalanced by loss of optimality due to a user-defined time restriction of the search process. An example of such techniques is Relief algorithm [3] based on a simple idea of repeated randomized sampling of one pattern followed by feature weights update. Combinations of randomized and greedy algorithms [4] seems to be better suited for high-dimensional tasks, than randomized methods based on Genetic algorithms, Simulated Annealing, and Tabu Search [5], which provide strong optimization mechanism, at the cost of long converge times. Method's over-fitting has been tackled by a random restriction of inter-feature dependencies evaluation by repeatable running FS process on various random subspaces in [6].

Finally a combination of ranking and randomization called Dependency-Aware Feature Ranking has been introduced in [7]. The idea of individually best ranking is generalized to evaluate features contributions in a sequence of randomly generated feature subsets. The method has been shown capable of selecting features reliably even in settings where standard feature techniques fail due to problem complexity or over-fitting issues and where individual feature ranking results are unsatisfactory. Several open questions, however, remain with respect to DAF applicability, that have not been addressed in [7]. The two most practically important are: a) *What is the right final subset size?*, and b) *How long is it necessary to let the random probe generation process run?*

The problem to specify the optimal number of features to be selected, is closely related to the number of available data, dimension of the feature space and also to the underlying classification complexity. It is well known that in case of infinitely large training sets we should use all features since by omitting features the classifier performance cannot be improved. If a multidimensional training set were not large enough then most classifiers would tend to over-fit with the resulting poor classification performance on the independent test data. In such a case the generalizing property of the classifier could be improved by selecting a subset of informative features. Obviously, the optimal choice of the final reduced dimensionality depends on the size of the training data set and the complexity of the underlying classification problem. In this sense the question a) is beyond the scope of this paper since the size of the training data set is not considered explicitly. For a more detailed discussion of dimensionality problems in the context of standard individual feature ranking see e.g. [8]. In the following we investigate some aspects of question b), i.e., we discuss different options specifying the stopping rule of the feature ordering process.

2 Dependency-Aware Feature Ranking

Denoting F the set of all features $F = \{f_1, f_2, \ldots, f_N\}$ 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. We assume the criterion $J(\cdot)$ to be bounded according to the most feature selection criteria (estimates of classification accuracy are typically bounded by [0,1]).

The starting point of dependency-aware feature ranking is a randomly generated sequence of feature subsets to be denoted *probe* subsets $\mathbb{S} = \{S_1, S_2, \ldots, S_K\}$, $S_j \subset F,\ \ j = 1, 2, \ldots, K$, where each subset is evaluated by the criterion function $J(\cdot)$. For details on probe generation see [7].

Given a sufficiently large sequence of feature subsets $\mathbb{S}$, we can utilize the information contained in the criterion values $J(S_1), J(S_2), \ldots, J(S_K)$ to assess how each feature adds to the criterion value. Therefore, we compare the quality of probe subsets containing f with the quality of probe subsets not including f.

We compute the mean quality μ_f of subsets $S \in \mathbb{S}$ containing the considered feature

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

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

$$\bar{\mu}_f = \frac{1}{|\bar{\mathbb{S}}_f|} \sum_{S \in \bar{\mathbb{S}}_f} J(S), \quad \bar{\mathbb{S}}_f = \{S \in \mathbb{S} : f \notin S\} \tag{2}$$

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

$$DAF(f) = \mu_f - \bar{\mu}_f, \quad f \in F. \tag{3}$$

The sequence of generated probe subsets can be arbitrarily long but the number of possible probes is finite. The probe subsets are generated randomly according to some fixed rules, for example the number of features in the subset may be fixed or bounded. If we denote $\mathbb{A}$ the class of admissible subsets which may occur in the sequence then, in view of the random generating procedure, the admissible subsets $S \in \mathbb{A}$ will occur in the sequence $\mathbb{S}$ repeatedly according to some fixed probabilities $\alpha(S)$. Thus, in long sequences of probes the admissible subsets $S \in \mathbb{A}$ will occur in $\mathbb{S}$ with the relative frequencies approaching $\alpha(S)$.

Like Eq. (1), (2) we denote $\mathbb{A}_f$ the class of admissible sets containing feature $f \in F$ and $\bar{\mathbb{A}}_f$ the class of admissible sets not containing feature f

$$\mathbb{A}_f = \{S \in \mathbb{A} : f \in S\}, \quad \bar{\mathbb{A}}_f = \{S \in \mathbb{A} : f \notin S\}, \quad f \in F. \tag{4}$$

It can be seen that, in view of above considerations, both the mean quality μ_f and $\bar{\mu}_f$ converge to some finite limit values. Considering Eq. (5) we can write

$$\lim_{|\mathbb{S}_f| \to \infty} \mu_f = \lim_{|\mathbb{S}_f| \to \infty} \frac{1}{|\mathbb{S}_f|} \sum_{S \in \mathbb{S}_f} J(S) = \sum_{S \in \mathbb{A}_f} \alpha_f(S) J(S) = \mu^* \tag{5}$$

where $\alpha_f(S)$ is the probability that the admissible subsets $S \in \mathbb{A}_f$ occur in the sequence $\mathbb{S}_f$ and μ^* is the corresponding limit value of μ_f. Similarly we can write analogous limit expression for the mean quality $\bar{\mu}_f$:

$$\lim_{|\bar{\mathbb{S}}_f| \to \infty} \bar{\mu}_f = \lim_{|\bar{\mathbb{S}}_f| \to \infty} \frac{1}{|\bar{\mathbb{S}}_f|} \sum_{S \in \bar{\mathbb{S}}_f} J(S) = \sum_{S \in \bar{\mathbb{A}}_f} \bar{\alpha}_f(S) J(S) = \bar{\mu}^* \tag{6}$$

with $\bar{\alpha}_f(S)$ denoting the occurrence probability of $S \in \bar{\mathbb{A}}_f$ in the sequence $\bar{\mathbb{S}}_f$. Consequently, the criterion value $DAF(f)$ has a finite limit for any $f \in F$:

$$\lim_{|\mathbb{S}_f| \to \infty} DAF(f) = \mu_f^* - \bar{\mu}_f^*, \quad f \in F. \tag{7}$$

It has been shown in [7] that selecting features according to highest DAF coefficients leads to significantly better results then selecting features according to individually best criterion values. This makes the method well suitable for scenarios where individual feature evaluation had been considered the only viable choice (i.e., very high-dimensional or small sample size problems).

In paper [7] the question of when to stop the process of randomized probe generation (i.e., what is the right value of K) is not specifically addressed. All presented results have been obtained using the ad-hoc stopping rules. The first obvious rule is *a user-specified time limit*, i.e., the computation is stopped after a pre-specified time limit. Here it is hoped that the number of probes that are evaluated in the time limit is sufficient with respect to the given problem. There is almost no way of guessing what time limit should suffice, except the generally applicable advice that the more time can be invested, the more accurate predictions can be made. Another problem here is the dependence on particular hardware, different computers would manage significantly different number of probes within the same time. The second trivial rule is *a user-specified limit of the number of probes*, i.e, the computation is stopped after a pre-specified number of probes has been investigated. Specifying the minimum necessary number of probes is as unreliable as specifying the time limit. Although this is independent on particular computer settings, there is still no guidance or interpretation available that would help to adjust the setting for particular problem.

3 Design of Novel Stopping Rules

In this section we consider two natural stopping rules that have not been considered in paper [7]. Both of them are based on evaluating a *function of change* while adding probes, which then can be thresholded to find the moment to stop.

Stopping Condition 1. *Change of Feature Order.* The adding of probes and recalculating DAF coefficients for each feature leads to changes in ordering of all features according to their DAF coefficients. Defining a threshold on the change would allow to stop adding probes when the ordering is not changing substantially any more.

Definition 1. *Let C denote the function to evaluate difference in feature ordering yielded by evaluating DAF coefficient in systems $\mathbb{S}_1$ and $\mathbb{S}_2$ where $\mathbb{S}_1 \subset \mathbb{S}_2$. Denoting $DAF(f)^{\mathbb{S}}$ the DAF coefficient of feature f computed on system $\mathbb{S}$, and assuming that features have been ordered according to descending $DAF(f)^{\mathbb{S}}$ values and the index of feature f in such ordering is denoted $DAF(f)^{\mathbb{S}}_{idx}$, we define* $C[\mathbb{S}_1, \mathbb{S}_2] = \frac{1}{N} \sum_{f=1}^{N} |DAF(f)^{\mathbb{S}_1}_{idx} - DAF(f)^{\mathbb{S}_2}_{idx}|$.

In Definition 1 we average the change in position in DAF-based ordering of features when a certain number of probes has been added to system $\mathbb{S}_1$ to obtain system $\mathbb{S}_2$. Naturally, with decreasing change in DAF based ordering of features we could assume at some point that no more probe adding is needed as it would not affect the resulting feature ranking.

In Stopping Condition 1 we assume the value of C asymptotically decreases with increasing $\mathbb{S}$ size. However, this may not be always true.

Proposition 1. *Assume we keep adding equally large groups of random probes to systems of subsets so as to obtain a series of systems* $\mathbb{S}_1 \subset \mathbb{S}_2 \subset \mathbb{S}_3 \ldots$. *For any* $i \geq 1$ *the value* $C[\mathbb{S}_i, \mathbb{S}_{i+1}]$ *can be arbitrary, there is no guarantee of it going close to zero. As a consequence, there is no guarantee that* C *would fall below given threshold when adding probes to a system of subsets indefinitely.*

Proof. The problem here is the fact that arbitrarily small change of DAF value can cause feature ordering to change. Imagine all features in the given problem to be equal. The feature selection criterion used to evaluate each probe would yield slightly different values for different probes because the estimate is done from finite training data susceptible to sampling errors. The process of computing DAF values would produce for each feature a DAF coefficient that would be arbitrarily close to each other, in some cases equal. Adding a probe could at any time cause an arbitrarily small change (possibly decreasing with the number of probes), but any arbitrarily small nonzero change would be capable of change DAF coefficient values of two features and change their mutual order.

It seems Stopping Condition 1 is thus useless in general case. We will test it, however, in our experiments as well, as the convergence problem should not show up in cases when a sufficient distinction among features can be identified.

Stopping Condition 2. *Change of Average DAF value.* The adding of probes and recalculating DAF coefficients for each feature leads to changes in DAF coefficient value for some or all features. Assuming that these changes would decrease with increasing number of probes, it should be possible to define a threshold on DAF value change to specify when the change is to be considered small enough to justify stopping the process.

Definition 2. *Let* $C2$ *denote the function to evaluate difference in average DAF coefficient values over all features, yielded by evaluating DAF coefficient in systems* $\mathbb{S}_1$ *and* $\mathbb{S}_2$ *where* $\mathbb{S}_1 \subset \mathbb{S}_2$. *Denoting* $DAF(f)^{\mathbb{S}}$ *the DAF coefficient of feature* f *computed on system* $\mathbb{S}$, *we define* $C2[\mathbb{S}_1, \mathbb{S}_2] = \frac{1}{N}\sum_{f=1}^{N} |DAF(f)^{\mathbb{S}_1} - DAF(f)^{\mathbb{S}_2}|$.

In Definition 2 we average the change in DAF coefficient values of features when a certain number of probes has been added to system $\mathbb{S}_1$ to obtain system $\mathbb{S}_2$. Naturally, with decreasing change in DAF coefficient values we could assume at some point that no more probe adding is needed as it would not affect the resulting feature ranking. Concerning the convergence properties of $C2$ we proof the following Lemma.

Lemma 1. *Assume we keep adding equally large groups of random probes to systems of subsets so as to obtain a series of systems* $\mathbb{S}_1 \subset \mathbb{S}_2 \subset \mathbb{S}_3 \ldots$. *Then, for arbitrarily small threshold value* $t > 0$ *there exists a size of subset system* $\mathbb{S}$ *(number of probes)* p *so that for any* $i > j > p$ *it is true that* $C2[\mathbb{S}_i, \mathbb{S}_j] < t$.

Proof. The proof is a simple consequence of the Bolzano-Cauchy theorem. The sequence of $DAF(f)^{\mathbb{S}}$ coefficients converges with the increasing number of probes in $\mathbb{S}$ and the same holds for the finite sum of coefficients $\sum_{f=1}^{N} DAF(f)^{\mathbb{S}}$. Therefore the corresponding Bolzano-Cauchy condition is satisfied which directly implies the assertion of the Lemma.

The remaining problem with Stopping Condition 2 is the necessity by user to specify a threshold based on DAF coefficient values. this may still be difficult to interpret. Therefore, we suggest to set relative instead of absolute threshold. The relative change can be evaluated with respect to the first recorded change in probe adding process. For this and also for computational reasons it is practical to evaluate function $C2$ not after each probe addition but after the addition of several probes.

Stopping Condition 2a. *Relative Change of Average DAF value.* The adding of probes to system of subsets $\mathbb{S}$ and recalculating DAF coefficients for each feature after the additions leads to changes in DAF coefficient value for some or all features. Stop probe adding when for the k-th added probe it is true that $\frac{C2[\mathbb{S}_k, \mathbb{S}_{k+1}]}{C2[\mathbb{S}_1, \mathbb{S}_2]} < t$ for a pre-specified threshold t.

In this case the threshold represents limit on the proportional change in average DAF coefficient values. In the next section we show on examples how the values C and $C2$ correspond with classification accuracy throughout the probe addition process.

4 Experimental Evaluation

We illustrate the proposed stopping rules on two datasets: Reuters-21578 text categorization benchmark data[3] (33 classes, 10105 features) and artificial Madelon data [9] (2 classes, 500 features, out of which 20 are informative and 480 noise). Our experiment setup followed the setup described in [7]. With Reuters data we used the estimated accuracy of linear SVM; both as probe evaluating criterion and the eventual evaluation of the quality of selected subsets. With Madelon data we used 3-NN for the same purpose.

Figures 1 and 2 show a 3D graph showing the achieved classification accuracy on independent test data at various stages of probe-adding process. As DAF ranking does not decide about the number of features, the d axis in graph represents results for various subset sizes obtained by using the first d best features according the current DAF coefficients. Both Figures 1 and 2 show very quick improvement of classification accuracy after a small number of initially added

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

probes, most of the remaining process of probe adding later led to very slow improvement (Fig. 1) or negligible improvements but stabilization (visible in Fig. 2 at least for subset sizes around 20 representing the informative features).

The experiments serve primarily to illustrate the behavior of functions C and $C2$ with respect to growing number of probes being added to $\mathbb{S}$. The C and $C2$ have not been computed after each single added probe but after each 400-th probe. This is to compensate for the fact that adding a single probe can not affect all features (probe size was limited to 200 features).

The function C converged very slowly in the case of Madelon data. Reaching a point of no changes in feature ordering proved unrealistic in this case of 500-dimensional data; with higher-dimensional Reuters data we did not even attempt. The function $C2$ though converges reasonably fast as can be seen in both experiments. The question of what would be the practical threshold can not be answered unanimously for the general case, but in all our experiments (on 5 different datasets from which only 2 are presented here) it showed practical to set the threshold roughly to $\frac{C2[\mathbb{S}_k,\mathbb{S}_{k+1}]}{C2[\mathbb{S}_1,\mathbb{S}_2]} < 0.01$, i.e., to stop when $C2$ values decrease roughly to 1% of their initial value.

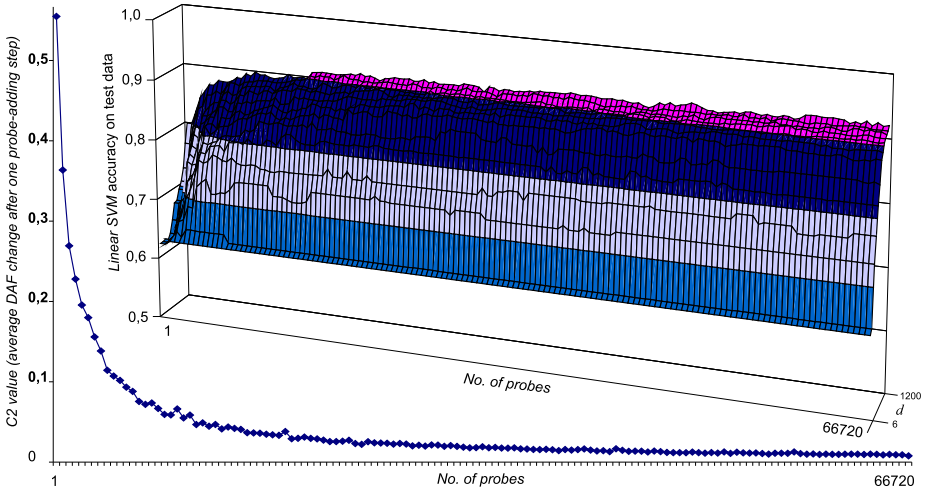

Fig. 1. Reuters data - SVM Classifier accuracy and $C2$ convergence during DAF probe generation.

5 Conclusions

We have investigated alternative stopping rules in Dependency-Aware Feature Ranking. We have shown that thresholding the averaged change in DAF value when adding probes to the considered subset system is preferable to other stopping rules in terms of interpretability, especially in cases when there is lack of knowledge of the underlying data. We have also demonstrated that DAF is fairly robust and does not require excessive numbers of randomized probes (as

8

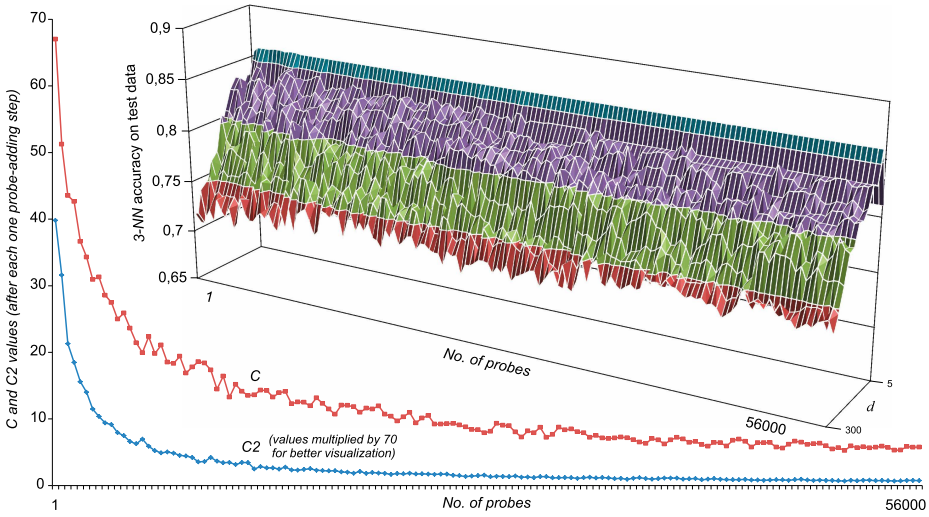

Fig. 2. Madelon data - 3-NN Classifier accuracy and C and $C2$ convergence during DAF probe generation.

expressed by change evaluating functions) in order to produce feature ranking that works well in independent test case.

Acknowledgements
This work has been supported by the Czech Science Foundation grants P403/12/1557 and P103/11/0335.

References

1. Fan, J., Li, R.: Statistical challenges with high dimensionality: Feature selection in knowledge discovery (2006)
2. Kuncheva, L.I.: A stability index for feature selection. In: Proc. 25th IASTED International Multi-Conference AIAP'07, ACTA Press (2007) 390–395
3. Kononenko, I.: 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) 171–182
4. Gheyas, I.A., Smith, L.S.: Feature subset selection in large dimensionality domains. Pattern Recognition **43**(1) (2010) 5–13
5. Glover, F.W., Kochenberger, G.A., eds.: Handbook of Metaheuristics. Volume 57 of Int. Series in Operations Research & Management Science. Springer (2003)
6. Lai, C., Reinders, M.J.T., Wessels, L.: Random subspace method for multivariate feature selection. Pattern Recogn. Lett. **27**(10) (2006) 1067–1076
7. Somol, P., Grim, J., Pudil, P.: Fast dependency-aware feature selection in very-high-dimensional pattern recognition. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC), IEEE (2011) 502–509
8. Liu, H., Motoda, H.: Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers, Norwell, MA, USA (1998)
9. Newman, D., Hettich, S., Blake, C., Merz, C.: UCI repository of machine learning databases (1998)