The Problem of Fragile Feature Subset Preference in Feature Selection Methods and A Proposal of Algorithmic Workaround

Somol P., Grim J.
Dept. of Pattern Recognition
Inst. of Information Theory and Automation, CAS
Pod vodárenskou věží 4, CZ 182 08 Prague 8
Email: {somol,grim}@utia.cas.cz

Pudil P.
Faculty of Management
Prague University of Economics
Jarošovská 1117/II, CZ 377 01, Jindřichův Hradec
Email: pudil@fm.vse.cz

***Abstract*—We point out a problem inherent in the optimization scheme of many popular feature selection methods. It follows from the implicit assumption that higher feature selection criterion value always indicates more preferable subset even if the value difference is marginal. This assumption ignores the reliability issues of particular feature preferences, over-fitting and feature acquisition cost. We propose an algorithmic extension applicable to many standard feature selection methods allowing better control over feature subset preference. We show experimentally that the proposed mechanism is capable of reducing the size of selected subsets as well as improving classifier generalization.**

***Keywords*-feature selection, machine learning, over-fitting, classification, feature weights, weighted features, feature acquisition cost**

I. Introduction

In feature selection (FS) the search problem of finding a subset of d features from the given set of D measurements, $d < D$, so as to optimize a chosen criterion [3] has been of interest for a long time. The aim of FS is to reduce data acquisition cost and/or improve pattern recognition performance.

In FS algorithm design it is generally assumed that *any* improvement in the criterion value leads to better feature subset. Nevertheless, this principle has been challenged [10]–[12] showing that the strict application of this rule may easily lead to overfitting and consequently to poor generalization performance even with the best available feature subset evaluation schemes. Unfortunately, there seems to be no way of defining FS criteria capable of avoiding this problem in general. In this paper we present an alternative workaround targeted specifically at improving the robustness of decisions about feature inclusion/removal in the course of feature subset search.

A. Common Process of Criterion Maximization

Following the common paradigm we assume that higher criterion value is meant to depict better subset. Many common sub-optimal FS algorithms can be viewed as generators of a sequence of candidate feature subsets and respective criterion values (see Fig. 1). Intermediate solutions are usually selected among the candidate subsets as the ones with the highest criterion value discovered so far. Intermediate solutions are used to further guide the search. The solution with the highest overall citerion value is eventually considered to be the result. In the course of search the candidate feature subsets may yield fluctuating criterion values while the criterion values of intermediate solutions usually form a nondecreasing sequence. The search generally continues as long as intermediate solutions improve, no matter how significant the improvement is and often without respect to other effects like excessive subset size increase. This type of scheme is followed by various sequential search algorithms [3], [9], genetic algorithms [6], simulated annealing [4] or tabu search [15], etc.

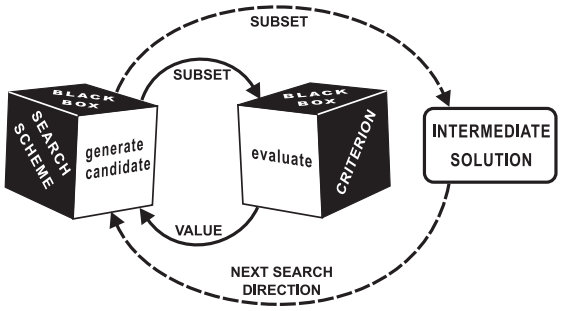

Figure 1. Feature selection algorithms can be viewed as black box procedures generating a sequence of candidate subsets with respective criterion values, among which intermediate solutions are chosen.

B. The Problem of Fragile Feature Preference

In many FS tasks it can be observed that the difference between criterion values of successive intermediate solutions decreases in time and often becomes negligible. Yet minimal change in criterion value may be accompanied by substantial changes in subset contents. This can easily happen, e.g., when many of the considered features are important but redundant to various degrees with respect to the chosen criterion, or when there is large number of features carrying limited but nonzero information (this is common, e.g., in text

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.1068

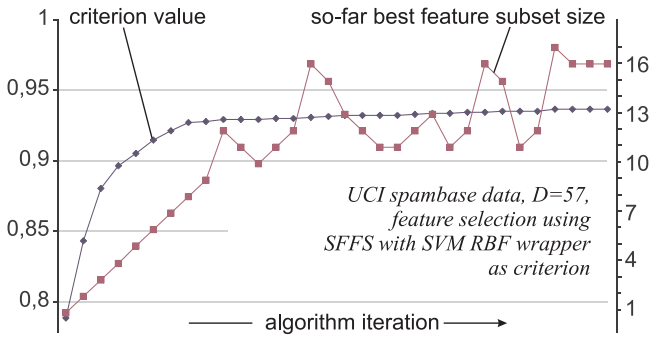

Figure 2. In many FS tasks very low criterion increase is accompanied by fluctuations in selected subsets; both in size and contents

categorization [13]). We illustrate this phenomenon in Fig. 2, showing the process of selecting features on *spambase* data [1] using SFFS algorithm [9] and estimated classification accuracy of Support Vector Machine (SVM, [2]) as criterion [8]. Considering only those tested subset candidates with criterion values within 1% difference from the final maximum achieved value, i.e., values from $[0.926, 0.936]$, their sizes fluctuate from 8 to 17. This sequence of candidate subsets yields *average Tanimoto distance* $\overline{S_S}$ [7] as low as 0.551 on the scale $[0, 1]$ (where 0 marks disjunct sets and 1 marks identical sets). This suggests that roughly any two of these subsets differ almost in half of their contents. Clearly, notable fluctuations in feature subset contents following from minor criterion value improvement are unlikely to lead to reliable final classification system. Correspondingly, Raudys [10] argues that to prevent overfitting it may be better to consider as FS result a subset with slightly lower than the best achieved criterion value.

II. Tackling The Problem of Fragile Feature Subset Preferences

Following the observations above, we propose to *treat as equal* (effectively indistinguishable) all subsets known to yield criterion value within a pre-defined (very small) distance from the maximum known at the current algorithm stage. Intermediate solutions then need to be selected from the treated-as-equal subset groups using a suitable *secondary criterion*. A good secondary criterion should be able to compensate for the primary criterion's deficiency in distinguishing among treated-as-equal subsets. Nevertheless, introducing the secondary criterion opens up alternative usage options as well, see Sect. II-B.

The idea of the secondary criterion is analogous to the principle of penalty functions as used, e.g., in two-part objective function consisting of goodness-of-fit and number-of-variables parts [5]. However, in our approach we propose to keep the evaluation of primary and secondary criteria separated. Avoiding the combination of two criteria into one objective function is advantageous as it a) avoids the problem of finding reasonable combination parameters (weights) of potentially incompatible objective function parts and b) enables to use the secondary criterion as supplement only in cases when the primary criterion response is not decisive enough. Remark: The advantage of separate criteria evaluation comes at the cost of necessity to specify which subset candidates are to be treated as equal, i.e., to set a threshold depending on the primary criterion. This, however, is transparent to define (see below) and, when compared to two-part objective functions, allows for finer control of the FS process.

A. Secondary Criterion Evaluation Mechanism

Let $J_1(\cdot)$ denote the primary FS criterion to be maximized by the chosen FS algorithm. Let $J_2(\cdot)$ denote the secondary FS criterion for resolving the "treated-as-equal" cases. Let $\lambda \in [0, 1]$ denote the *equality threshold* parameter. Throughout the course of search two pivot subsets, X^{max} and X^{sel}, are to be updated after each criterion evaluation. Let X^{max} denote the subset yielding the maximum J_1 value known so far. Let X^{sel} denote the currently selected subset (intermediate solution). When the search process ends, X^{sel} is to become the final solution.

The chosen backbone FS algorithm is used in its standard way to maximize J_1. It is the mechanism proposed below that simultaneously keeps selecting an intermediate result X^{sel} among the currently known "treated-as-equal" alternatives to the current X^{max}, allowing $\mathrm{X}^{sel} \neq \mathrm{X}^{max}$ if X^{sel} is better than X^{max} with respect to J_2 while being only negligibly worse with respect to J_1, i.e., provided $J_1(\mathrm{X}^{sel}) \geq (1-\lambda) \cdot J_1(\mathrm{X}^{max}) \wedge J_2(\mathrm{X}^{sel}) > J_2(\mathrm{X}^{max})$.

FS Algorithm Extension

Whenever the backbone FS algorithm evaluates a feature subset X (depicting any subset evaluated at any algorithm stage), the following update sequence is to be called:

if $J_1(\mathrm{X}) > J_1(\mathrm{X}^{max})$ **then**
  make X the new X^{max}
  {*now the current* X^{sel} *may not be valid any more*}
  if $J_1(\mathrm{X}^{sel}) < (1-\lambda) \cdot J_1(\mathrm{X}^{max})$
  or $J_2(\mathrm{X}^{sel}) \leq J_2(\mathrm{X}^{max})$ **then**
    make X also the new X^{sel}
  end if
else {X *still may be better than the current* X^{sel}}
  if $\big(J_1(\mathrm{X}) \geq (1-\lambda) \cdot J_1(\mathrm{X}^{max})$ **and** $J_2(\mathrm{X}) > J_2(\mathrm{X}^{sel})\big)$
  or $\big(J_2(\mathrm{X}) = J_2(\mathrm{X}^{sel})$ **and** $J_1(\mathrm{X}) > J_1(\mathrm{X}^{sel})\big)$ **then**
    make X the new X^{sel}
  end if
end if

The proposed mechanism does not affect the course of search of the primary FS algorithm; it only adds a form of lazy solution update. Note that the presented mechanism is applicable with a large class of FS algorithms (cf. Sect I-A).

Table I
RESULTS – LOWER-DIMENSIONAL DATA [*Dimensionality*](*No. of classes, No. of all samples*) – CLASSIFIER ACCURACY FOR VARIOUS λ, SFFS

SFFS		dermatol. [*D*=34] (6, *358*)			spectf [*D*=44] (2, *267*)			wdbc [*D*=30] (2, *569*)			spambase [*D*=57] (2, *4601*)		
crit.	λ	feat. subs. size	train acc.	test acc.	feat. subs. size	train acc.	test acc.	feat. subs. size	train acc.	test acc.	feat. subs. size	train acc.	test acc.
SVM RBF	0	8	.977	.917	2	.827	.761	7	.943	.923	16	.937	.883
	0.001	dtto	dtto	dtto	dtto	dtto	dtto	dtto	dtto	dtto	dtto	dtto	dtto
	0.005	dtto	dtto	dtto	dtto	dtto	dtto	5	.940	.923 •	10	.934	.879
	0.01	dtto	dtto	dtto	dtto	dtto	dtto	4	.936	.926 •	9	.930	.884
	0.02	7	.966	.922 •	dtto	dtto	dtto	2	.926	.912	8	.921	.870
	0.03	6	.955	.922 •	dtto	dtto	dtto	dtto	dtto	dtto	6	.912	.872
	0.04	5	.944	.933 •	1	.797	.791 •	dtto	dtto	dtto	5	.904	.871
	0.05	dtto	dtto	dtto	dtto	dtto	dtto	2	.919	.919	4	.896	.866
3NN	0	16	.994	.933	11	.948	.769	5	.954	.937	30	.930	.871
	0.001	dtto	dtto	dtto	dtto	dtto	dtto	dtto	dtto	dtto	24	.930	.872 •
	0.005	dtto	dtto	dtto	dtto	dtto	dtto	3	.951	.937 •	20	.926	.871 •
	0.01	dtto	dtto	dtto	dtto	dtto	dtto	dtto	dtto	dtto	18	.923	.876 •
	0.02	6	.983	.950 •	dtto	dtto	dtto	2	.937	.930	14	.913	.867
	0.03	5	.966	.939 •	7	.925	.776 •	dtto	dtto	dtto	9	.905	.856
	0.04	dtto	dtto	dtto	dtto	dtto	dtto	dtto	dtto	dtto	8	.896	.828
	0.05	dtto	dtto	dtto	6	.910	.716	1	.912	.881	7	.887	.807

B. Secondary Criterion Usage Options

The J_2 criterion can be utilized for various purposes. Depending on particular problem it may be possible to define J_2 to *distinguish better among subsets* that J_1 fails to distinguish reasonably enough.

The simplest yet useful alternative is to utilize J_2 for emphasising the *preference of smaller subsets*. To achieve this, J_2 is to be defined as $J_2(\mathrm{X}) = -|\mathrm{X}|$. Smaller subsets not only mean lower measurement cost, but more importantly in many problems the forced reduction of subset size may help to remove features that over-fit, what consequently leads to better generalization (see Section III). More generally, J_2 can be used to incorporate *feature acquisition cost* minimization into the FS process. Provided a weight (cost) $w_i, i = 1, \ldots, D$ is known for each feature, then the appropriate secondary criterion can be easily defined as $J_2(\mathrm{X}) = -\sum_{x_i \in \mathrm{X}} w_i$.

III. APPLIED EXAMPLE

We illustrate the potential of the proposed methodology on a series of experiments where J_2 was used for emphasising the *preference of smaller subsets* (see Sect. II-B). For this purpose we used several data-sets from UCI repository [1] and one data-set – xpxinsar satellite – from Salzburg University. Table I collects results obtained using the extended version (see Sect. II) of the Sequential Forward Floating Search (SFFS, [9]). Table II collects results obtained using the extended version (see Sect. II) of the Dynamic Oscillating Search (DOS, [14]). Both methods have been used in wrapper setting [8], i.e., with estimated classifier accuracy as FS criterion. For this purpose we have used Support Vector Machine (SVM) with Radial Basis Function kernel [2] and 3-Nearest Neighbor classifier accuracy estimates. To estimate final classifier accuracy on independent data we split each dataset to equally sized parts; the training part was used in 3-fold Cross-Validation manner to evaluate wrapper criteria in the course of FS process, the testing part was used only once for independent classification accuracy estimation.

We repeated each experiment for different *equality thresholds* λ, ranging from 0.001 to 0.05 (note that due to the wrapper setting both considered criteria yield values from $[0, 1]$). Tables I and II show the impact of changing equality threshold to classifier accuracy on independent data. First row ($\lambda = 0$) equals standard FS algorithm operation without the extension proposed in this paper. By black bullets we emphasize cases where the proposed mechanism led to improvement, i.e., the selected subset size has been reduced with better or equal accuracy on independent test data. Note that positive effect of nonzero λ can be observed in significant number of cases. Note in particular that in many cases the number of features could be reduced to less than one half of what would be the standard FS method's result (cf. in Table I the dermatology–3NN case and in Table II the gisette–SVM, xpxinsar–SVM, spambase–SVM and madelon–3NN cases). However, it can be also seen that the effect is strongly case dependent. It is hardly possible to give general recommendation about suitable λ value, except that improvements in some of our experiments have been observed for various λ values up to roughly 0.1.

IV. CONCLUSION

We have pointed out a problem of feature subset preference fragility (over-emphasized importance of negligible criterion value increase) as one of factors that make many FS methods more prone to over-fitting. We propose an

Table II
RESULTS – HIGHER-DIMENSIONAL DATA [*Dimensionality*](*No. of classes, No. of all samples*) – CLASSIFIER ACCURACY FOR VARIOUS λ, DOS($\Delta = 15$)

DOS(15)		gisette [*D*=5000](2, *1000*)			madelon [*D*=500](2, *2000*)			xpxinsar [*D*=57](7, *1721*)			spambase [*D*=57] (2, *4601*)		
crit.	λ	feat. subs. size	train acc.	test acc.	feat. subs. size	train acc.	test acc.	feat. subs. size	train acc.	test acc.	feat. subs. size	train acc.	test acc.
SVM RBF	0	10	.922	.856	21	.841	.804	12	.873	.863	16	.930	.873
	0.001	9	.921	.860 •	dtto	dtto	dtto	dtto	dtto	dtto	11	.929	.877 •
	0.005	7	.918	.862 •	17	.837	.817 •	9	.871	.867 •	8	.927	.878 •
	0.01	5	.914	.854	15	.833	.812 •	7	.866	.897 •	dtto	dtto	dtto
	0.02	3	.906	.852	13	.825	.816 •	6	.864	.896 •	6	.915	.877 •
	0.03	dtto	dtto	dtto	dtto	dtto	dtto	5	.856	.871 •	5	.905	.873 •
	0.04	2	.890	.856 •	12	.811	.793	4	.840	.845	4	.896	.866
	0.05	dtto	dtto	dtto	dtto	dtto	dtto	dtto	dtto	dtto	dtto	dtto	dtto
3NN	0	15	.958	.904	18	.891	.844	16	.847	.854	36	.930	.859
	0.001	dtto	dtto	dtto	dtto	dtto	dtto	14	.847	.854 •	32	.929	.858
	0.005	13	.954	.898	13	.888	.842	12	.844	.848	23	.926	.868 •
	0.01	11	.950	.892	9	.883	.850 •	10	.840	.847	21	.922	.869 •
	0.02	8	.940	.892	7	.877	.848 •	9	.837	.825	14	.913	.866 •
	0.03	6	.930	.874	6	.869	.847 •	5	.823	.842	12	.907	.836
	0.04	5	.922	.89	5	.858	.854 •	dtto	dtto	dtto	dtto	dtto	dtto
	0.05	4	.914	.87	dtto	dtto	dtto	4	.812	.837	8	.892	.836

algorithmic workaround applicable with many standard FS methods. Moreover, the proposed algorithmic extension enables improved ways of standard FS algorithms' operation, e.g., taking into account feature acquisition cost. We show just one of possible applications of the proposed mechanism on a series of examples where two sequential FS methods are modified to put more preference on smaller subsets in the course of search. Although the main course of search is aimed at criterion maximization, smaller subsets are permitted to be eventually selected if their respective criterion value is negligibly lower than the known maximum. The examples show that this mechanism is well capable of improving classification accuracy on independent data.

ACKNOWLEDGMENT

We thank Helmut Mayer from Salzburg University for providing the xpxinsar data. The work has been supported by grants of the Czech Ministry of Education 2C06019 ZIMOLEZ and 1M0572 DAR, and the GACR Nos. 102/08/0593 and 102/07/1594.

REFERENCES

[1] A. Asuncion and D. Newman. UCI mach. learn. repository, http://archive.ics.uci.edu/ml/, 2007.

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

[3] P. A. Devijver and J. Kittler. *Pattern Recognition: A Statistical Approach*. Prentice Hall, 1982.

[4] F. W. Glover and G. A. Kochenberger, editors. *Handbook of Metaheuristics*, volume 57 of *Int. Ser. in Operat. Research & Management Science*. Springer, 2003.

[5] I. Guyon and A. Elisseeff. An introduction to variable and feat. sel. *J. Mach. Learn. Res.*, 3:1157–1182, 2003.

[6] F. Hussein, R. Ward, and N. Kharma. Genetic algorithms for feature selection and weighting, a review and study. *icdar*, 00:1240, 2001.

[7] A. Kalousis, J. Prados, and M. Hilario. Stability of feature selection algorithms. *Knowledge and Information Systems*, 12(1):95–116, 2007.

[8] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artif. Intell.*, 97(1-2):273–324, 1997.

[9] P. Pudil, J. Novovičová, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11):1119–1125, 1994.

[10] Š. J. Raudys. Feature over-selection. In *Proc. S+SSPR, LNCS 4109*, pages 622–631. Springer-Verlag, 2006.

[11] J. Reunanen. A pitfall in determining the optimal feature subset size. In *Proc. 4th Int. Workshop PRIS*, pages 176–185, Porto, Portugal, 2004.

[12] J. Reunanen. Less biased measurement of feature selection benefits. In *SLSFS 2005, Revised Selected Papers*, volume LNCS 3940, pages 198–208. Springer, 2006.

[13] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, March 2002.

[14] P. Somol, J. Novovičová, J. Grim, and P. Pudil. Dynamic oscillating search algorithms for feature selection. In *Proc. ICPR*. IEEE Computer Society, 2008.

[15] M. A. Tahir, A. Bouridane and F. Kurugollu. Simultaneous FS and feature weighting using hybrid tabu search/k-nearest neighbor classifier. *Patt. Rec. Lett.*, 28(4):438–446, 2007.