Improving Stability of Feature Selection Methods*

Pavel Křížek[1], Josef Kittler[2], and Václav Hlaváč[1]

[1] Czech Technical University in Prague, Center for Machine Perception,
Karlovo nám. 13, 121 35 Prague 2, Czech Republic
[2] University of Surrey, Centre for Vision, Speech, and Signal Processing,
GU2 7XH Guildford, United Kingdom
{krizekp1, hlavac}@fel.cvut.cz, j.kittler@surrey.ac.uk

Abstract. An improper design of feature selection methods can often lead to incorrect conclusions. Moreover, it is not generally realised that functional values of the criterion guiding the search for the best feature set are random variables with some probability distribution. This contribution examines the influence of several estimation techniques on the consistency of the final result. We propose an entropy based measure which can assess the stability of feature selection methods with respect to perturbations in the data. Results show that filters achieve a better stability and performance if more samples are employed for the estimation, i.e., using leave-one-out cross-validation, for instance. However, the best results for wrappers are acquired with the 50/50 holdout validation.

Keywords: Feature selection, stability, entropy.

1 Introduction

Many tasks in statistical pattern recognition are characterised by high dimensional data which have to be processed and analysed using statistical tools. A data sample is a vector formed generally by several hundreds of measurements, called features. Examples of such data are measurements arising in text recognition, genetic engineering, astronomy, etc. The problem of analysing and processing such multivariate sensory information can be aggravated by a relatively small sample size available for learning. A small sample statistics results in inaccurate parameter estimates of the data models. Thus, a poor generalisation is achieved on unknown data. This phenomenon is known as the *curse of dimensionality* [1]. A common solution is to reduce the dimensionality and employ, for example, only those features that are relevant to a given problem. This task is known as the *feature selection* [1].

There are many approaches to feature selection, however, all in principle involve two main ingredients: i) an *objective function* (criterion) which reflects

* This work was supported by the EU INTAS project PRINCESS 04-77-7347 and by the Czech Ministry of Education under Project 1M0567.

W.G. Kropatsch, M. Kampel, and A. Hanbury (Eds.): CAIP 2007, LNCS 4673, pp. 929–936, 2007.

informativeness of feature subsets, and ii) a *search strategy*. The search strategy is fixed and employs usually *feature ranking* [7] or *subset search* [1,11] methods. Both approaches are premised on certain principles which guide the search through the feature space and determine what results can be ideally achieved. The objective function guiding the search can be either *classifier specific* [9,5] (so called *wrappers*) or *classifier independent* [7,5] (so called *filters*). In general, learning algorithms designed using features selected by wrappers have been shown to achieve a better predictive accuracy. Nevertheless, wrappers are computationally very intensive and tend to over-fit [9]. Filters can be seen more as a preprocessing step for a subsequent learning, because the objective function is not directly linked to minimising the error rate of a particular classifier. Filters usually execute quite fast and provide a general approach to feature selection.

One issue that has been relatively neglected in the literature is the *stability of feature selection algorithms*, i.e., the sensitivity of the solution (selected features) to perturbations in the input data. The motivation behind exploring the stability is to provide an evidence that the selected features are relatively robust to slight changes in the data. Feature selection methods producing consistent solution on the given data are preferable to those with highly volatile outputs. Possibly, the only related publications discussing the feature selection stability topic are [2,6,10]. However, the proposed stability measures have many limitations, unclear motivation, and empirically estimated bounds. Furthermore, these papers do not clearly motivate remedies for improving the stability and performance. We show that for any given feature selection algorithm, the stability and performance can be significantly improved by a careful use of the data available.

The paper is organised as follows. Section 2 formulates the stability problem. In the same section, we propose and theoretically justify a measure which can assess the stability of feature selection algorithms with respect to perturbations in the data. We also suggest a concept how the stability and performance can be improved. The experimental set-up is described in Section 3. Section 4 presents and discusses the numerical results. Conclusions are drawn in Section 5.

2 Stability of Feature Selection Methods

2.1 The Stability Problem

It is not often realised in the literature on feature selection that values of the objective function guiding the search for the best feature set are random variables with some probability distribution, no matter what type of the search strategy or criterion is adopted. This originates from the randomly sampled data at the feature selection input. The exact criterion value is unknown and the search strategy has to work just with an estimate acquired on the available data. Note that the accuracy of the criterion estimate can considerably influence the result of the feature selection.

Suppose that we carry out $T \in \mathbb{N}$ runs of a feature selection algorithm on randomly sampled data. We would like to find some kind of measure which would allow us to assess the feature selection method stability with respect to perturbations in the data, i.e., to assess the sensitivity of the selected features to variations in the input data. We will consider a concept which reflects the most frequently selected feature subsets.

It has to be emphasised that the *stability does not say anything about the performance* of the selected features. It *just indicates the sensitivity* of a feature selection method output to random perturbations in the input data. Large variations in the selected feature subsets signify that something is wrong. For instance, the feature selection algorithm is not appropriate for a given data, or there are not enough samples, or too many correlated variables or feature subsets with very similar information content, etc. Thus, less confidence should be assigned to feature sets that change radically with slight variations of the input data or perhaps it is advisable to refrain completely from the feature selection.

2.2 Stability of Feature Sets

Notice that various feature selection techniques select different feature subsets with a certain probability if the input data for training are randomly sampled from the original data set. One extreme case is a *random feature selection* which selects every feature subset with the same probability and thus produces a uniform probability distribution. The other extreme is a *perfectly stable feature selection* method which all the time selects the same feature subset and thus creates a single peak probability distribution. It appears that the stability of feature selection algorithms can be assessed through the properties of the generated probability distributions of the selected feature subsets. Our interest is, of course, in feature selection algorithms that produce probability distributions far from the uniform and close to the peak one.

A convenient measure quantifying randomness of a system is *entropy* [12]. Entropy is a real function defined on a set of probability distributions. In information theory, the concept of entropy indicates the amount of uncertainty about an event associated with a given probability distribution. The entropy is maximal for a uniform probability distribution (i.e., outcome of random feature selection). If the event is certain (i.e., outcome of perfectly stable feature selection) then the entropy is zero.

There are several entropy measures in information theory. We derive the stability measure from the *Shannon entropy*, see [12],

$$H(X) = -\sum_{i=1}^{m} p(x_i) \log p(x_i) \, . \qquad (1)$$

Here X is a discrete random variable with possible states $X = \{x_1, x_2, \ldots, x_m\}$ (i.e., particular feature subsets), $m \in \mathbb{N}$ is the number of all possible states (i.e., the number of all different feature subsets), and $p(x_i)$ is the probability of the i-th state occurrence (i.e., probability of selecting a particular feature subset).

Let $n \in \mathbb{N}$ be the problem dimensionality, $k \in \{1, 2, \ldots, n\}$ indicates the size of the feature subset, and $T \in \mathbb{N}$ is the number of evaluation trials with randomly sampled data. The frequencies of selected feature subsets are recorded over T trials in a histogram given by a structure with entries $G_{jk} \in \mathbb{Z}^+$, where $\mathbb{Z}^+$ are non-negative integers, $j = 1, 2, \ldots, C(n,k)$, and $C(n,k) = \binom{n}{k}$ is the number of all possible feature combinations. The histogram structure can be implemented by recording different feature subsets and their frequencies G_{jk}. The probability estimates of particular feature subsets occurrence can be determined by normalising the histogram entries by the number of trials, i.e., $\overline{G}_{jk} = G_{jk}/T$. Thus, all bin values $\overline{G}_{jk}$ are scaled into the interval $[0,1]$ and $\sum_{j=1}^{C(n,k)} \overline{G}_{jk} = 1$.

Based on the Shannon entropy (1), the following stability measure can be constructed for a feature subset of a fixed size k,

$$\gamma_k = - \sum_{j=1}^{C(n,k)} \overline{G}_{jk} \log \overline{G}_{jk} \ . \tag{2}$$

In reality, the histogram structure is sparse, because the number of evaluation trials T is small compared to the theoretical combinatorial amount of all possible feature combinations which yields the maximal number of the histogram entries $G_{jk} > 0$ to be $j_{\max} = \min[T, C(n,k)]$. Thus, the stability measure (2) ranges in the interval $0 \leq \gamma_k \leq \log \min[T, C(n,k)]$. Deriving both bounds is a straightforward task. The lower bound expresses that there is only one uniquely selected feature subset of size k over all T trials which corresponds to a perfectly stable feature selection algorithm, hence zero entropy. The theoretical upper bound is based on the assumption that an arbitrary feature subset of size k is selected over T trials with the same probability $\overline{G}_{jk} = (\min[T, C(n,k)])^{-1}$. It can be seen that the stability measure (2) creates an ordering and thus, the stability of examined feature selection algorithms can be compared.

2.3 Improving Stability and Performance

The primary key for the improved stability is an appropriate estimate of the objective function values. With a better criterion estimate, the search algorithm is also more likely to converge to its optimal solution with respect to the unknown data underlying probability distribution. Selected features thus achieve a better generalisation and performance. Surprisingly, it is very hard to find a research paper which would follow this basic rule. However, the price we may have to pay for a better estimate is an exponentially increasing execution time.

The commonly adopted estimation methods in statistical pattern recognition are data re-sampling techniques like cross-validation, holdout validation, or bootstrap, see [1,3,8]. Only wrapper design sometimes applies five or ten-fold cross-validation to avoid over-fitting of a classifier employed in the objective function definition, see [9,10]. No estimation is ever done in filter approaches.

3 Experiment Design and Description

The motivation behind the following experiments is to investigate how the choice of a data re-sampling technique for the objective function estimation influences the stability and performance of filter and wrapper feature selection approaches.

3.1 Data Used in Experiment

A simple artificial two-class problem is synthesised for the purpose of this study, since we would like to have a control all over the experiment. The data are designed so that ordinary ranking or greedy feature selection methods fail to find an optimal feature subset of a minimal size. Samples are derived from a 20-dimensional normal probability distribution with a common covariance matrix Σ and mean values $\mu = \mu_1 = -\mu_2$. First class consists of a component $N(\mu, \Sigma)$, and second class of a component $N(-\mu, \Sigma)$. The common covariance matrix and the mean values comprise several blocks which simulate different qualities of features.

The first block contains statistically independent features with identical discriminatory ability of the particular features. The parameters are the following

$$\Sigma^{1,\dots,3} = I_3 \quad \text{and} \quad \mu^{1,\dots,3} = [0.635, 0.635, 0.635]^\top ,$$

where I_d is the $d \times d$ identity matrix for $d \in \mathbb{N}$. Upper indices in $\Sigma^{i,j,k,\cdots}$ and $\mu^{i,j,k,\cdots}$ indicate the corresponding coordinates of the block.

A nested pair of features with indices $\{4, 6\}$ is hidden in the second block. The parameters are given by

$$\Sigma^{4,\dots,6} = \begin{bmatrix} 1.05 & 0.48 & 0.95 \\ 0.48 & 1.0 & 0.20 \\ 0.95 & 0.20 & 1.05 \end{bmatrix} \quad \text{and} \quad \mu^{4,\dots,6} = \begin{bmatrix} 0.5 \\ 0.4 \\ 0 \end{bmatrix} .$$

The third block contains statistically independent features with decreasing discriminatory ability of the particular features. The parameters are

$$\Sigma^{7,\dots,13} = I_7 \quad \text{and} \quad \mu^{7,\dots,13} = [0.636, 0.546, 0.455, 0.364, 0.273, 0.182, 0.091]^\top .$$

The last block contains only noise with parameters

$$\Sigma^{14,\dots,20} = I_7 \quad \text{and} \quad \mu^{14,\dots,20} = [0, 0, 0, 0, 0, 0, 0]^\top .$$

Finally, all the dimensions of the problem are randomly permuted to make sure that an examined feature selection algorithm will not follow feature ordering created in the design above. Data in our experiment contain 500 samples per class. It is interesting to note that the theoretical classification error of such data is about 2.3%.

3.2 Experiment Set-Up

In experiments, we examine a filter and wrapper variant of the state-of-the-art feature selection technique known as the Sequential Forward Floating Search (SFFS) algorithm [11]. The filter version applies the Mahalanobis distance [1] in the objective function definition. The wrapper form uses prediction accuracy of a linear decision rule created by the Gaussian classifier [1]. Both feature selection algorithms fit the data so they should find the exact solution.

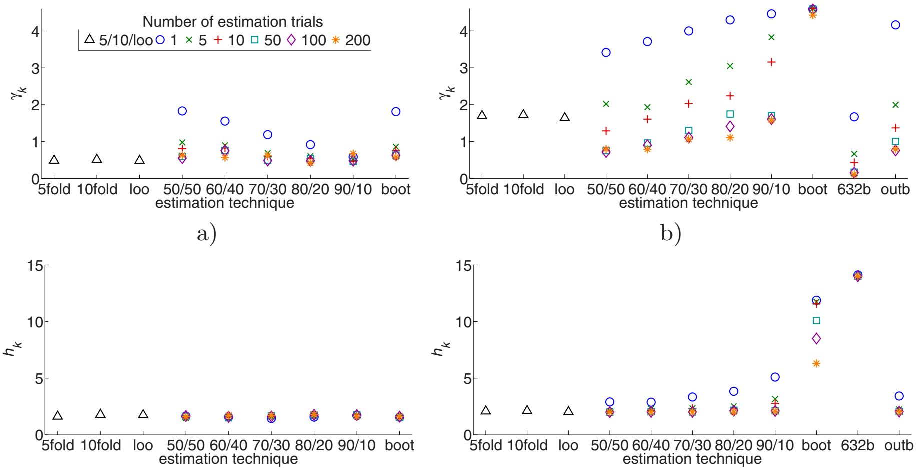
For the objective function estimation, we consider five-fold, ten-fold, and leave-one-out cross-validation, repeated holdout validation scheme with 50/50, 60/40, 70/30, 80/20, and 90/10 splits of data to the training/validation part, respectively, and estimation methods based on sampling with replacement such as bootstrap, .632 bootstrap, and out-of-bootstrap, see [1,3,8] for more details. Holdout and bootstrap techniques employ 1, 5, 10, 50, 100, and 200 trials, respectively, for the objective function estimation. All methods are *stratified* [8] so the class a priori probabilities are preserved within the data re-sampling.

The feature selection is applied in all experimental set-up conditions described above. To acquire a good statistics on the probability estimates of the selected feature subsets, one hundred evaluation trials are performed with 90% of the data randomly sampled from the original data set for training. The stability measure in Equation (2) is determined from the acquired probability estimates.

The performance is assessed by the Hamming distance [4,2] between the exact solution and the selected features in order to check the real quality of the result. The exact solution is found by the SFFS algorithm which employs the true data probability distribution and the Mahalanobis distance in the objective function definition. The final performance is averaged over all one hundred trials.

4 Numerical Results and Discussion

The stability results for the filter and wrapper are depicted in Figures 1a and 1b. The performance of the selected features measured by the average Hamming distance is shown in Figures 1c and 1d. Graphs are depicted only for a feature subset of size $k = 8$ as similar behaviour was observed for all subset sizes.



a) b)

c) d)

Fig. 1. The stability factor γ_k for a) filter, b) wrapper, and the averaged Hamming distance h_k for c) filter, d) wrapper, displayed with respect to the re-sampling technique involved in the objective function estimation for a feature subset of size $k = 8$

The filter variant of the SFFS algorithm achieves better stability if more samples are employed in the objective function estimation, i.e., using techniques like ten-fold or leave-one-out cross-validation, for instance. Such result can be interpreted by the general fact that more samples available for learning lead to better parameter estimates and thus to a more consistent solution. This effect is clearly visible for the holdout validation, where the stability gets worse with less samples available for the estimation. However, the estimate gets better with the increasing number of trials. It appears that at some point the absolute stability factor saturates and becomes more or less independent of the examined re-sampling techniques if the number of estimation trials is sufficiently large.

The situation is quite different for the wrapper approach. The best stability and the performance result is obtained for the repeated 50/50 holdout validation employed in the objective estimation. The popular ten-fold or leave-one-out cross-validation does not achieve such a good solution. Our interpretation is as follows. Although a classifier designed with more samples has better parameter estimates, the objective function is based on the prediction accuracy using the validation data. Having fewer samples available for validation implies higher variance of the prediction accuracy. Thus, the feature selection algorithm becomes more sensitive to random perturbations in the data and fails to find a consistent solution. The number of the training and validation samples compete against each other which explains why the 50/50 holdout validation achieves the best results. It seems again, that the stability factor saturates at some point if the number of estimation trials is sufficiently large.

Wrappers appear to be much more sensitive to the correct objective function estimate than filters. Notice that the .632 bootstrap achieved the best stability factor, however, its performance is by far the worst. Bootstrap techniques are supposed to give estimates with low variance [3] which explains a good stability. Nevertheless, the bias of the estimate is high and as a result the wrapper converged to a wrong solution.

As for the performance, the non-zero Hamming distance indicates that none of the solutions is identical with the theoretical best feature subset. The slight bias (Hamming distance “2”) is probably caused by an in-accurate estimate of the weakest features discriminatory power.

5 Conclusions

The feature selection results should always be strengthen by some confidence in the solution. For this purpose, we designed and theoretically justified an entropy based measure which can assess the stability of any kind of a feature selection algorithm with respect to random perturbations in the input data. Furthermore, we derived bounds on the stability measure analytically. For a given feature selection method and data, the stability factor has to be determined empirically over a number of evaluation trials with randomly sampled training data. A large number of trials is required to guarantee a sufficient accuracy of the probability estimates of the selected feature subsets occurrence.

The stability and performance of feature selection methods can be improved to a certain degree by a suitable algorithm design. This can be achieved by an appropriate estimate of the objective function guiding the search for the best subset of features. Our experiments showed that filters achieved better stability and performance if more samples were employed in the objective function estimation, i.e., using re-sampling techniques like ten-fold or leave-one-out cross-validation, or 90/10 holdout validation. For wrappers, however, the best estimation technique appeared to be the 50/50 holdout validation. Wrappers also turned up to be much more sensitive to incorrect objective function estimation than filters.

Nevertheless, a good stability of a feature selection algorithm on the given data is just a necessary condition for a good performance of the selected features. Both, the stability and performance should always be analysed together, because the stability itself does not say anything about the selected features quality.

References

1. Devijver, P.A., Kittler, J.: Pattern Recognition: A Statistical Approach. Prentice Hall, Englewood Cliffs (1982)
2. Dunne, K., Cunningham, P., Azuaje, F.: Solutions to instability problems with sequential wrapper-based approaches to feature selection. Technical Report TCD-CD-2002-28, Dept. of Computer Science, Trinity College, Dublin, Ireland (2002)
3. Efron, B., Tibshirani, R.: Estimating the error rate of a prediction rule: Improvement on cross-validation. Technical Report TR-477, Dept. of Statistics, Stanford University (1995)
4. Hamming, R.W.: Error detecting and error correcting codes. Bell System Technical Journal 26(2), 147–160 (1950)
5. Jain, A., Zongker, D.: Feature selection: Evaluation, application and small sample performance. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(2), 153–158 (1997)
6. Kalousis, A., Prados, J., Hilario, M.: Stability of feature selection algorithms. Proceedings of the 5th IEEE International Conference on Data Mining, Houston, Texas, pp. 218–225. IEEE Computer Society, Los Alamitos (2005)
7. Kirra, K., Rendell, L.A.: The feature selection problem: Traditional methods and a new algorithm. In: Proceedings of the 10th National Conference on Artificial Intelligence, San Jose, CA, pp. 129–134. MIT Press, Cambridge, MA (1992)
8. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proceedings of the Joint Conference on Artificial Intelligence, pp. 1137–1145. Morgan Kaufmann, Montreal, Canada, San Mateo, CA (1995)
9. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence 97(1-2), 273–324 (1997)
10. Kuncheva, L.I.: A stability index for feature selection. In: Proceedings of the 25th International Multi-Conference on Artificial Intelligence and Applications, pp. 390–395 (February 2007)
11. Pudil, P., Novovičová, J., Kittler, J.: Floating search methods in feature selection. Pattern Recognition Letters 15, 1119–1125 (1994)
12. Shannon, C.: Mathematical theory of communication. Bell System Technology Journal 27, 379–423, 623–656 (1948)