Feature Selection Based on Mutual Correlation

Michal Haindl[1], Petr Somol[1], Dimitrios Ververidis[2], and Constantine Kotropoulos[2]

[1] Institute of Information Theory and Automation,
Academy of Sciences CR,
Prague, CZ182 08, Czech Republic
{haindl,somol}@utia.cas.cz http://ro.utia.cz
[2] Dept. of Informatics, Aristotle Univ. of Thessaloniki
Box 451, Thessaloniki 541 24, Greece
{jimver,costas}@aiia.csd.auth.gr http://poseidon.csd.auth.gr

Abstract. Feature selection is a critical procedure in many pattern recognition applications. There are two distinct mechanisms for feature selection namely the wrapper methods and the filter methods. The filter methods are generally considered inferior to wrapper methods, however wrapper methods are computationally more demanding than filter methods. A novel filter feature selection method based on mutual correlation is proposed. We assess the classification performance of the proposed filter method by using the selected features to the Bayes classifier. Alternative filter feature selection methods that optimize either the Bhattacharyya distance or the divergence are also tested. Furthermore, wrapper feature selection techniques employing several search strategies such as the sequential forward search, the oscillating search, and the sequential floating forward search are also included in the comparative study. A trade off between the classification accuracy and the feature set dimensionality is demonstrated on both two benchmark datasets from UCI repository and two emotional speech data collections.

1 Introduction

Feature selection is defined as the process of selecting D most discriminatory features out of $d \geq D$ available ones [1]. Feature subset selection aims to identify and remove as much irrelevant and redundant information as possible. Feature transformation is defined as the process of projecting the d measurements to a lower dimensional space through a linear or non-linear mapping. Principal component analysis and linear discriminant analysis are probably the most common feature transformations [4]. Both feature extraction and feature transformation reduce data dimensionality and allow learning algorithms to operate faster and more effectively on large datasets and even to improve classification accuracy in some cases. Depending on the available knowledge of class membership, the feature selection can be either supervised or unsupervised.

The feature selection problem is NP-hard. So, the optimal solution is not guaranteed to be found unless except exhaustive search in the feature space is

performed [1]. Two approaches to feature selection are commonly used namely the wrapper methods and the filter methods. The former use the actual classifier to select the optimal feature subset, while the latter select features independently of the classifier. The filter methods use probability based distances independent of the classification such as the Bhattacharyya distance, the Chernoff distance, the Patrick Fisher distance, and the divergence. Both filter and wrapper methods may employ efficient search strategies such as branch and bound, best individual N method, sequential forward selection (SFS), sequential backward selection (SBS), and sequential floating forward search (SFFS).

A novel filter feature selection method based on mutual correlation is proposed. Both filter and wrapper techniques have their advantages as well as drawbacks. The major problem with wrapper methods and filter methods employing search strategies is their high-computational complexity, when applied to large data sets. For feature sets of large dimensionality, any feature selection method that would approximate an exhaustive search in these large data spaces is infeasible due to the many possible combinations

$$\frac{d!}{(d-D)!\,D!} \ .$$

On the other hand, any non-exhaustive search method is not guaranteed to find the optimal feature set. We can only hope to reach a reasonable local optimum. While the literature has shown no clear superiority of any particular feature selection method, some feature selection methods are more suitable for large-dimension applications than others.

2 Correlation-Based Method

Correlation is a well-known similarity measure between two random variables. If two random variables are linearly dependent, then their correlation coefficient is ± 1. If the variables are uncorrelated, the correlation coefficient is 0. The correlation coefficient is invariant to scaling and translation. Hence two features with different variances may have the same value of this measure. Let us have n d-dimensional feature vectors

$$X_i = [{}^i x_1, \ldots, {}^i x_d] \qquad i = 1, \ldots, n$$

from K possible classes. The mutual correlation for a feature pair x_i and x_j is defined as

$$r_{x_i,x_j} = \frac{\sum_k {}^k x_i \, {}^k x_j - n\bar{x}_i\bar{x}_j}{\sqrt{(\sum_k {}^k x_i^2 - n\bar{x}_i^2)(\sum_k {}^k x_j^2 - n\bar{x}_j^2)}} \tag{1}$$

If two features x_i and x_j are independent then they are also uncorrelated, i.e. $r_{x_i,x_j} = 0$. Let us evaluate all mutual correlations for all feature pairs and compute the average absolute mutual correlation of a feature over δ features

$$r_{j,\delta} = \frac{1}{\delta} \sum_{i=1, i\neq j}^{\delta} |r_{x_i,x_j}| \ . \tag{2}$$

The feature which has the largest average mutual correlation

$$\alpha = \arg\max_{j} \; r_{j,\delta} \tag{3}$$

will be removed at each iteration step of the feature selection algorithm. When feature x_α is removed from the feature set, it is also discarded from the remaining average correlations, i.e.

$$r_{j,\delta-1} = \frac{\delta\, r_{j,\delta} - |r_{x_\alpha,x_j}|}{\delta - 1} \;\; . \tag{4}$$

2.1 Proposed Feature Selection Algorithm

The proposed correlation based feature selection algorithm can be summarized as follows.

1. Initialize $\delta = d - 1$.
2. Discard feature x_α for α determined by (3).
3. Decrement $\delta = \delta - 1$, if $\delta < D$ return the resulting D dimensional feature set and stop. Otherwise,
4. Recalculate the average correlations by using (4).
5. Go to step 2.

The algorithm produces the optimal D-dimensional subset from the original measurements with respect to the correlation criterion

$$X = [x_1, \ldots, x_D] \;\; .$$

The algorithm is very simple and so it has low computational complexity.

3 Evaluation Criteria

The presented method was compared with three wrapper based alternatives: SFS [9], SFFS [9], and oscillating search (OS) [10] used to directly optimize the Bayes error when each class probability density function is modeled by a single Gaussian. We also compared it with the Bayes error committed by two filter methods that select optimal feature subsets either with respect to the Bhattacharyya distance

$$B = \frac{1}{8}(\mu_i - \mu_j)^T \left(\frac{\Sigma_i + \Sigma_j}{2}\right)^{-1} (\mu_i - \mu_j) + \frac{1}{2}\ln \frac{|\frac{\Sigma_i+\Sigma_j}{2}|}{\sqrt{|\Sigma_i||\Sigma_j|}}, \tag{5}$$

or the divergence (assuming normality)

$$\begin{aligned} DIV = \; & (P_i - P_j)\ln \frac{P_i|\Sigma_j|^{\frac{1}{2}}}{P_j|\Sigma_i|^{\frac{1}{2}}} + \frac{1}{2}\, tr\{[P_i\Sigma_i + P_j\Sigma_j][\Sigma_j^{-1} - \Sigma_i^{-1}]\} + \\ & \frac{1}{2}(\mu_i - \mu_j)^T \left(P_i\Sigma_j^{-1} + P_j\Sigma_i^{-1}\right)(\mu_i - \mu_j) \;\; , \end{aligned} \tag{6}$$

where Σ_i and μ_i are the class covariance matrices and mean vectors, respectively and P_i are prior class probabilities. The criterion functions (5) and (6) are extended for multi-class problems by summing the criterion values for all combinations of 2 out of K classes.

4 Experimental Results

4.1 UCI datasets

In this section, we demonstrate results computed on 2-class datasets from the UCI repository [8] namely the SPEECH data originating from British Telecom (15 features, 682 utterances of the word "yes" and another 736 utterances of the word "no") and the mammogram Wisconsin Diagnostic Breast Center (WDBC) data (30 features, 357 benign and 212 malignant samples). The parameters of the two datasets are summarized in Table 1.

Table 1. UCI repository set parameters.

Parameter	SPEECH	WDBC
K	2	2
D	15	30
n_1	682	357
n_2	736	212
n	1418	569

The progress of the algorithm at the several iterations of the proposed algorithm is illustrated in Table 2.

Although the proposed method selects less optimal feature subsets on average for specific numbers of retained features, as can be seen from Tables 3 and 4, the corresponding Bayes error increases up to 7%. The latter deterioration in accuracy is compensated by the speed of the method.

4.2 Emotional speech data collections

In this section, the Bayes error committed by the subset of features determined with respect to the mutual correlation is compared to that of filter methods employing B or DIV and wrapper methods employing SFS, and SFFS on 2 emotional speech data collections. The first data collection is Danish Emotion Speech (DES) containing recordings of speech utterances expressed by 4 actors in 5 emotional states [13]. The second data collection uses a subset of Speech Under Simulated and Actual Stress (SUSAS) data collection which includes words uttered under low and high stress conditions as well as speech in various talking

Table 2. Recalculated average correlation at the several iterations of the proposed algorithm for the SPEECH dataset.

step	class 1	class 2
1	$r_{6,15} = 0.59$	$r_{7,15} = 0.54$
2	$r_{7,14} = 0.57$	$r_{10,14} = 0.51$
3	$r_{4,13} = 0.54$	$r_{11,13} = 0.48$
4	$r_{9,12} = 0.51$	$r_{4,12} = 0.47$
5	$r_{3,11} = 0.50$	$r_{3,11} = 0.44$
6	$r_{11,10} = 0.49$	$r_{8,10} = 0.43$
7	$r_{5,9} = 0.46$	$r_{12,9} = 0.41$
8	$r_{10,8} = 0.44$	$r_{14,8} = 0.39$
9	$r_{15,7} = 0.44$	$r_{1,7} = 0.38$
10	$r_{1,6} = 0.39$	$r_{6,6} = 0.37$
11	$r_{8,5} = 0.37$	$r_{15,5} = 0.34$
12	$r_{13,4} = 0.32$	$r_{5,4} = 0.31$
13	$r_{2,3} = 0.30$	$r_{9,3} = 0.24$
13	$r_{12,2} = 0.25$	$r_{2,2} = 0.21$
14	$r_{14,1} = 0.16$	$r_{13,1} = 0.13$

styles expressed by 9 native American English speakers [14, 15]. Several statistics of *pitch, formants*, and *energy* contours were extracted as features [16]. In Table 5, the parameters of DES and SUSAS are summarized. For DES, $n_k = 72$, $k = 1, 2, \ldots, 5$, while for SUSAS $n_k = 630$, $k = 1, 2, \ldots, 8$.

The feature selection methods are evaluated according to their execution time and the classification error achieved by the Bayes classifier that classifies the speech segments into emotional states. The crossvalidation method was used to obtain an unbiased error estimate [17]. For wrapper techniques based on SFS and SFFS, the crossvalidation method has been speeded up by two mechanisms that reduce its computational burden and improve its accuracy [16]. In the experiments, feature set A is declared to be better than feature set B, if the error achieved by using A is smaller than that obtained using B by at least 0.015. The error difference 0.015 was chosen according to observations made in [16] and the available computational power.

A comparison of the execution time needed by each feature selection method is made in Table 6 for each data collection. Filter methods such as those employing correlation, B, and DIV are 50 times faster than wrapper ones based on SFS and SFFS. The execution time for correlation and DIV is comparable, whereas the filter method based on B is twice slower.

To evaluate the efficiency of the proposed filter method based on correlation, we compare the classification errors measured on DES and SUSAS. The classification errors on DES are plotted in Figure 1 for the number of retained features (SFS,SFFS) and the number of discarded features (correlation, B, DIV). It is seen that SFS and SFFS achieve about 48% classification error, whereas the error for filter methods is about 10% higher. The lowest error rates achieved by wrap-

Table 3. Bayes error for different feature selection algorithms on SPEECH dataset.

Number of retained features	Correlation	SFS	OS	B	DIV
14	0.077	0.074	0.074	0.081	0.081
13	0.082	0.068	0.066	0.076	0.073
12	0.092	0.069	0.062	0.076	0.076
11	0.089	0.066	0.060	0.072	0.077
10	0.084	0.060	0.056	0.079	0.089
9	0.115	0.061	0.058	0.074	0.087
8	0.113	0.055	0.050	0.074	0.098
7	0.108	0.052	0.052	0.087	0.102
6	0.092	0.053	0.053	0.086	0.118
5	0.113	0.053	0.052	0.076	0.108
4	0.118	0.068	0.061	0.079	0.098
3	0.108	0.081	0.081	0.111	0.111
2	0.119	0.119	0.119	0.187	0.226
1	0.345	0.139	0.139	0.221	0.221
average	0.118	0.073	0.070	0.099	0.112

pers are for 10-15 retained features. Similarly, the lowest error rates obtained by filter methods are accomplished when 60-70 features are removed from the entire feature set. From the error rates of the Bayes classifier plotted in Figure 1, we infer that correlation method is equivalent to the other filter methods but it is clearly inferior to wrapper methods.

From the experimental results on data collection SUSAS plotted in Figure 2, it is inferred that the lowest error rates are achieved when almost all the features are selected, either in the first steps of filters or the last steps of wrappers. So, feature selection here is not used to reduce error rates but to remove redundant features. The optimal feature set for wrappers as well for filters is achieved after 20-30 iterations. Wrappers select 20-30 features, whereas filters remove 20-30 features out of the 90 initial ones. Therefore, wrappers yield a smaller feature set than filters. Regarding the time requirements, wrappers select the optimal feature subset of 20 features within 2000 sec., whereas filters based on correlation and divergence can yield a subset of 50 features yielding comparable error rates to wrappers within 150 sec. There is a great difference between the results obtained for DES and SUSAS. By using all features in DES for classification, the error is at random level, whereas the error rates in SUSAS are minimized when the entire feature set is employed. This abnormal behavior of classification error regarding the size of feature set could be a topic of further research.

Table 4. Bayes error for different feature selection algorithms on WDBC dataset.

Number of retained features	Correlation	SFS	OS	B	DIV
30	0.053	0.059	0.084	0.079	0.089
29	0.053	0.052	0.053	0.056	0.053
28	0.053	0.049	0.042	0.053	0.049
27	0.056	0.049	0.032	0.046	0.042
26	0.056	0.053	0.028	0.049	0.049
25	0.053	0.053	0.025	0.046	0.063
24	0.060	0.053	0.021	0.046	0.049
23	0.056	0.046	0.018	0.056	0.060
22	0.067	0.039	0.018	0.053	0.067
21	0.063	0.032	0.014	0.046	0.063
20	0.056	0.028	0.018	0.042	0.067
19	0.056	0.021	0.018	0.039	0.056
18	0.053	0.018	0.011	0.039	0.056
17	0.074	0.014	0.014	0.035	0.053
16	0.056	0.014	0.014	0.042	0.046
15	0.077	0.011	0.011	0.053	0.046
14	0.088	0.014	0.011	0.035	0.056
13	0.074	0.011	0.011	0.039	0.053
12	0.077	0.011	0.014	0.053	0.046
11	0.070	0.011	0.007	0.046	0.053
10	0.074	0.018	0.007	0.053	0.046
9	0.063	0.018	0.004	0.053	0.060
8	0.102	0.018	0.007	0.053	0.062
7	0.105	0.018	0.007	0.053	0.042
6	0.109	0.025	0.011	0.063	0.063
5	0.250	0.028	0.021	0.056	0.053
4	0.253	0.042	0.032	0.077	0.077
3	0.274	0.046	0.042	0.067	0.067
2	0.372	0.049	0.056	0.077	0.077
1	0.345	0.084	0.084	0.109	0.105
average	0.098	0.032	0.025	0.054	0.059

5 Conclusions

A filter method for feature selection based on mutual correlation has been proposed. Being a filter method, it yields features independent of the classifier to be used. Hence, in principle, the proposed method can only approach the feature selection quality of methods based on direct estimation of the Bayes classifier error rate (i.e. wrapper methods with SFS or OS, filter methods using B or DIV). At the same time, the proposed filter method can easily cope with classification tasks in feature spaces of large dimensionality. The method is extremely

Table 5. Parameters of emotional speech data collections.

Parameter	DES	SUSAS
K	5	8
D	90	90
n_k	72	630
n	360	5040

Table 6. Execution time (in sec).

Method	Databases	
	DES	SUSAS
SFFS	18107	53494
SFS	9446	21092
correlation	276	458
B	351	633
DIV	292	454

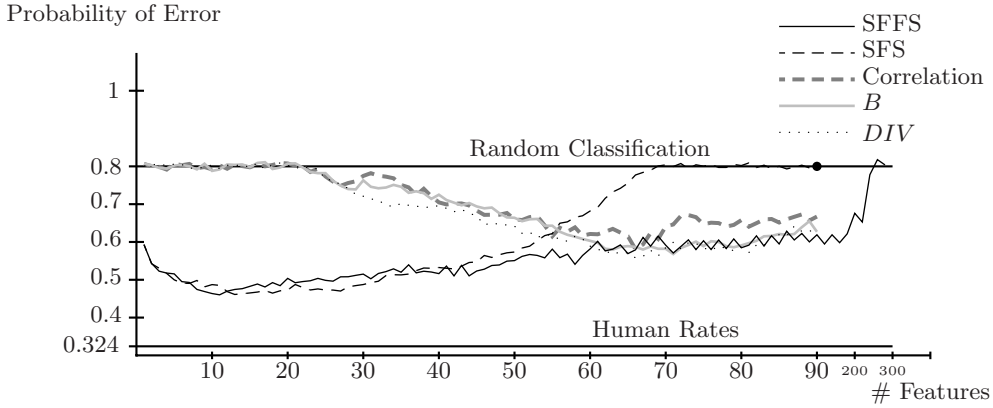

Fig. 1. Probability of classification error versus the number of features retained/discarded by feature selection method on DES.

fast in comparison with the other compared methods (except DIV). The presented method can also be used when alternative filter methods based on B or DIV cannot be applied due to limited measurements which prevent the robust estimation of necessary covariance matrices. The method can be used either in supervised or unsupervised mode.

Acknowledgments

This research was supported by the EC project no. FP6-507752 MUSCLE, grants No.A2075302, 1ET400750407 of the Grant Agency of the Academy of Sciences CR and partially by the MŠMT grant 1M0572 DAR.

References

1. Devijver PA, Kittler J Pattern Recognition: A Statistical Approach, Prentice-Hall, (1982)
2. Duda RO, Hart PE, Stork DG Pattern Classification, 2nd Ed., Wiley-Interscience, (2000)
3. Ferri FJ, Pudil P, Hatef M, Kittler J Comparative Study of Techniques for Large-Scale Feature Selection, Gelsema ES, Kanal LN (eds.) Pattern Recognition in Practice IV, Elsevier Science B.V., (1994) 403–413

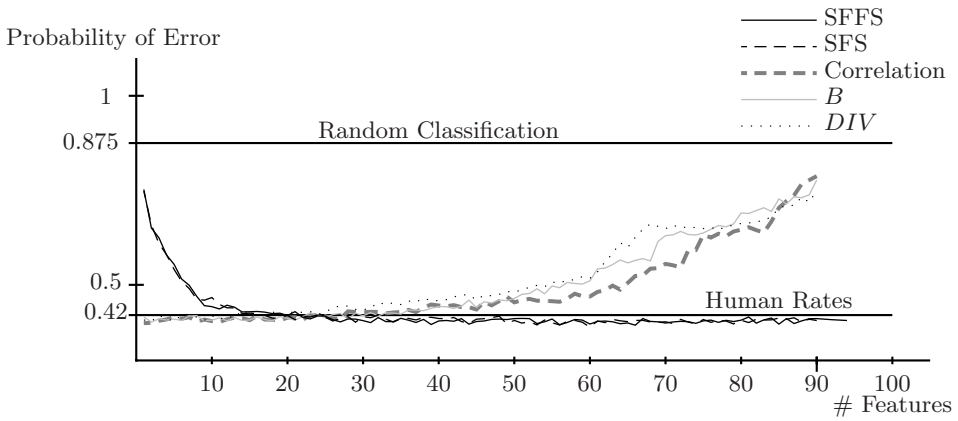

Fig. 2. Probability of classification error versus the number of features retained/discarded by feature selection method on SUSAS.

4. Fukunaga K Introduction to Statistical Pattern Recognition, Academic Press, (1990)
5. Jain AK, Zongker D Feature Selection: Evaluation, Application and Small Sample Performance, IEEE Transactions on Pattern Analysis and Machine Intelligence 19(**2**): (1997) 153–158
6. Kohavi R, John GH Wrappers for Feature Subset Selection. Artificial Intelligence 97(**1-2**): (1997) 273–324
7. Kudo M, Sklansky J Comparison of Algorithms that Select Features for Pattern Classifiers, Pattern Recognition 33(**1**): (2000) 25–41
8. Murphy PM, Aha DW UCI Repository of Machine Learning Databases [ftp.ics.uci.edu]. Univ. of California, Dept. of Information and Computer Science, Irvine, CA, (1994)
9. Somol P, Pudil P Feature Selection Toolbox. Pattern Recognition 35(**12**): (2002) 2749–2759
10. Somol P, Pudil P Oscillating Search Algorithms For Feature Selection, In: Proc 15th IAPR International Conference on Pattern Recognition, Barcelona, Spain, (2000) 406–409
11. Theodoridis S, Koutroumbas K Pattern Recognition, 2nd Ed., Academic Press, (2003)
12. Webb A Statistical Pattern Recognition, 2nd Ed., John Wiley & Sons, (2002)
13. Engberg IS, Hansen AV Documentation of the Danish Emotional Speech Database (DES), Techn. Report, Center for Person Kommunikation, Aalborg Univ., (1996)
14. Womack BD, Hansen JHL N-Channel Hidden Markov Models for combined stressed speech classification and recognition, IEEE Trans. Speech and Audio Processing 7 (**6**): (1999) 668–667
15. Bolia RS, Slyh RE Perception of stress and speaking style for selected elements of the (SUSAS) database, Speech Communication (**40**): (2003) 493–501
16. Ververidis D, Kotropoulos C Sequential forward feature selection with low computational cost, In: Proc 13th European Signal Processing Conf., Antalya, Turkey, (2005)
17. Efron B, Tibshirani RJ An Introduction to the Bootstrap, Chapman & Hall/CRC, (1993)