Benchmarking Feature Selection for E-Noses

M. Pardo[1], G. Sberveglieri[1], J.W. Gardner[2]
[1]INFM-CNR & University of Brescia, Italy, pardo@ing.unibs.it
[2]Warwick University, UK

Abstract

We present the results of an extensive feature selection benchmarking study, which includes the testing of different search strategies and different feature subsets scoring criteria (the test set error of different classifiers) on four datasets (three electronic nose, one GC dataset) posing diverse computational challenges and spanning diverse difficulty levels. Notable findings are: 1) floating searcg methods consistently give higher performances at the expense of greater computational times, 2) the non-parametric kNN classifiers, while more prone to the curse of dimensionality than linear and quadratic classifiers, normally reaches better performances with few features.

1. Introduction

Selecting good sensor subsets can help both in the understanding of the sensors themselves and in enhancing the data analysis (e.g. classification performance) through a more stable data representation. In statistical pattern recognition the phenomenon of the "curse of dimensionality" has been often observed, where the sparseness of the data in high dimensional spaces causes a bad classification performance. At the same time, two trends in sensor systems are the increase of the sensor numbers and the extraction of complex patterns from each sensor (e.g. dynamical features).

Feature selection (FS) needs two ingredients: a criterion for judging a feature subset (classical ones are the ratio between between-class class distances divided by within-class scatter or the test set error for a given classifier) and a search strategy through which successive subsets are examined.

2. Methods

The study includes:

1. The comparison (benchmarking) of different search strategies: i) exhaustive search, ii) branch and bound, iii) forward selection, iv) backward selection and v) floating search methods [2,3].
2. The benchmarking of different wrappers based on the test set error: i) kNN (with k=1,3,5), ii) linear classifier, iii) quadratic classifier, iv) Fisher linear discriminant
3. The test of the different search strategies and selection criteria on four different datasets posing different computational challenges and spanning diverse difficulty levels.

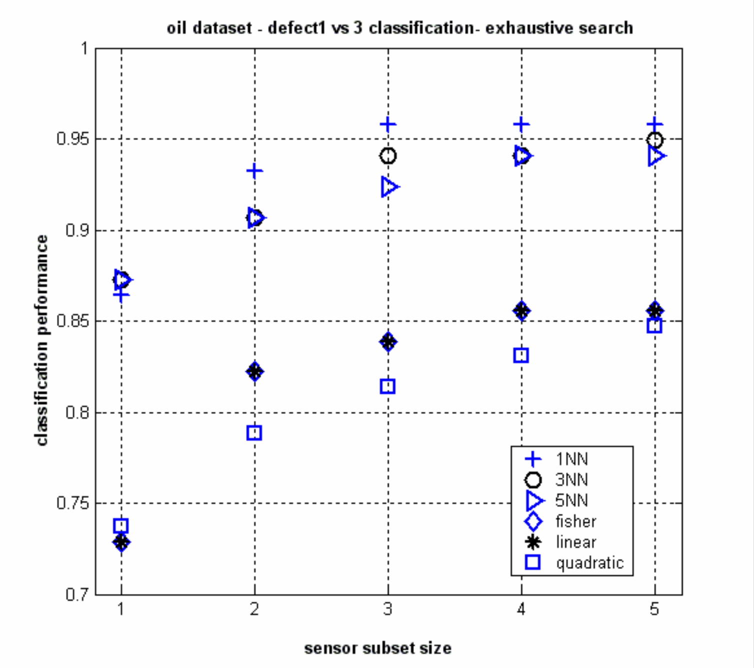

Figure 1 Test set performance of different classifiers for the best subsets composed of 2, 3, 4, 5 sensors for the discrimination between two oil defects

3. Results

The comparison between the best results of the different classifiers, for the discrimination between two oil defects, is displayed in **Figure 1**. We note that: 1) Performance increases, or at least doesn't decrease, with subsets size: the curse of dimensionality hasn't set in yet; 2) the classifiers perform in the order ('>' means 'is better than') 1NN>3NN>5NN>fisher=linear>quadratic: the

more flexible the classifier, the better. Note that, as known from the theory, linear and Fisher classifiers implement the same decision rule for two class problem and therefore have the same performance; 3) parametric (i.e. Fisher, linear, quadratic) methods give worse results: they are probably too rigid for this data, i.e. they introduce too much bias.

The results given by suboptimal search methods are shown in

Figure 2:

- The classifiers are ranked:1NN>3NN>5NN> fisher = linear > quadratic. This is the same order found for the exhaustive search in **Figure 1**.
- Forward and floating search have comparable performances and are better than backward search.
- There is a 8% - 16% increase in classification performance by suboptimally selecting the best features with respect to using all the features.
- A single selected feature is better than all features for kNN: the non-parametric kNN is prone to the curse of dimensionality. This is less the case for the parametric classifiers.
- By comparing **Figure 1** with
- Figure 2 we see that performances are similar between exhaustive search over 5 sensors and the best suboptimal search methods. Yet, search times are much shorter for suboptimal methods (minutes vs. hours). Therefore suboptimal methods are advantageous.

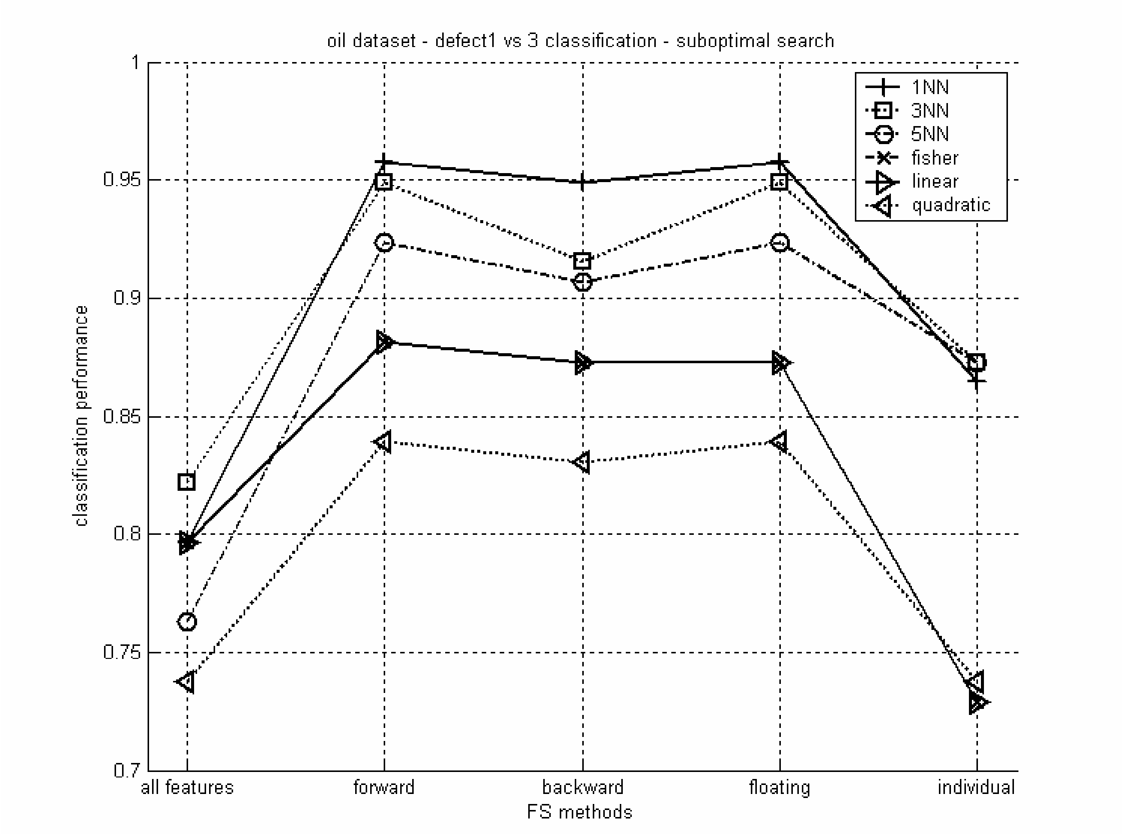

Figure 2 Test set performance of different classifiers for the best subsets –obtained with different suboptimal methods- for the discrimination between two oil defects.

4.Conclusions

Feature selection is proably *the* most important data analysis topic, as far as data interpretation is concerned, and has been therefore widely investigated in every application branch. For electronic noses, FS permits to find the best sensors and the best features extracted from the response curve and can hence guide electronic nose development. This paper assessed a variety of FS options and gave indications on the best of them.

Acknowledgments

Study partially performed during a visit of M. Pardo to the Warwick University thanks to the NOSE II exchange program. This work has also been partly supported by the European Commission in the project NMP4-CT-2003-001528 “Nanostructured sold-state gas sensors with superior performance” (NANOS4).

References

1 M. Pardo, L.G. Kwong, G. Sberveglieri, K. Brubaker, J. F. Schneider, W.R. Penrose, J.R. Stetter. Data Analysis for a Hybrid Sensor Array; accepted for Sensors and Actuators B (Gopel prize at ISOEN 03 in Riga)

2 P. Pudil, J. Novovicov´a, and J. Kittler. Floating search methods in feature selection. Pattern Recognition Letters, 15:1119–1125, November 1994

3 A. Jain and D. Zongker. Feature selection: evaluation, application, and small sample performance. IEEE Trans. On Pattern Analysis and Machine Intelligence, 19(2):153–158, 1997