DOI 10.5162/IMCS2012/P1.0.21

Optimal Feature Selection for Classifying a Large Set of Chemicals Using Metal Oxide Sensors

Thomas Nowotny[1], Amalia Z. Berna[2], Russell Binions[3], Stephen Trowell[2]
[1] School of Engineering and Informatics, University of Sussex, Falmer, Brighton BN1 9QJ, UK
[2] CSIRO Food Futures Flagship and Ecosystem Sciences Division, GPO Box 1700 Canberra, ACT 2601, Australia
[3] School of Engineering and Materials Science, Queen Mary University of London, London E1 4NS, UK

Abstract:
We investigated the feature selection problem for the application of all-against-all classification of a set of 20 chemicals using metal oxide sensors and linear support vector machines. We defined a set of possible features in terms of identity of sensors and sampling times and tested all possible combinations of such features. We found that performance is clearly increased by feature selection compared to previous results [1] but that, contradictory to naïve expectation, using the maximal number of different sensors and all available data points for each sensor does not necessarily yield the best results. Similarly, the standard method of taking one data point from all sensors also underperforms.

Key words: Feature selection, metal oxide sensors, classification, support vector machines, electronic nose

Introduction

For any chemical sensing application, an important choice to make is which types of sensors, and, if using an array of sensors, how many sensors of any particular type to use. Further choices apply to how to sample data from the sensors and how to pre-process the collected raw data. It is well-known in machine learning that this process of feature selection (FS) is very important for the eventual success of the overall classification (recognition) system.

Intuitively, we would expect that using more sensors can only improve performance, as long as the sensors are not fully redundant (identical) or fully uncorrelated with the problem (equivalent to noise). Here, we systematically investigate this feature selection problem for fully classifying a set of 20 chemicals (Tab. 1) using metal oxide sensors (MOS) [2] and linear support vector machines (SVMs) [3] in a "wrapper" approach.

The set of analytes consists of five chemicals each from four chemical groups: alcohols, aldehydes, esters and ketones. They were chosen from a larger set of chemicals used in a comparative study between metal oxide sensors and biological sensors [4]. To detect the individual chemicals we used a 12-sensor array, comprising six standard, doped tin dioxide (SNO_2), five novel zeolite-coated, doped chromium titanium oxide (CTO) sensors [5] and one non coated CTO sensor. Every subset of the 12 sensors was sampled at six representative time points (Fig. 1) which comprised the set of all possible features. Feature selection thus consisted of choosing a subset of sensors and data points, each choice constituting a particular "feature set".

We tested the performance of all possible 257,985 such feature sets in all-against-all classification of the data set using a standard linear SVM algorithm [6] in 10-fold cross-validation.

Tab. 1: Analytes used and their chemical classes

Alcohols	1-Pentanol	Aldehydes	Acetaldehyde
	1-Hexanol		Butanal
	Z2-hexen-1-ol		Hexanal
	1-Octen-3-ol		E2-hexenal
	3-Methylbutanol		Furfural
Esters	Ethylhexanoate	Ketones	Acetone
	Ethylacetate		2-butanone
	Isopentylacetate		2-pentanone
	Methylacetate		2-heptanone
	Ethylbutyrate		2,3-butanedione

DOI 10.5162/IMCS2012/P1.0.21

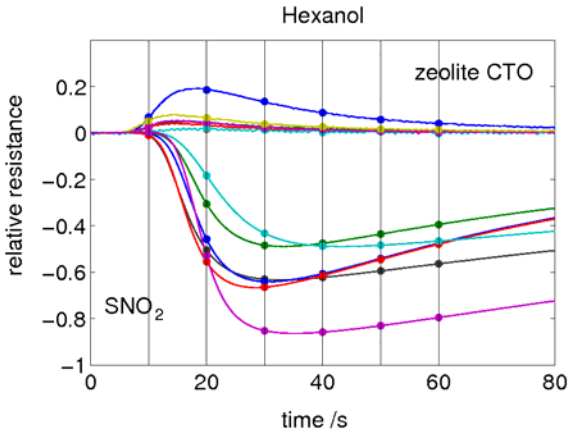

Fig. 1: Example of response from the FOX Enose fitted with the twelve-sensor array. Vertical lines mark the available sampling times.

Results

Figure 2 shows the performance of feature sets of different size constraint, e.g. 2,3 is 2 data points each from 3 sensors. We note that the best performing feature sets do much better than previously reported classifiers based on this sensor array [1] and that the best performance is not achieved with the naïvely expected maximal sensor- and data-use (12 sensors, 6 data points, top line of Fig. 2).

To control for selection biases, we chose a group of well-performing sensors (column "top10") and repeated cross-validation for this group ("top10 rerun"). The superior performance of smaller feature sets over the (6,12) choice remains intact. Note, however, that for smaller feature sets, classification success depends critically on the choice of the used feature sets from the pool of all potential sets of a given size. This is illustrated by the much lower worst performance (see column "worst" and the low outliers in the left column).

The data is presented ordered by the number of possible feature sets for each given size constraint, ranging from a single (6,12) feature set on the top to 18480 possible (3,6) feature sets at the bottom. The prevalence of excellent "best", "top 10" and "top 10 rerun" performance at the bottom of the graph illustrates that size constraints with many different feature set choices are more likely to have well-performing sets, even though this is not an absolute rule, see e.g. the 2970 (1,4) feature sets that are much less successful than the 20 (3,12) sets.

In order to identify possible explanations for the improved performance for some feature sets over others we compared the classification performance for each set with the quality of clustering given this feature choice (Fig. 3). For this purpose we defined the quality of clustering as the quotient d_{inter}/d_{intra} of the average Euclidean distance between average class vectors,

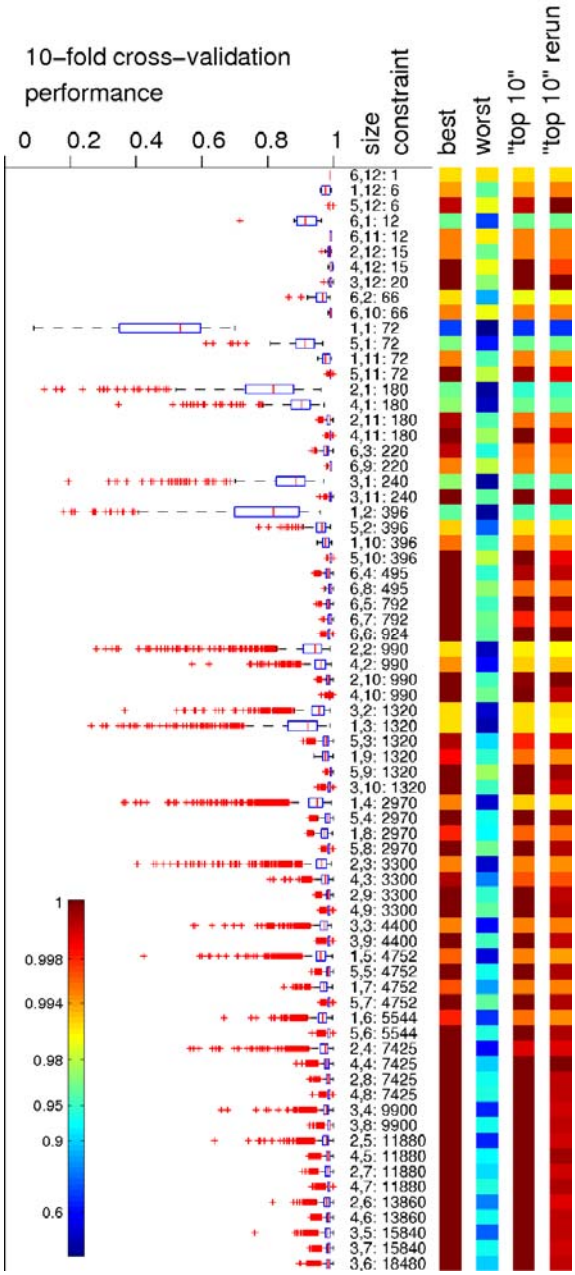

Fig. 2: Fractional prediction accuracy of 10-fold cross-validation using linear SVMs for all possible sub-samplings of 6 time points and the 12 available sensors. The box plots show the median and 25% and 75% quantiles, the estimated overall range (whiskers) and identified outliers (red crosses). The numerical columns give the number of time points and the number of sensors used followed by the resulting number of such combinations that was tested. The colored columns give the best observed performance, worst performance, performance of the "top 10" group of feature choices and the performance of this group in a repeated 10-fold cross-validation. Note the highly non-linear color code.

$$d_{\text{inter}} = \left\langle \left\| \langle \vec{x} \rangle_i - \langle \vec{x} \rangle_j \right\|_2 \right\rangle_{i,j} \quad (1)$$

and the average Euclidean distance of vectors within a class to the average class vector,

$$d_{\text{intra}} = \left\langle \left\| \vec{x}_{i,k} - \langle \vec{x} \rangle_i \right\|_2 \right\rangle_k . \quad (2)$$

Here, $\vec{x}_{i,k}$ denotes the *k*th measurement of chemical (class) *i*, $\langle . \rangle_i$ denotes taking the average over the index *i* and $\|.\|_2$ denotes the Euclidean norm. While the data in Fig. 3 shows a noticeable positive correlation, particularly for the extremes of very low clustering quality and very low performance (lower left), the clustering quality apparently does not fully explain the classification results. The overall correlation coefficient between classification performance and clustering quality as defined here is positive but only 0.205.

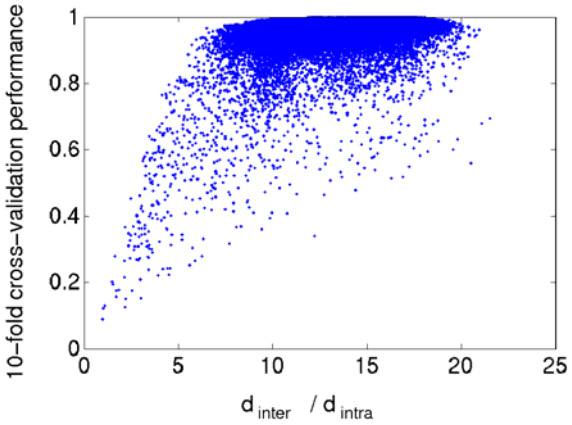

Fig. 3: Classification performance of linear SVMs for all possible feature choices in 10-fold cross-validation plotted against the clustering quality (ratio of inter-class to average intra-class Euclidean distance, see main text). A clear correlation is noticeable, in particular the absence of points with low clustering quality and high performance (upper left corner) or low performance and high clustering quality (lower right corner).

Discussion

The results illustrated in Fig. 2 suggest that it may be beneficial to design a sensor array specifically for each envisioned application domain and if doing so, that a few well-chosen sensors and data sampling times may outperform using the maximal array and many data points. However, it is worth noting that choosing the correct sensors and data sampling times is critical. For example, the median performance of (3,6) feature set (0.985) is actually worse than the performance of the single (6,12) choice. This implies that taking just an arbitrary (3,6) feature set would likely not improve the overall success.

We notice that a large number of classification results are almost optimal and some of the differences we base our conclusions on amount to discrepancies of a single error in classifying 200 measurements of chemicals. This indicates that the array we used is capable of this quite challenging classification problem. Future work will need to extend the results to even more challenging applications including lower or multiple concentrations, and measurements taken over an extended period of time.

As pointed out above, intuitively we would have expected that classification performance can only increase when additional data (information) is added. In the worst case one would have expected unchanged performance if the additional data was not useful. Here, however we saw that adding additional data can decrease the accuracy of classification. The likely explanation of this phenomenon is over-fitting. The additional data may provide additional information for the training data, but this can lead to too specific classifiers that may not generalize as well to new testing data as "less informed" ones. This trade-off between optimal classification on the training data and optimal ability to generalize to new test data, the so-called over-fitting problem, is a classic topic in machine learning. Future work will focus on unraveling what the optimal solutions are for given practical problems and the degree to which these are generalisable within or beyond a problem set.

The work reported here was conducted with a specific classification method, i.e. a linear support vector machine. One could argue that the observed phenomenon of better classification with smaller feature sets may be specific to this particular method. While we cannot fully exclude this possibility, the effects of over-fitting are known to affect all approaches to classification. While the details may differ for other classification methods, the principal results are likely to apply to a variety of such methods.

Conclusion

We set out to systematically assess the question of feature selection for arrays of MO sensors in a classification task, using standard machine learning methods. We found that feature selection can improve classification performance and that the best-performing feature sets are not necessarily the naively expected ones.

In future work we plan to analyze in depth why particular combinations of sensors are very successful and whether this translates to classification methods other than linear support vector machines.

Acknowledgements

This work was partially supported by an OCE Distinguished Scientist Award of CSIRO to TN.

References

[1] A. Z. Berna, et al., Evaluating Zeolite-Modified Sensors: Towards a Faster Set of Chemical Sensors, In AIP Conference Proceedings 1362, 50-52 (2011) ; doi: 10.1063/1.3626302

[2] A. Berna, Metal Oxide Sensors for Electronic Noses and Their Application to Food Analysis, Sensors 10, 3882-3910 (2010); doi:10.3390/s100403882

[3] C. Cortes, V. Vapnik, Support-Vector Networks, Machine Learning. 20(3), 273-297 (1995); doi: 10.1007/BF00994018

[4] A.Z. Berna, A. R. Anderson, S. C. Trowell, Bio-benchmarking of electronic nose sensors, PlosONE 4, e5406 (2009)

[5] R. Binions, et al., Zeolite-Modified Discriminating Gas Sensors, Journal of . Electrochemical Society. 156(3), J46-J51 (2009); doi: 10.1149/1.3065436

[6] C.-C. Chang. C. –J. Lin, LIBSVM: a library for support vector machines (2001); Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm