Proceedings of the 14th International Symposium on Olfaction and Electronic Nose, New York City, NY, USA, 2-5 May 2011, Pages 9-10. Editor: Perena Gouma.

Experiences in Pattern Recognition for Machine Olfaction

C. Bessant

Bioinformatics Group, Cranfield University, Bedfordshire MK43 0AL, UK.

Abstract. Pattern recognition is essential for translating complex olfactory sensor responses into simple outputs that are relevant to users. Many approaches to pattern recognition have been applied in this field, including multivariate statistics (e.g. discriminant analysis), artificial neural networks (ANNs) and support vector machines (SVMs). Reviewing our experience of using these techniques with many different sensor systems reveals some useful insights. Most importantly, it is clear beyond any doubt that the quantity and selection of samples used to train and test a pattern recognition system are by far the most important factors in ensuring it performs as accurately and reliably as possible. Here we present evidence for this assertion and make suggestions for best practice based on these findings.

Keywords: Chemometrics, validation, design of experiments.
PACS: 07.05.Kf, 07.05.Fb, 07.07.Df.

INTRODUCING PATTERN RECONGNITION TECHNOLOGIES

Most of the pattern recognition systems developed for olfactory sensor systems are based on multivariate classification, where multiple measured variables are reduced to a single value representing the class membership of the sample under study. ANNs can theoretically provide the optimal solution linking measured variables to the desired outputs, but we have seen that the much simpler method of partial least squares discriminant analysis (PLS-DA) can yield better results in practice because it is easier to optimize [1]. More recently, SVMs have become our technology of choice as they generally produce results as good as or better than PLS-DA and ANNs while being relatively easy to optimize [2]. However, our experience strongly suggests that the specific classification technology used has a much lesser effect than other decisions that are made during the creation of a pattern recognition system.

SELECTION OF DATA FOR TRAINING AND TESTING

Of primary importance is the selection of data used to train and test the classification models that underpin the recognition system. There are two issues to consider: the number of samples used, and the variation captured by those samples.

The aim of most olfactory pattern recognition studies is to distinguish *cases* from *controls*. In this context, a case is taken to be a sample exhibiting a particular characteristic and is usually easy to define, e.g. it's taken from a patient suffering from

a particular disease, a sample that was spiked, or a piece of food known to be bacterially spoiled. The definition of a control is much less clear – in many published medical studies it means *healthy control*, i.e. a person free of not just the disease in question, but free of disease altogether. We have shown that this can artificially simplify the problem under study [3], presumably because the biochemical differences between the two classes are exaggerated. Furthermore, the use of healthy controls results in limited applicability to the real world as the classifier may be picking up generic host responses to disease instead of disease-specific markers.

In terms of determining the optimum number of samples, it is not possible to use a traditional power calculation when dealing with the multivariate data typical of olfaction. Our solution is to determine the optimum number of samples empirically by building classifiers using increasing amounts of data, ideally while the data is being collected. When the performance of the classifier reaches a stable plateau with respect to the number of training samples it can be assumed that no further data is needed.

EVALUATION OF MODEL PERFORMANCE

In our experience, the particular way in which the behavior of a classification model is evaluated (often referred to as *model validation*) is the biggest factor influencing the perceived performance of the model. Well established methods such as plotting PLS scores, and traditional cross validation, have been discredited in recent years [4]. More representative performance metrics are based on permutation tests, whereby hundreds of different models are produced from random permutations of available data. Statistical analysis of the results from these models yields information about the actual classification accuracy that can be achieved, the robustness of approach, and how significant the result is compared to random chance.

This use of multiple models creates a dilemma as to which model to ultimately deploy to identify new samples, but this can be turned to our advantage by using a consensus classification from multiple models, with the level of agreement between different models being used as a confidence measure.

ACKNOWLEDGMENTS

The author would like to thank the many colleagues whose work has contributed to the findings reported, and is grateful for financial support from ISOEN 2011, the European Commission (SYMBIOSIS-EU FP7 project) and Amerderm Research Trust.

REFERENCES

1. E. Z. Panagou, F. R. Mohareb, A. A. Argyri, C. Bessant, G. J. E. Nychas, *Food Microbiology* **in press**. doi:10.1016/j.fm.2010.05.014.
2. R. G. Brereton and G. R. Lloyd, *Analyst* **130**, 230-267 (2010).
3. C. M. Weber, M. Cauchi, M. Patel, C. Bessant, C. Turner, L. E. Britton and C. M. Willis, *Analyst* **136**, 359-364 (2011).
4. J. A. Westerhuis, H. C. J. Hoefsloot, S. Smit, D. J. Vis, A. K. Smilde, E. J. J. van Velzen, J. P. M. van Duijnhoven and F. A. van Dorsten, *Metabolomics* **4**, 81-89 (2008).