DIMENSIONALITY REDUCTION METHODS OF ELECTRONIC NOSE DATA FOR BACTERIA DISCRIMINATION

Geoffrey C. Green, Adrian D. C. Chan, and Rafik A. Goubran
Department of Systems and Computer Engineering, Carleton University, Ottawa, ON

INTRODUCTION

Electronic nose (EN) technology has emerged in the last decade as a viable means for analyzing and classifying samples based on odour. An EN instrument consists of an array of gas sensors that respond to a sample's odorant molecules. A wide range of sensor materials now exist, including conducting polymers, metal oxides, and quartz crystal microbalance [1]. The array is constructed so that a wide range of compounds will invoke a response from the entire array, with individual sensor elements responding differently for samples from each odour category. The array response forms a unique "smellprint" that can be used to discern samples. Among the important advantages of EN systems are reduced sample preparation effort, decreased processing time, and simplicity of operation. As a result, they have been deployed for tasks such as process monitoring and quality control in several industries including pharmaceutical, food, and packaging [2].

The ability to process *biological* samples with EN has garnered interest of late – potential uses include food safety (detection of bacterial contaminants) and clinical diagnosis (pathogen identification). In both of these applications, EN testing will allow more expeditious results compared to traditional laboratory techniques (*e.g.* plating/culturing), thus allowing effective timely responses (*e.g.* issuing recalls and warnings on tainted food, or early patient treatment before a disease progresses). A growing body of research elicits optimism in this regard – the detection of food contaminants and pathogens responsible for several medical conditions (*e.g.* urinary tract infection, bacterial vaginosis) appear to be viable with commercially available EN systems [2,3].

In this paper, we used two EN modalities to investigate the detection and discrimination behaviour when processing bacteria samples of varying concentration provided by the Canadian Food Inspection Agency (CFIA), Ottawa, ON. In the food safety application, high sensitivity is of obvious importance. *E. coli* O157:H7, for instance, can cause infection at a dose of only 10 cells/mL [4]. Testing for the presence and/or type of bacteria in a candidate sample should ideally yield accurate results over a wide range of concentrations.

In EN analysis, feature vectors for each sample are often highly dimensional, causing an exponential increase in volume of the feature space with each added dimension. This makes the number of degrees of freedom large in such cases, in turn causing analysis problems. This *curse of dimensionality* is a significant impediment in machine learning systems [5]. In this work, we employ uncorrelated LDA (ULDA) [6] feature reduction to mitigate this problem in the context of detecting and identifying bacteria of different concentrations, and compare its performance with principal component analysis (PCA). Several pitfalls that must be avoided for effective feature reduction are also discussed.

METHODS

2.1 Data Acquisition & Feature Extraction

The following two commercially available EN instruments were used in this work:

1. metal oxide conductivity sensors (MOS) – AlphaMOS FOX [7], and
2. fingerprint mass spectrometer (FMS) – AlphaMOS Kronos [7].

In order to promote the release of volatile organic compounds (VOCs) into the headspace of the 10mL vial containing the sample (thereby giving a stronger signal), samples were heated to 100^{o}C then agitated at 500rpm for 900s prior to EN injection (we used 1.0mL injection volume for the MOS, 4.0mL for the FMS). Raw data for the MOS consists of a 300s time series for each of 12 sensor response curves. For the FMS, the response intensity of each mass fragment between 45-150amu was recorded for 120s. Feature extraction on the MOS required: 1) a fractional difference calculation (to eliminate baseline drift); 2) selecting the maximum absolute value of the resulting curve; and 3) sensor normalization. On the FMS, the features of interest were the areas under the intensity *vs.* time curves during the time interval for which the intensity was greater half of its maximum value, followed by sensor normalization. These features were calculated as described in our previous work [8].

2.2 Bacteria Samples

The samples used were *E. coli DH5α* and *Listeria innocua* (non-pathogenic bacteria strains) cultured in a nutrient broth in the same manner as described in [8]. Concentrations of 10^8, 10^7, and 10^6 cells/mL were obtained through serial dilution of dense cultures with additional broth. There were 3 runs performed (one for each concentration). During each run, 18 samples (6 samples from each of the three classes: E (*E. coli*), L (*Listeria*), B (broth)) were prepared at the same time (sample aliquot 2.0 mL) and processed by the EN. Samples from each class were presented in alternating order.

2.3 Dimensionality Reduction

PCA – This unsupervised method creates a feature vector in a lower dimensional space with a linear transformation that represents the original in the new space in a least squares sense. Conceptually, PCA creates a new basis from linear combinations of the original dimensions along which the scatter of the data points is greatest. Generally, two or three components are sufficient [5].

LDA and ULDA – Unlike PCA, these are supervised methods (they use the category labels for each sample). Linear transformations are calculated that discriminate the feature vectors between classes in the new space, by maximizing inter-class variation and simultaneously minimizing intra-class variation. In LDA, the criterion function requires a non-singular scatter matrix. ULDA finds application when the scatter matrices of the samples are singular (and thus non-invertible) – a scenario that can arise when there are many more features than samples or when the features are highly correlated. ULDA also ensures that the new features in the transformed space are uncorrelated, and this has the potential to give increased classification rates [5,6].

2.4 Classification and Validation

Scatter plots of the data in the dimension-reduced space, while useful for visualization, are not sufficient to quantify performance. Instead, a classifier is used to assign a category label to a new sample (represented with its feature vector) using a previously trained model. In this paper, we used two types of classifiers: linear and k-nearest neighbours (kNN). In order to validate our classification models and to estimate the system performance, we adopted the commonly used "leave-one-out" (LOO) cross validation technique, with overall classification accuracy averaged over all possible partitions [5].

RESULTS

3.1 MOS EN

As the bacteria concentration decreases, we expect classification to be more difficult because at lower concentrations, the bacteria samples more closely resemble the bacteria-free nutrient broth. This appears to be validated in Figure 1, where we can see better separation of classes at 10^8 cells/mL than at 10^6 cells/mL.

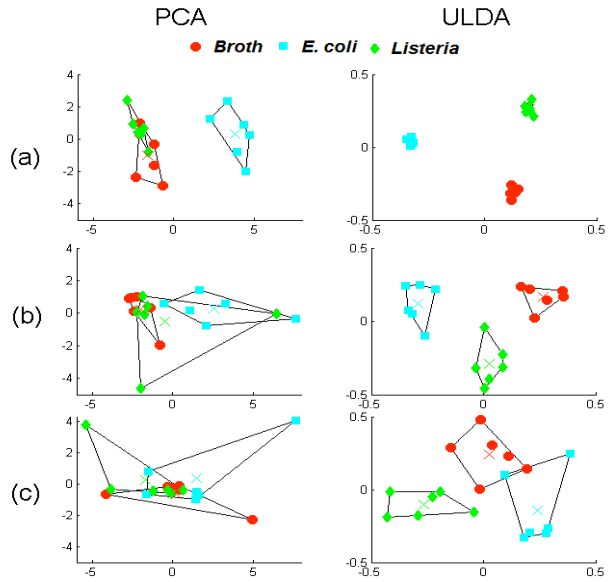

Figure 1: MOS EN clustering results showing the first two components in the dimension-reduced space. (a) 10^8 cells/mL, (b) 10^7 cells/mL, (c) 10^6 cells/mL. Cluster centroids are denoted with an 'X'.

Table 1: Scatter and Classification Accuracy results. Normalized intra-cluster scatter is defined as the average Euclidean distance between each scatter point and the group's centroid. Mean inter-class distance is measured between class centroids. k=3 was used in the nearest neighbours classifier.

Conc. (cells/ mL)	**Intra-cluster scatter (as % of mean interclass distance) – (MOS)**			**Classification Accuracy, % (MOS)**		**Classification Accuracy, % (FMS) - using kNN**	
	Broth	*E.coli*	*Listeria*	*Linear*	*kNN*	*ALL*	*Reduced*
PCA							
10^8	32.1	33.8	18.0	94.4	66.7	33.3	50.0
10^7	33.1	70.2	96.5	61.1	66.7	27.8	61.1
10^6	82.6	119.6	100.6	22.2	27.8	55.6	72.2
ULDA							
10^8	5.0	4.0	5.3	94.4	100	66.7	66.7
10^7	17.5	23.3	23.1	61.1	55.6	55.6	50.0
10^6	36.1	48.2	30.2	22.2	33.3	50.0	61.1

Table 1 quantifies this behaviour with measures of intra-class cluster size (normalized by the inter-class separation). Notice that the supervised method, ULDA, performs better than PCA. Upon inspection of Figure 1, we would expect that the classification accuracy of the system would be quite high at all concentrations; however, Table 2 indicates otherwise – at the lowest concentration, the accuracy of the system is similar to the result of

random guessing. Furthermore, while ULDA gives 100% accuracy in a specific case (using kNN at 10^8 cells/mL), at lower concentrations it does not give the improvement over PCA that we might expect based on the clustering results.

The reason for this seemingly paradoxical result is as follows. Figure 2 shows a series of plots indicating the scatter in the training samples. In each plot, a different, randomly selected sample is withheld (LOO cross-validation), meaning that the training model is built with 17 out of the 18 samples. The position and shape of each category's cluster is highly variable within each series (this behaviour is more prominent at the lower concentration), indicating that the model is sensitive to the omission/inclusion of a single sample. This variation certainly affects the transformation matrix that is subsequently used to classify the withheld test sample, resulting in misclassifications.

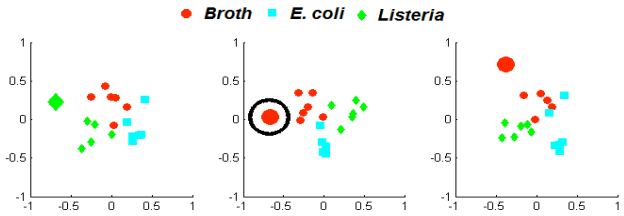

Figure 2: MOS EN clustering during training, 10^6 cells/mL. In each plot, one randomly chosen sample is withheld and ULDA is performed on the remaining 17 samples (small markers). The test sample (large marker) is then projected onto the model. The shape/colour of the test sample corresponds to the *calculated* category (using the linear classifier) – errors are circled in black.

3.2 FMS EN

Figure 3 shows the clustering when all of the mass spectrometer data (mass fragments 45-150 amu) are used. Using PCA, class separation is clearly not possible, but the extremely tight ULDA clusters at all concentrations might suggest excellent classification. The results in Table 3 indicate otherwise – in this case, the explanation is not solely due to the model's sensitivity to a small number of training samples.

Figure 4 shows raw FMS data, from which it is evident that only a few of the lower weight mass fragments (<100amu) are responding significantly – the others are simply measuring noise. Because the feature space is of very high dimension (D=106 in this case), ULDA can operate with a large number of degrees of freedom to find a projection that separates the clusters. This creates problems when novel samples are projected onto this model. The noise data permits erroneous degrees of freedom enabling the training clusters to be tightly defined, which causes poor generalization. In order to filter out the features that are essentially meaningless (causing invalid clustering), we removed those mass fragments which were below a noise floor (empirically determined; Figure 4) for all samples, resulting in 12 retained features. Results are shown in Figure 5 and the resulting classification rates are, in general, markedly higher (Table 3).

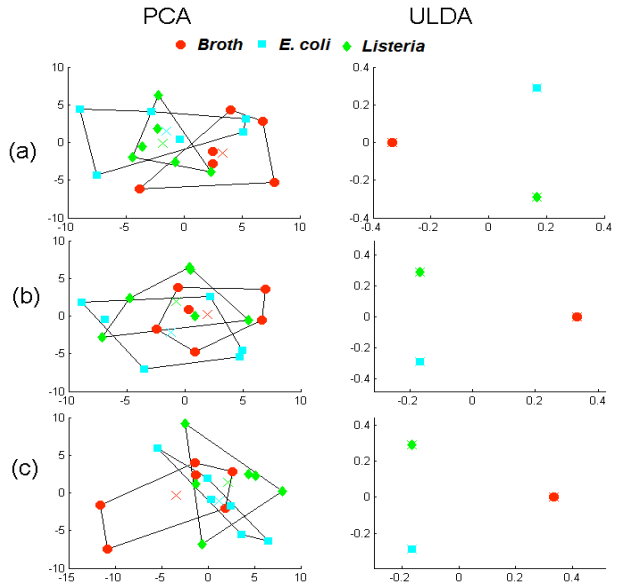

Figure 3: FMS EN clustering results (all fragments) showing the first two components in the dimension-reduced space. (a) 10^8 cells/mL, (b) 10^7 cells/mL, (c) 10^6 cells/mL. Cluster centroids are denoted with an 'X'.

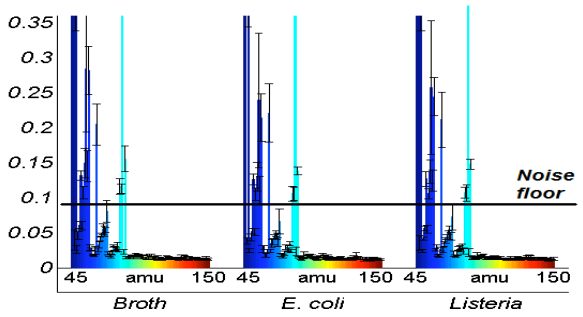

Figure 4: Raw data (FMS EN) for the 10^8 cells/mL samples (top truncated) with error bars showing standard deviation among 6 samples in each class.

CONCLUSIONS AND DISCUSSION

In this paper, we have described an application that stands to benefit from the sensible use of dimensionality reduction. The clustering results that we have presented demonstrate that supervised methods such as ULDA generate better separation between classes than the unsupervised PCA method. Using the MOS EN, discrimination becomes more difficult as the concentration decreases (from 100% accuracy at 10^8 cells/mL to 33.3% at 10^6 cells/mL, using ULDA and a kNN classifier). Curiously, the FMS EN did not exhibit this trend. Indeed, with the reduced set of mass fragments, similar classification accuracies (roughly 50-70%)

were seen across all concentrations (using ULDA). The clustering behaviour, however, exhibited excellent separation in all cases (using ULDA). FMS is generally considered to be a more sensitive technique than MOS [2]. One possible parameter that might contribute to this unexpected result is the number of mass fragments used (*i.e.* noise floor) – this, along with the repeatability of the instrument, will be a subject of further study.

We also investigated the EN's ability to *detect* bacteria (as opposed to *discriminate*) – here, the *E. coli* and *Listeria* samples were combined into one category. The best classification results obtained were – at 10^8 cells/mL, 94.4%, at 10^7 cells/mL, 94.4%, and at 10^6 cells/mL: 77.8%.

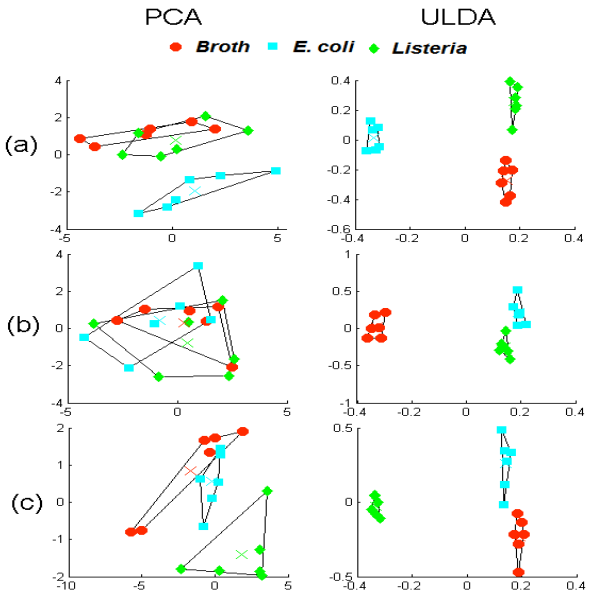

Figure 5: FMS EN clustering results (after elimination of noisy fragments) showing the first two components in the dimension-reduced space. (a) 10^8 cells/mL, (b) 10^7 cells/mL, (c) 10^6 cells/mL. Cluster centroids are denoted with an 'X'.

The performance of linear and kNN classification methods were similar. This implies that it is possible to compare dimensionality reduction methods without the results being too sensitive to the classifier type. Linear classification has the advantage of being simple (no parameters to choose), but the kNN classifier is able to generate complicated decision boundaries and might be useful for more difficult applications, along with other tools such as multilayer perceptron [5].

During the course of our investigation, we uncovered several potential difficulties that must be avoided when using the methods described herein:

1. Techniques that show promising clustering results during training do not necessarily generalize well with novel samples. We demonstrated this with the MOS EN, in which a small number of training samples led to classification models that are overly sensitive to individual data points.
2. It is important to eliminate noisy features before dimensionality reduction; otherwise the training model is fallaciously optimistic (due to the increased degrees of freedom), and will not generalize well. This was shown with the FMS.

We should emphasize that these pitfalls can be partially overcome by using a greater amount of training data. Experimentation (with synthetic data) has shown that classification models built with many more samples are more robust, and lead to higher classification rates (that are more in line with what we expect based on promising clustering results). The relatively small number of samples (n=18 at each concentration) used in this study contributed to the lower classification rates. This study represented preliminary work in this application and did not allow us to precisely determine the concentration thresholds of bacteria detection and discrimination. Further work is presently being done (at lower concentrations than 10^6 cells/mL) to quantify these thresholds. We have, however, demonstrated that supervised dimensionality reduction methods (like ULDA) can increase the classification performance at lower bacteria concentrations when compared to PCA. The cost for this potential improvement is vigilance in the manner in which they are applied.

ACKNOWLEDGEMENT

We would like to thank CFIA for providing the bacteria samples. This research was supported by the Natural Sciences and Engineering Research Council of Canada, the Canada Foundation for Innovation, and the Ontario Innovation Trust.

REFERENCES

[1] H.T. Nagle, R. Gutierrez-Osuna, S.S. Schiffman, "The how and why of electronic noses," *IEEE Spectrum*, vol. 35, pp. 22-31,1998.
[2] T.Pearce, S.Schiffman, H.Nagle and J.Gardner (Ed),*Handbook of Machine Olfaction*, Wiley, Weinheim, Germany, 2003.
[3] A.K. Pavlou, A.P. Turner, "Sniffing out the truth: clinical diagnosis using the electronic nose," *Clin. Chem. Lab. Med.*, vol. 38, pp. 99-112, 2000.
[4] E.C. Alocilja, N.L. Ritchie, and D.L. Grooms, "Protocol development using an electronic nose for differentiating *E. coli* strains," *IEEE Sensors Journal*, vol. 3, pp. 801-5, 2003.
[5] R. O. Duda, P.E. Hart, and D.G Stork, *Pattern Classification*, 2nd Edition, Wiley Interscience, New York, USA, 2001.
[6] J. Ye, T. Li, T. Xiong, and R. Janardan, "Using uncorrelated discriminant analysis for tissue classification with gene expression data", *IEEE/ACM Trans. Computat. Biol and Bioinform.*, vol. 1, pp. 181-190, 2004.
[7] http://www.alpha-mos.com
[8] G.C. Green, A.D.C. Chan, R.A. Goubran, "An investigation into the suitability of using three electronic nose instruments for the detection and discrimination of bacteria types", 28th Annual

International Conference of the IEEE-EMBS, New York, USA, 1850-1853, 2006.