A dimensionality-reduction technique inspired by receptor convergence in the olfactory system

A. Perera, T. Yamanaka, A. Gutiérrez-Gálvez, B. Raman and R. Gutiérrez-Osuna
Department of Computer Science, Texas A&M University, College Station, TX, USA
email: aperera@cs.tamu.edu

Abstract

In this paper we propose a new technique for feature extraction/selection based on the projection of sensor features in class space and taking into account the sensor variance. The proposed technique is inspired by the organization of the early stages in the biological olfactory system, and proves to be highly suitable for high-dimensional feature vectors with small number of training samples. We demonstrate the method on experimental data from two metal oxide sensors driven by a sinusoidal temperature profile.

1 Introduction

In a large number of applications using smart chemical sensors (sometimes referred as electronic noses) the goal can be defined as a categorization into a limited number of discrete classes from the information given by a set number of descriptors or features. Signal processing, and pattern recognition algorithms are useful tools applied in many steps after the chemical signals are translated to the digital domain. However, there are characteristics found in certain applications that configure harder problems for pattern recognition algorithms. These characteristics typically translate into a lack of generalization of the algorithm to unseen data. This is commonly found when the sensor system generates high dimensional input spaces.

Several authors have proposed different methods to generate high dimensional data from chemical sensors. Some make use of the dynamic information extracted from the evaluation of sensor transients [1,2] or sensor responses under certain temperature modulation profiles [3]. Other authors expanded the chemical information by means of a different sensor technology foundation. In 1996, White et al. presented a new sensor device built with an array of fiber-optic based chemosensors. Changes in dye fluorescence were recorded using a CCD device obtaining 256 channels [5]. High density configurations reaching 20k-fibers were achieved two years later by Michael et al. [6]. These works hint the future availability of sensor systems providing large dimensionalities.

Also, high dimensional spaces penalize the generalization performance of most classifiers. In 1968, Hughes showed that with a fixed design pattern sample, recognition accuracy increased with the number of measurements made, but decreased when the measurement complexity was higher than some optimum value [4]. This points out to an optimum number of features for a given a training set size which, unfortunately, is not known a priori. Furthermore, it has been shown that data distributed in high dimensional spaces has some interesting properties. Jimenez and Landgrebe proved that the volume of a hypercube concentrates in the corners as the dimensionality increases, suggesting than most of the multivariate dataset space is empty and that data distribution on high dimensional spaces might be counterintuitive, making density estimation a more difficult problem [7]. Additionally, the required size of the training set increases as a function of the dimensionality, linearly for a linear classifier and to the square of the dimensionality for a quadratic classifier [8].

However, typical algorithms found in electronic nose literature for feature extraction (e.g., Fisher's Linear Discriminant Analysis) or selection (e.g., Sequential Forward Floating Selection) are prone to over-fitting or computational ill-conditioning when the ratio of dimensionality to samples is large.

High dimensional sensory data is also found on the biological side. In the mammalian olfactory bulb, a very large number of receptors are processed through a topographic mapping: olfactory sensory neurons (OSN) expressing the same receptor project onto a single or few glomeruli [9]. It is know that mammals develop around 2 million olfactory sensory neurons in where each neuron expresses only a type of odorant receptor gene out of a repertoire of up to 1000 genes [10,11]. The study, characterization and modeling of this signal pathway [12,13], has lead to the development of neuromorphic signal-processing techniques for gas sensor arrays and may lead to new methods for processing of high dimensional data.

In this paper we propose an alternative statistical formulation inspired on the convergence principle by means of a convergence map created from the training set. We demonstrate its suitability for high-dimensional problems with small training set sizes.

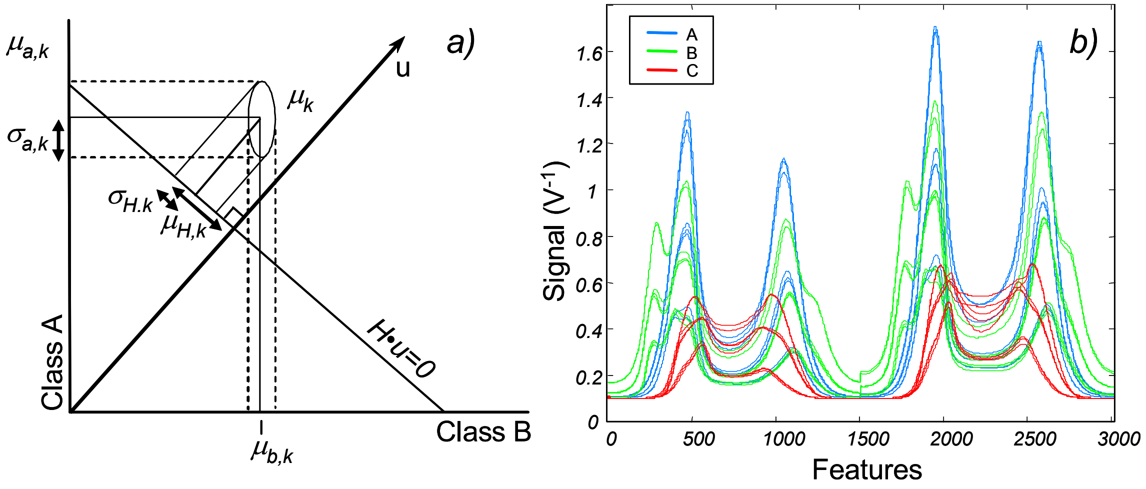

Figure 1. (a) Illustration of the projection of features ion the subspace *H*, orthogonal to *u*. (b) Response of two sensors under a sinusoidal temperature profiles and three analytes.

2 Method

We can define the response of the sensor system as a vector $\mathbf{x}$ in a D-dimensional feature space $\Re^D$, where D corresponds to the complete number of features extracted from the sensor array. Assuming that all classes have Gaussian likelihoods, data distribution for the training set can be described by the mean response of a given sensor (or feature) k to class c, $\mu_{c,k}$, and its standard deviation $\sigma_{c,k}$. Assuming C classes, a vector μ_k in a C-dimensional *class space* $\Re^C$ is defined to represent the mean response of each feature across all classes, where $\mu_k=\{\mu_{1,k},..., \mu_{C,k}\}$. Note that this *class space* is the dual to the conventional *feature space*, in which an odor sample is represented by its response across all sensors. The standard deviations for each feature ($\sigma_{c,k}$) define confidence figures for μ_k values. This confidence values can be generally expressed by a covariance matrix Σ_k in $\Re^C$.

The variance of a given feature k in class space is symmetric with respect to the axis. This is due to the fact that a feature can be placed in *class space* using any pattern from class c_i training set against any pattern from class c_j training set. This enforces a diagonal covariance matrix $\Sigma_{k,c} = \sigma^2_c$.

The projection of the features into class space provides some interesting options due to the space properties. Features alongside the hyper-diagonal $\mathbf{u}=[1\ 1 ... 1]^T$ (see Figure 1(a)) in class space contain no discriminatory information since they provide the same response to all the classes. Therefore, a measure of the discriminatory information can be obtained by the projection of the mean vector μ_k onto the subspace H, namely μ_{Hk}, where H is a C-1 dimensional subspace orthogonal to $\mathbf{u}$. Identically, to measure the confidence in the information content of feature k, we project the covariance matrix Σ_k onto the subspace H orthogonal to $\mathbf{u}$ (Fig.1 *(a)*), obtaining the projected mean μ_{Hk} and projected $\Sigma_{Hk,}$, both in $\Re^{C-1}$. Information about the discriminatory information, represented by the mean μ_{Hk}, and the its uncertainty, represented by its variance Σ_{Hk}, can be combined to find a set of values $\mathbf{w}_k=f(\mu_{Hk}, \Sigma_{Hk})$.

The selection of $f(\bullet)$ determines the method used to weight the discriminatory information once the uncertainty of the affinities is known from the training set. In the following we propose a heuristic method for $f(\bullet)$, although other options are possible. This heuristic is based on the intuitive idea that the relative relationship between μ_{Hk}, and the projection of Σ_{Hk} on H axis, σ_{Hk}, determine the confidence on the value of μ_{Hk}. This is computed element-wise using the function $f(\mu_{Hk},\sigma_{Hk})=\mu_{Hk}exp(|\mu_{Hk}| -\sigma_{Hk})$, which rewards features that are distant from the hyper-diagonal $\mathbf{u}$ and have low standard deviation.

We propose two different alternatives in the use of the resulting factors. For feature selection, similar discriminant vectors $\mathbf{w}_k$ are grouped using a partition of the subspace H (referred as ConvI). As a result, features with similar behavior (similar affinities to the set of C classes) are clustered together, in the same manner that olfactory receptor neurons expressing the same receptor converge onto the same glomeruli. Note that this is done via a *partition* of H, as opposite to a *clustering* (e.g. with a self Organizing Map (SOM)) in order to avoid a density estimation process in class-space. Estimation methods like a SOM would split high-density clusters creating multiple nodes although they would provide very similar discriminative power.

A second algorithm is proposed by using the set of factors $\mathbf{w}_k$ directly as a projection matrix, W. This projection maps $\Re^D$ into $\Re^{C-1}$ in a feature extraction sense (we refer to this method as ConvII).

The first method needs of the assignment of a feature to a group or partition in the subspace *H*. This step can be computationally expensive when the number of classes defined in the problem is high (class space will show high dimensionalities). The second method avoids this partition and therefore will be computationally cost-effective.

3 Robustness to over-fitting

Both PCA and LDA use the data covariance for building the linear projection. In the case of PCA, the complete dataset covariance is computed whereas LDA computes per-class scatter matrices. These covariance matrices result in strong computational issues under high dimensional datasets. The proposed algorithm evades covariance matrices computation and builds a linear projection by using only variance-mean relationship information. This is important in the case where covariance based methods are prone to over-fit the training dataset, degrading their generalization properties. In order to show this property, a synthetic experiment was designed with three normal distributions generated in a three-dimensional space. The dimensionality was extended by padding feature vectors with normal noisy channels $N(0,\alpha)$ (with $\alpha=0.1$), to achieve a 200 dimensional space. Each class was generated with a normal distribution $N(\mu,\Sigma)$ with a selection of parameters that provides with a dataset with discriminatory information in both mean and covariance, and defined by,

$$\Sigma_1=\begin{bmatrix}4 & 0 & 0\\ 0 & 0.6 & 0\\ 0 & 0 & 0.1\end{bmatrix}\ \Sigma_2=\begin{bmatrix}1.45 & -1.47 & 0\\ -1.47 & 3.15 & 0\\ 0 & 0 & 0.1\end{bmatrix}$$

$$\Sigma_3=\begin{bmatrix}1.45 & 1.47 & 0\\ 1.47 & 3.15 & 0\\ 0 & 0 & 0.1\end{bmatrix}\ \mu_1=\begin{bmatrix}-3\\ 0\\ 2\end{bmatrix}\ \mu_2=\begin{bmatrix}3\\ 0\\ 1\end{bmatrix}\ \mu_3=\begin{bmatrix}0\\ 3\\ 0\end{bmatrix} \tag{1}$$

The performance of a k-NN classifier in validation was computed after a reduction to two principal components using PCA, LDA and Convergence (Conv II). The number of samples per class used in the training set was varied from 66 to 600, corresponding to a $\rho=0.33$ to $\rho=3$, where ρ is defined as the number of samples per class over the dimensionality. Classification ratio was computed by means of a validation set with 1000 samples generated with the aforementioned distributions. Each experiment was repeated a total of 100 times. Results are shown in Figure 2. Note that PCA holds a stable performance for all computed values of ρ. On the other side, although Convergence and LDA provide similar classification rates at $\rho=3$, there is a noticeable performance drop of LDA at low ρ due to ill-conditioned with-in scatter matrix ($S^{-1}{}_{w}$) in the Fisher criterion. Results for this simple setup suggest that Convergence achieves similar performance to LDA, and even improves its performance avoiding data over-fitting effects.

4 Results

Proof-of-concept for the proposed method is illustrated with experimental data from two metal oxide sensors modulated in temperature. The sensors were exposed to dilutions of three different analytes (A,B and C), and their responses under a sinusoidal temperature profile (0-7 V; 2.5min period; 10Hz sampling frequency) were recorded (Figure 1(b)). As a result, the input space dimensionality was $D=3{,}000$. The proposed feature selection algorithm was performed with a uniform grid of four units covering the subspace *H*. As shown in Figure 3, the algorithm splits the temperature profile into different regions, which depend on their discriminatory information. Note that peaks characteristics of class B are grouped together under the G1 group, and features that provide low response to C but high response to A are grouped as G3. Overall the method is providing a rich combinatorial selection of which features are helpful for obtaining discriminant projections. After the convergence mapping is constructed, training and test data can be projected using grouping data using the convergence information:

$$g_i=\frac{1}{N_i}\sum_{k=1}^{D}c_{ik}x_k \tag{2}$$

where g_i is the output of the group *i*, N_i the number of features grouped in the set *i*, x_k is the feature *k* and c_{ik} takes *1* if the feature *k* converges to set *i,* and *0* otherwise.

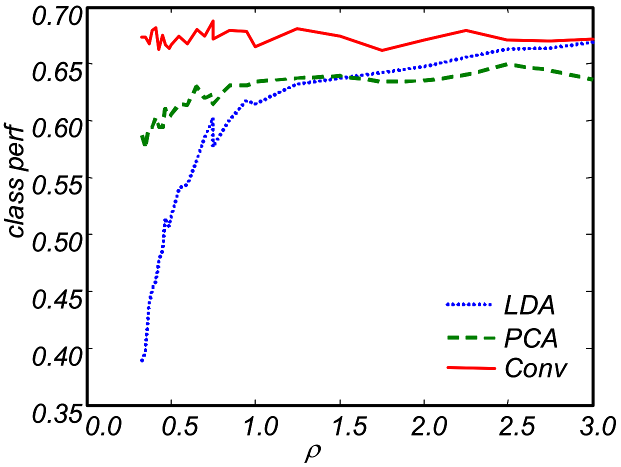

Figure 2. Comparative response of PCA, LDA and Convergence (II) under over-fitting conditions.

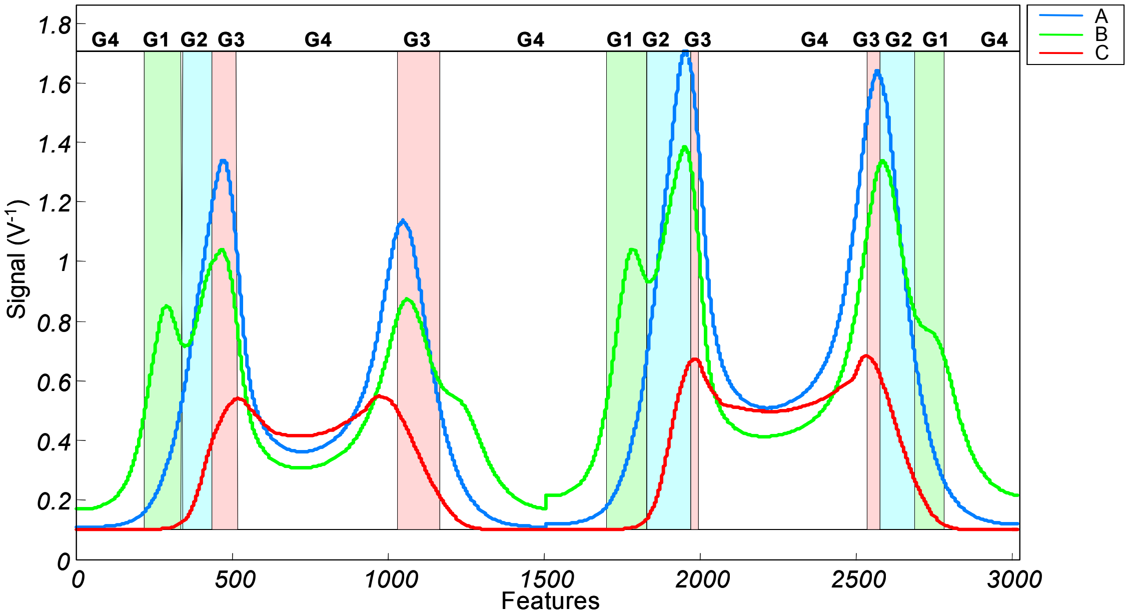

Figure 3. Result of the feature grouping on class-space. For visualization purposes, only the response to the highest concentration to each analyte is shown.

In Figure 4 we show the score plots from the output of PCA (a) and LDA (b) compared with those generated by the two algorithms proposed (ConvI in (c). ConvII in (d)). LDA projection was computed using a decimation of the waveform of the order 1:6. In the four algorithms, the models were trained using the lower dilutions (numbered as A1, B1 and C1) and the rest of dilutions were projected onto the respective models (2 and 3) as validation samples. In the PCA plot, it is noticeable that analytes A and B are confused at low concentration (dilutions 1 and 2) whereas in LDA those are correctly separated. The convergence based algorithms also separate the three analytes and show a similar behavior as LDA, with the difference that Convergence was computed using the complete sensor waveforms *D=3000*.

5 Conclusions

The proposed algorithms shown this paper are a result of two efforts: the study of the convergence seen from the population of olfactory sensory neurons to the glomerular layer form a signal processing view, and the construction of this convergence under the constrains given by the existence of a training set available to build the convergence map. This idea has led to the proposal of an algorithm based in grouping of features in class-space constructed with information that takes into account the relationship between mean and variance for each feature. An alternative method is derived from the first that avoids the feature clustering and builds a direct linear projection from the class space to a *C-1* dimensional space. The algorithms are computationally efficient under high dimensionalities and well suited for small-sample-set input spaces since it does not involve the computation of covariance matrices in feature space. Further work will consider a more theoretical formulation of the method, more quantitative analysis and will explore its generalization characteristics for other fields like image processing or genomics.

Acknowledgments

This work was supported by the National Science Foundation under CAREER Grant No. 9984426/0229598

References

1. Y. Hiranaka, T. Abe, and H. Murata "Gas dependant response in the temperature transient of SnO2 gas sensors," Sensors and Actuators B, Chemical, 9(3):177, 1992
2. R. Gutierrez-Osuna, H. T. Nagle, and S. S. Schiffman "Transient response analysis of an electronic nose using multi-exponential models," Sensors and Actuators B, Chem., 61(1-3):170–182, 1999
3. A. P. Lee and B. J. Reedy. "Temperature modulation in semiconductor gas sensing,". Sensors and Actuators B, Chemical, 60(1):35–42, November 1999.
4. G. F. Hughes, "On the mean accuracy of statistical pattern recognizers," in IEEE Trans. on inf. theory, vol. 14, 1, 55—63, 1968
5. J. White, J. S. Kauer, T. A. Dickinson and D. R. Walt, "Rapid analyte recognition in a device based on optical sensors and the olfactory system," Anal. Chem. Vol. 68, pp 2191—2202 , 1996

6. K. L. Michael, L. C. Taylor, S. L. Shultz and D. R. Walt "Randomly ordered addressable high-density optical sensor arrays," Anal. Chem. Vol. 70, pp 1242—1248, 1998
7. L. Jimenez and D. Landgrebe, "Supervised classification in high-dimensional space: geometrical, statistical and asymptotical properties of multivariate data," IEEE Transactions on Systems, Man, Cybertetics C, vol. 28, pp 39—54, Jan 1998
8. K. Fukunaga, "Effects of sample size in classifier design," IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 11, no. 8, pp 873—885, 1989
9. R. Vassar, S. K. Chao, R. Sitcheran, J. M. Nuñez, L. B. Vosshall and R. Axel, "Topographic Organization of Sensory Projection to the Olfactory Bulb," in Cell, vol. 79, 981—991, Dec 1994
10. R. Axel, "The molecular logic of smell," Scientific American, vol. 273, no. 4, pp. 154—159, 1995
11. R. Vassar, S. K. Chao, R. Sitcheran, J. M. Nuñez, L. B. Vosshall and R. Axel, "Topographic Organization of Sensory Projection to the Olfactory Bulb," in Cell, vol. 79, 981—991, Dec 1994
12. R. Gutierrez-Osuna "A Self-organizing Model of Chemotopic Convergence for Olfactory Coding," Proceedings of the 2nd Joint EMBS-BMES Conference, Houston, TX, 23-26 Oct 2002.
13. B. Raman, R. Gutierrez-Osuna, A. Gutierrez-Galvez and A. Perera, "Sensor-based machine olfaction with a neurodynamics model of the olfactory bulb," in Proc. 2004 IEEE/RSJ Intl. Conf. on Intl. Robots and Systems, Sendai, Japan, Sept 28 – Oct 2, 2004

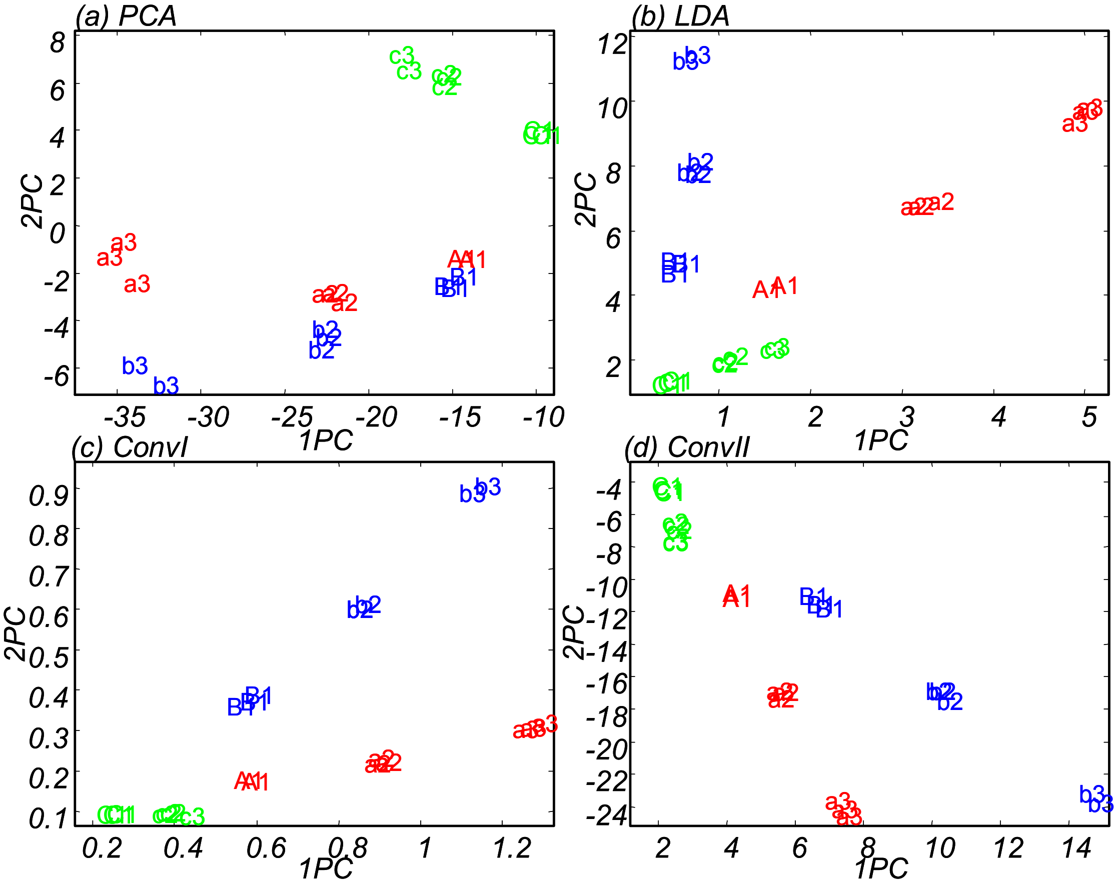

Figure 4. Scatter-plots for the four methods. (a) Principal Components Analysis, (b) Linear Component Analysis, (c) Convergence type I, (d) Convergence type-II