Machine Hound - Creating an artifical nose

Jacques de Chalendar, Marguerite Graveleau, Clément Renault*
Mentor: Youssef Ahres

December 2015

Introduction

Will a machine be able to follow an olfactory trail? That is, after learning an odor, is a computer able to recognize it, and predict the variation of its concentration? This challenge is offered on the website tunedit.org[1], with associated data.

The dataset represents object recognition in the olfactory domain. The testing data comprises the activity of a simulated gas sensor "traversing" an environment containing multiple odorant sources, as well as a persistent background odorant. In the training data the sensor is traversing the environment with just one odorant source present in isolation at a time[2]. The model will have for objective to learn the signature of that object during training (which is learned in mixture with a background odorant that is different from the background that will be present in testing). This task is analogous to the capability of biological systems to learn the smell of an object in one environment, and recognize it in multiple different contexts.

The training datasets are comprised of values representing sensor activation (input features) and odorant concentration (output) at 4000 different timepoints. The test is repeated for another machine (with the same odor environment but with different sensors). During the training recording, only one odor is present at a time, with a noisy background. The response of the sensors to a time-varying concentration signal is recorded for each odor. The testing set however, includes all 4 odors. The main challenge is therefore not to classify each odor, but to be able to recognize the concentration of each. We want to start off by trying machine learning techniques taught in CS 229. Nevertheless, neural networks might prove very efficient on this kind of problem, and it could be interesting to compare several techniques.

This problem is analogous to invariant object recognition in clutter and background. However, in our setting the object is a chemical signature detected by the activation of an array of olfactory sensors across which each odorant produces a distinct pattern.

An interesting application of this project would be the application of sensory networks trained by this data to enable neural sensorimotor control of the agent in the virtual worlds.

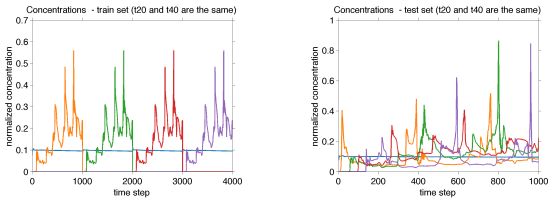
1 Study of the data

Before conducting any regression or machine learning algorithms, we want first to have a long look at the data, and study its shapes and main variations.

1.1 Output variables : concentration of odorant

The output variables are the concentration of each odorant, and the concentration of a background odor signature. During the training recording, each odor is present, but one at a time with the background (Figure 1a). However the testing data is a mixed signal of all the odors (Figure 1b). We have two sets of training+testing sensor activation, for two different machines which are named T20 and T40. They are trained on the same odor environment.

*jdechalendar@stanford.edu, mgravele@stanford.edu, clementr@stanford.edu
[1]website: http://tunedit.org/challenge/artificialOlfaction
[2]video demonstration: https://www.youtube.com/watch?v=kCCe88OMpA8



(a) Training set (b) Test set

Figure 1: Concentration profiles

1.2 Activation of sensors

Figure 2 describes the typical response of some of the sensors to the training odors. Some of the sensors seem to have all-or-nothing type behaviors while others follow the odor concentration signal.

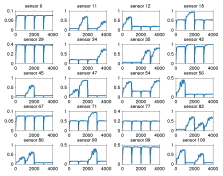
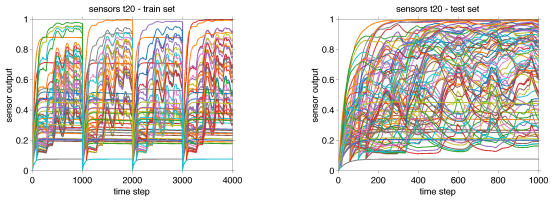


Figure 2: Activation of some of the sensors in T20 during the training

As we can see in Figure 3 (we plot the signal from 20 randomly chosen sensors), the activation of each sensor can almost be seen as binary. In presence of some odors, they will or will not go above a threshold (which is the same for each of them). However, they are mostly not specific of only one odor. This kind of 'activation' behavior could lead us to try to implement a neural network method later.



(a) Training set (b) Test set

Figure 3: Sensors on the T20 machine

1.3 Correlation of sensor readings with odor concentration

This study is mainly made using R (code is in appendix). In figure 4 we plot the activation of the 4 sensors the most correlated with odor 1 in function of the concentration of odor 1. We see that the 4 sensors most correlated with the response variables are also correlated together (visually the shapes are extremely similar and more specifically correlation of these sensors' readings are greater than 0.98). It seems that correlations are associative in this dataset.

It seems normal that some sensors are highly correlated, indeed they are activated by some kind of chemical molecules, if molecules specific to two sensors are close to each other, these sensors will tend to be activated at the same time. However, with correlation of almost 1, having those two features is not effective for our model. It could be interesting later to consider feature selection (forward or backward) which is likely to only retain one of those highly correlated features.

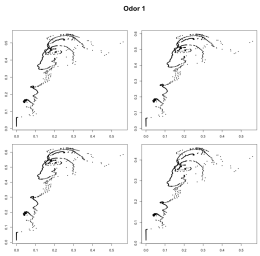


Figure 4: Correlation plots for Odor 1

2 Establishing a baseline

This preliminary work was done using R.

2.1 Linear regression

We first applied simple linear regression to the training set. In this model, each of the odor concentrations is predicted separately using all the sensor signals as features. Predicted versus true values are given in Figure 5 for odor 1. The results are similar for the other odors.

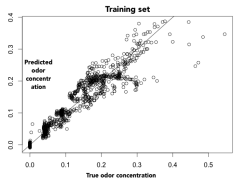


Figure 5: Base case linear regression for Odor 1

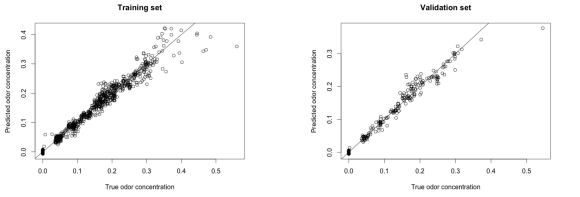
2.2 Forward selection on 1st order interaction terms

Scott, James and Ali (Data analysis for electronic nose systems, Microchim Acta 156, 183–207 (2007)) give a

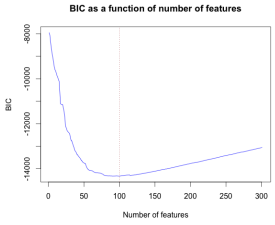
review of techniques used to analyze data from electronic nose systems. They emphasize that this problem is ripe for the application of feature extraction techniques. There are 100 sensors in our dataset. A quick look at the data shows that there are 10 duplicate sensors (both in the training and test data sets). In our preliminary modelling, we rejected the duplicate sensors. Other than that, we used the signals from the sensors as features directly. We then moved on to trying to reduce the number of features we use as inputs, using a forward selection algorithm.

We performed this analysis for the concentration of odor 1. The main findings here are that the interaction terms seem to be important. Indeed, if we consider as features the input signals and first-order interaction terms between the signals, and apply forward stepwise selection, we notice that out of the 100 first features that are selected only 5 are sensor signals, all the others being interaction terms. This makes sense if we consider the odor from a chemical perspective. The chemicals activate the sensors, and looking at combinations of sensors tells us which chemicals were present.

We plot the BIC obtained for each model against the number of features in the model in Figure 6c. We observe a U-shaped curve with a minimum at 100 features (We initially started with no feature and added one at a time so that the extended model is as good as possible).



(a) Base case (b) After forward selection



(c) BIC curve for the forward selection analysis

Figure 6: Linear regression results

Once we found the best model using feature selection we predicted the concentration of odor 1 using this new model, just like we had done with the first linear regression in 2.1. By comparing the regression in figure 6a and 6b it is clear that feature selection leads to much more accurate model. The spread around the first diagonal is very low in this new model. Indeed comparing Mean Absolute Error (MAE) between the model from 2.1, which has an MAE of 0.00788, and the model with feature selection, which has an MAE of 0.00407, we see that the MAE has nearly been divided by 2.

2.3 Limits

Although these results are very encouraging, one has to keep in mind that the actual test set will be comprised of a background odor, and a combination of the 4 odors. Whereas the training and validation sets are only comprised of the background odor, and one odor at a time. This means that a combination of sensor activations from different odors simultaneously present may pose a bigger challenge to our models.

3 Reducing dimensionality of the input space

3.1 Principal Components Analysis

The importance of interaction terms lead us to try out a Principal Component Analysis. We observe that indeed only the first 5 pc vectors explain more than 95% of the variance.
By plotting the first 5 principal components we can see that one seem to be specific of the bacjkgournd, and the 4 next are each specific of one odor.

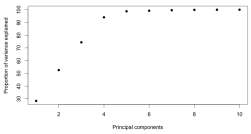
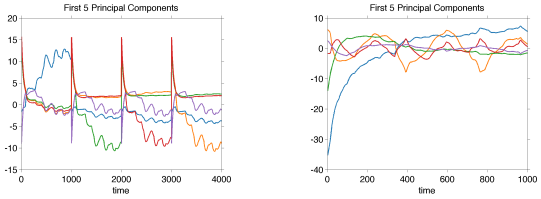


Figure 7: Variance explained in function of the number of principal components



(a) Training set (b) Test set

Figure 8: Plot of the first 5 pc in function of time

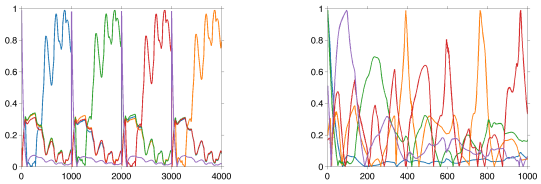
3.2 Linear regression models on the first principal components

Running a linear regression only on these 5 pc reduces the dimension of our input by a factor of 20, and only increases the mean squared error from 0.79% to 0.84%.

3.3 Independent Components Analysis

This problem is analogous to invariant object recognition in clutter and background, and called for an independent component analysis with the 5 first principal components as inputs (we normalize our data). In this particular study, let's notice that we don't use the information of the output (the actual concentration) in the training set. In order to get better results, we train our model on shuffled data (we go through the data randomly and not in function of time)

The assumption we are making in this model is that the first 5 PCs are linear combinations of the odor concentrations.



(a) Training set (b) Test set

Figure 9: ICA results : plot of the 5 "sources" in function of time

For the training we clearly observe 4 trends specific of each odor. It is not as clear on the testing set. However, by having a closer look, we can see that peaks are well positioned : for example the orange curve here represents the green odor of the testing set, it reaches its maximums at the same time, and always varying the same (increasing when the concentration increases). In order to follow a trail we only need to capture the variations of concentration. Nevertheless, the relative concentration of odorant are wrong. It may come from the normalization of our data, but without it the results are not conclusive.

4 Predicting the dominant odor

4.1 Training a model

The problem can also concern only determining which is the predominant odor (by opposition to outputting each individual concentration of odorant). By training a logistic regression on 80% of the dataset and validate it n the reminder we obtain less that 2% on the classification error (Figure 10 drescribes the results : the first bar separates the data of he training from the data on which we test the regression, the second bar gives the output classification).

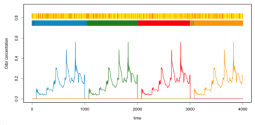


Figure 10: Classification using logistic regression

4.2 Using the predictions

From the ICA we can get the signal profiles, but since ICA is ambiguous with regards to the order of the sources, we do not know which is which. Using predictions from this model allows us to match the signals.

5 Possible future work

5.1 Additional pre-processing

Applying further data pre-processing techniques could provide better results. Ideas from the Scott et al. paper mentions hierarchical clustering analysis for example. Additionally, one of the difficulties of our data set is the presence of a background odor, and the fact that this background will be different in the training set. So simply removing the background when training will probably not be a great help. A baseline substraction of the sensors could help us however.

5.2 Time dependence of the signals

One aspect of the problem we have left to the side until now is the fact we are considering time series. We have been considering that the concentration of each odor can be predicted using the information from the sensor array at the same time step. In the test set, the variables are continuous, and a Markov chain approach could prove very efficient. Note that this sort of time dependence should be treated with caution, however, as we could be looking at data where the sensor array is

brutally taken from one environment and plunged into another.

5.3 Neural Networks

Given the shape of the activation of each sensor, being activated or not by some odors, we would expect a neural network approach to give good results. This data set actually comes from a challenge aiming to compare the performance of ML algorithms against neural networks. We tried to rapidly implement a neural network (1 hidden layer, 10 neurons), and got good results on the training set, but again on our testing sets the results were not conclusive. However our neural networks model was very simple, and called a for further work. One of the major difficulties of the data set seems to be that the test set is very different from the training set, not only in the shape of the signals but also with regards to the background odor.

Conclusion

- Simple linear regression models seem to give pretty good results on a test set drawn randomly from the training set
- Dimension reduction techniques, through feature selection or PCA, is well adapted to our data because sensors share a lot of information and are correlated by nature
- After selecting the first 5 principal components of the data, using an ICA approach allows us to recover signals that are visually close to the original ones. To resolve the inherent ambiguity of ICA (we don't know which odor is which), using logistic regression type methods to identify the dominant odor seems to work pretty well. The main remaining weakness of this approach is that the output signals are normalized and we lose the possibility of predicting an actual concentration value. Depending on the exact nature of the problem we are trying to solve, however, this may not actually be limiting.
- The test set appears different from the training set in two ways: (a) the odors are mixed together rather than separate (and their concentration is often below that of the background) and (b) the background odor is different. The second is one of the constraints of the problem. However, using acquiring a training set closer to the test in the shape of the signals could help make predictions on the test set.