

A neural network analysis of a dataset obtained through a carbon nanotube sensor array for breathomics applications

MICHELE ZANOTTI

Department of Mathematics and Physics, Università Cattolica del Sacro Cuore - via della Garzetta 48, 25123 Brescia, Italy

received 29 January 2022

Summary. — In this study, a dataset collected with a carbon nanotube based sensor array is analyzed to discriminate healthy from sick patients affected by chronic obstructive pulmonary disease (COPD). The dataset is analyzed with two approaches, the Principal Component Analysis (PCA) and the Neural Network. A comparison between the results obtained with these two methods is made. The applicability of the Neural Network is discussed, along with methods used to avoid the over-fitting problem.

1. – Introduction

An e-nose can be regarded as an array of sensors where the set of data collected by this device is analyzed by machine learning tools. This analysis is usually aimed to provide a classification of data according to specific requirements. For example in breathomics, the classification is aimed to identify patients with a specific disease, in food control the classification is aimed to identify fresh food samples, or the origin of specific products (coffee, tea, honey, wine, olive oil, cereals, and so on) [1-4]. In addition to sensor arrays, also analytical instruments such as gas chromatographers, mass spectrometers, and infrared spectrometers are often regarded as e-noses, because the spectral fingerprint of the investigated samples can be analysed with the statistical approaches used for sensor arrays data classification. Figure 1 summarizes the most common analytical tools, which are usually classified as branches of the more general machine learning approach.

Neural networks are among the possible tools to treat the data set produced by e-noses. Although PCA is the most diffused method to classify datasets, equipping also commercially available systems, neural networks can offer a robust classification of datasets outperforming, if the network is properly trained, PCA algorithms [5]. Among the sensors that can be used to build an e-nose, the most diffused are chemiresistive sensors and electrochemical sensors [6]. In particular, chemiresistive nanostructured carbon based sensing layers have recently attracted much attention due to good sensitivity,

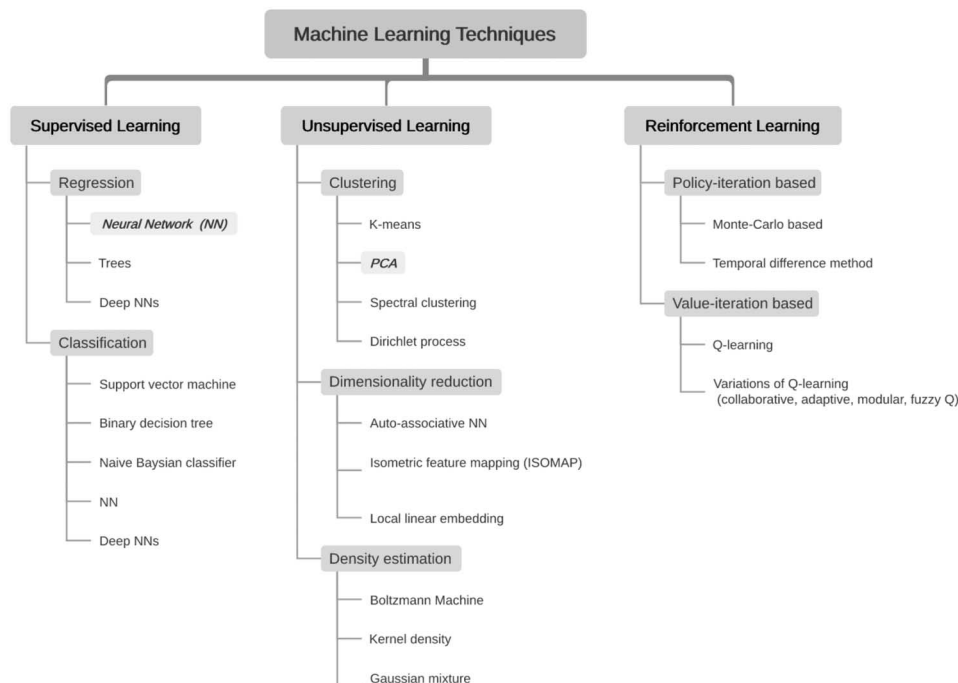


Fig. 1. – Classification of existing Machine/Deep Learning techniques.

stability, and the capability of operating at room temperature. Several applications have already been devised and some devices based on carbon nanotubes (CNTs) or graphene oxide (GO) have been brought up to a proof-of-concept stage, including the use in breathomics [7-10]. Indeed, exhaled breath contains a number of volatile organic compounds (VOCs) as gaseous molecules that are products of physiological and pathophysiological metabolic processes. Hundreds of VOCs can be detected in human breath, and their concentration may be altered due to infectious and metabolic diseases, genetic disorders, and various forms of cancer. When the concentration of specific VOCs is altered, these VOCs can be regarded as biomarkers of a particular disease. Therefore, VOCs analysis with e-noses addresses many of the issues related to breathomics, as it can be used for health diagnosis as a non-invasive, inexpensive, and widespread alternative to regular health screening campaigns.

In the present study, a dataset collected with a CNT-based sensor array [9, 11] is analyzed through PCA and neural network approaches. We discuss the feasibility of neural networks to discriminate healthy from sick patients affected by chronic obstructive pulmonary disease (COPD). Indeed, CNT layers are known to be quite sensitive to nitrogen dioxide and ammonia, which are usually regarded as COPD biomarkers in the exhaled breath.

2. – Experimental and computational details

The sensor array used to collect data was composed of 8 layers based on single-walled carbon nanotube (SWCNT) layers functionalized with organic molecules. The 8 sensors

were set on a properly designed board with 8 independent read out channels for the simultaneous detection of each sensor’s response. Relative humidity and temperature were also collected. The sensor signal was acquired using a script written in the LabVIEW environment. The sensing properties of the array upon gas/volatile exposures were analysed in the chemo-resistive configuration, where the presence of gases/volatiles is detected by monitoring the change in the resistance value of the sensitive element, *i.e.*, bundles of SWCNTs (bare or functionalized). Breath samples were collected (after signed consent) from 11 volunteers aged 22–88 years. Among them, 7 volunteers suffered from COPD, while 4 were healthy control volunteers. All volunteers were recruited within a research project funded by the Università Cattolica del Sacro Cuore in the frame of the 2016–2018 D 3.2 Strategic Program. For each volunteer, several samples were collected in different days. An overall number of 52 samples were analysed. Breath sampling was carried out in a disposable bag (volume = 0.6 liters), containing the sensor array, and inflated by breath through a disposable plastic straw. The exposure time was set at 3 minutes to let all sensors fully interact with the breath sample. All details about the sensor array and the data collection have been reported in [11] and [9].

The Neural Network was developed in `neuralnet`, an R package which allows training neural networks using backpropagation. The package permits to customize the error and activation function. The used neural network activation function is the logistic function,

$$(1) \quad f(x) = \frac{1}{1 + e^{-x}}.$$

The activation function is employed to process incoming information in every neuron and pass it through the network. In this work, the activation function is applied also to the output neurons. To understand the role of the activation function, it can be imagined as a process that sums all the input signals and determines whether the sum reaches the threshold. If it does, then the signal is passed through the network.

The training algorithm used is the standard backpropagation algorithm, in the most general form this algorithm iterates many cycles of two processes. Every cycle is called an epoch. In the very first cycle, the network contains no knowledge, the starting weights are typically set randomly. The algorithm goes on until a stopping criterion is reached. Each epoch includes:

- Forward phase: Neurons are activated from the input layer to the output layer, applying each neuron’s weights and activation function. When the signal reaches the final layer, the output signal is produced.
- Backward phase: The output signal resulting from the forward phase is compared to the true target value in the training data. The difference between the output signal and the true value results in an error that is propagated backward in the network. This error modifies the connection weights between neurons and reduces the final error.

This algorithm aims to find the minimum of the error function. The gradient of the error function is calculated with respect to the weights in order to find a root. Weights are modified going in the opposite direction of the partial derivatives until a local minimum is reached. If the partial derivative is negative the weight is increased, if the partial derivative is positive the weight is decreased. The algorithm stops when the partial derivatives of the error function reach the threshold (0.01). One problem with this algorithm is that it may find a local minimum.

TABLE I. – *Results obtained from the PCA in terms of variance that each component represents and the standard deviation.*

Parameter	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	2.6097	1.0664	0.61722	0.47172	0.41058	0.32764	0.25966
Proportion of variance	0.7567	0.1263	0.04233	0.02472	0.01873	0.01193	0.00749

3. – Data analysis

All the work done refers to the paper [11], and the data required to write this article were kindly provided by the authors. The data analysis was performed using R which is a programming language for statistical computing.

All data used were properly normalized by subtracting the mean from each column to center the data, and by dividing each column by its standard deviation to scale the data.

The first technique used for this analysis is the Principal Component Analysis (PCA). The first step in the analysis was decomposing the data in the space of principal components. The results obtained from this decomposition are presented in table I, for every principal component the standard deviation and the portion of variance explained are shown.

From table I we can see that the first two principal components describe 88% of the variance, so that using only these two components gives a good representation of the dataset. The first two components were used to discriminate healthy from sick patients. Figure 2 presents a graph of the two components plotted, on the x -axis there is the first component, and the second one is on the y -axis. Circles represent sick patients affected by COPD, and triangles represent healthy patients. The line is used to highlight the two clusters and to show that they are linearly separable.

As we can see from fig. 2 patients cluster as their condition, *i.e.*, all sick patients are on the left side of the graph and all healthy patients on the right. We can say that the two clusters are linearly separable except for one patient who appears to be positioned between the two clusters. If we draw a line to divide these two clusters, this sick patient is predicted to be healthy (we can see in fig. 2). The PCA was also used in the original article [11]. To compare the results and to see which method performs better. Another technique is also applied in this study. The second technique used to analyze the dataset is the Neural Network, which was developed using the R package neuralnet. There is not a reliable rule to determine the number of hidden layers and the number of neurons in each layer. The appropriate configuration depends on the number of input nodes, the amount of training data, and the complexity of the task. To choose the appropriate number of hidden layers and neurons, 8 different topologies were tested, in fig. 3 a scheme of all the topologies tested is presented, each configuration is associated to a letter, usually topologies are also labeled with a sequence of comma-separated digits representing the number of neurons for each layer, for example the configuration *A* can also be labeled as (9, 3, 2).

Since the dataset is composed of only 52 measures (coming from 11 volunteers), to

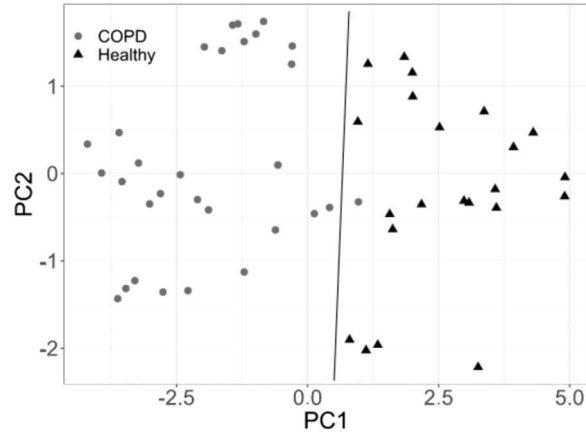


Fig. 2. – Result of the PCA analysis with the first two principal components. The circles are sick patients affected by COPD and triangles are healthy patients. The line is used to show that, except for one patient, clusters are linearly separable.

avoid the problem of over-fitting some precautions were adopted:

- When preparing the training and testing dataset, data from some patients (including the ones who were wrongly predicted by the PCA) was used only in the test dataset to see how the neural network behaves with unseen data.
- When preparing the training dataset, few measures per patient were used attempting to insert as many different patients as possible, consistently with the previous point.

When using these two precautions, the training dataset is homogeneous and the testing dataset contains unseen patients. This should have minimized the over-fitting probability. To choose the best topology among the eight ones that have been (fig. 3) tested, the SSE (Sum of Squared Errors) was considered. This error is defined as the sum of square differences between actual and predicted values and tells us how much the model fits the training dataset. Since the dataset used is quite small having a perfect fit on the

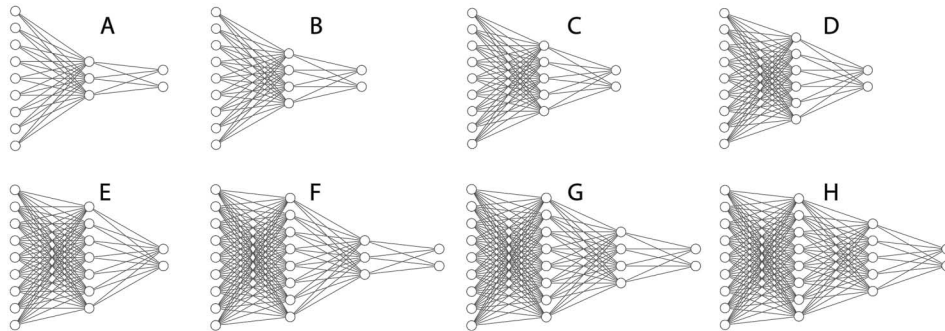


Fig. 3. – Topologies of the 8 configurations tested.

TABLE II. – *SSE and results of the predictions for the different 8 topologies.*

Configuration	A	B	C	D	E	F	G	H
Sick predicted healthy	0	0	0	0	0	0	0	0
Healthy predicted sick	3	3	1	1	3	1	1	0
Total errors	3	3	1	1	3	1	1	0
SSE error	0.011	0.004	0.011	0.005	0.009	0.011	0.006	0.012

training dataset, it can result in a model too specialized on the training dataset which cannot handle data coming from an unseen patient. This will be taken into account when choosing the configuration. As the SSE error gives us no information about how well the model works on the testing dataset, the number of correctly predicted measures on the actual test dataset was also used. Table II shows all the results obtained for the eight configurations.

A significant result is that all the wrongly predicted patients are healthy subjects which are predicted to be sick. From table II we can see that configuration H (with a 9, 8, 5, 2 sequence of neurons in the four layers) is the one that better fits our needs, as it is the one with the higher SSE error (less specialized on the training dataset) and it is the topology that predicts all patients accurately (handles well unseen subjects). A remarkable fact is that all the topologies are correctly predicting the patient that the PCA was failing to predict. This model improves its performance and reliability when we enlarge the dataset, so it is the best choice for a large-scale application.

4. – Conclusion

Data collected with a carbon nanotube based sensor array was analyzed using two techniques, PCA and Neural Network, to discriminate healthy from sick patients affected by COPD (chronic obstructive pulmonary disease). The feasibility of the Neural Network was discussed. To choose the best configuration, eight different topologies of neural network were tested. We discussed the problem of over-fitting and presented the methods adopted to avoid it. The results obtained using the two methods were compared. Unlike PCA, the Neural Network classifies all the patients correctly. A remarkable result is that, for all the tested configurations, wrongly predicted patients were always healthy subjects that were predicted to be sick. Neural networks improve their performance and reliability as we enlarge the dataset, therefore this technique is promising to handle data from large-scale screening campaigns.

* * *

The author acknowledges Giovanni Drera, Sonia Freddi, Aleksei V. Emelianov, Ivan I. Bobrinetskiy, Maria Chiesa, Stefania Pagliara, Fedor S. Fedorov, Albert G. Nasibulin, Paolo Montuschi, Luigi Sangaletti for making the dataset of [11] available for the NN analysis. A special thanks to Professor Luigi Sangaletti for helping in writing this article and for precious advice.

REFERENCES

- [1] PARK S. Y. *et al.*, *InfoMat*, **1** (2019) 289.
- [2] KANG M. *et al.*, *ACS Sensors*, **7** (2022) 430.
- [3] SANAEIFAR A. *et al.*, *TrAC Trends Anal. Chem.*, **97** (2017) 257.
- [4] WILSON A. *et al.*, *Sensors*, **9** (2009) 5099.
- [5] ZHANG L. *et al.*, *Electronic Nose: Algorithmic Challenges* (Springer) 2018, <https://doi.org/10.1007/978-981-13-2167-2>.
- [6] STAERZ A. *et al.*, *Electronic Nose: Current Status and Future Trends*, in *Surface and Interface Science: Applications of Surface Science I*, first edition, edited by WANDELT K. (Wiley-VCH Verlag GmbH and Co. KGaA) 2020, <https://doi.org/10.1002/9783527822492>.
- [7] ELLIS J. E. *et al.*, *ChemPlusChem*, **81** (2016) 1248.
- [8] CHATTERJEE S. *et al.*, *J. Mater. Chem. B*, **1** (2013) 4563.
- [9] FREDDI S. *et al.*, *Adv. Healthc. Mater.*, **9** (2020) 2000377.
- [10] FREDDI S. *et al.*, *Analyst*, **144** (2019) 4100.
- [11] DRERA G. *et al.*, *RSC Adv.*, **11** (2021) 30270.