Semi-Supervised Learning Techniques in Artificial Olfaction: A Novel Approach to Classification Problems and Drift Counteraction

Saverio De Vito, *Member, IEEE*, Grazia Fattoruso, Matteo Pardo, Francesco Tortorella, *Senior Member, IEEE*, and Girolamo Di Francia

***Abstract*— Semi-supervised learning is a promising research area aiming to develop pattern recognition tools capable to exploit simultaneously the benefits from supervised and unsupervised learning techniques. These can lead to a very efficient usage of the limited number of supervised samples achievable in many artificial olfaction problems like distributed air quality monitoring. We believe it can also be beneficial in addressing another source of limited knowledge we have to face when dealing with real world problems: concept and sensor drifts. In this paper we describe the results of two artificial olfaction investigations that show semi-supervised learning techniques capabilities to boost performance of state-of-the art classifiers and regressors. The use of semi-supervised learning approach resulted in the effective reduction of drift-induced performance degradation in long-term on-field continuous operation of chemical multisensory devices.**

***Index Terms*— Artificial olfaction, data streams, drift counteraction, dynamic environments, electronic noses, semi-supervised learning.**

I. Introduction

THE last decade have seen machine learning research focusing significant attention to Semi-Supervised Learning (SSL) techniques [1]. SSL aims to obtain superior machine learning performances by extending the supervised knowledge, provided by a small labeled training set, with the information content about samples distribution provided by an unsupervised sample set. In this sense SSL can be defined as mixing traditional supervised and unsupervised approaches. More specifically, it is based on the assumption of samples probability distribution $p(x)$ being meaningful to compute labeling conditional probability $p(y|x)$.

SSL now spans a multitude of approaches, either generative or not, that actually exploit both unlabeled and labeled samples' distribution in features space obtaining very interesting performances. It should be noted that the use of semi-supervised learning techniques require significant effort in order to choose the right model for the devised application. Each model requires strong assumptions on problem structure, consequently, if problem structure does not properly match with model assumptions this can easily lead to performance degradation instead of improvements [2], [3]. It is also remarkable to note that recent findings advocates for SSL being a relevant learning strategy that humans regularly use [2].

The opportunity to exploit unsupervised samples is interesting in many ways since obtaining supervised samples can be difficult and costly. Artificial olfaction (AO) problems are no exception to supervised samples shortage. Actually, the recording of a single supervised sample at calibration time requires significant amount of time typically involving skilled personnel. On the contrary, unlabeled samples are relatively cheap to obtain in multiple AO applications. Using SSL techniques, our data analysis algorithm could continue to learn after a brief training phase exploiting the unlabeled samples they will encounter in an e-nose operative life.

Another interesting point is that SSL techniques could be used for addressing the common solid state chemical sensors' drift issue that hampers the performance of the-state-of-the-art learning algorithms and seriously affects the diffusion of electronic nose technologies. Actually, the sensors and concept drift issues are the results of slow changes that affect sample probability distribution in feature space, driven by sensors transfer function, environmental conditions and sensed phenomena modifications during time. These changes result in the failure of the obtained supervised calibration due to inadequacy of the learned conditional $p(y|x)$ distribution. Interestingly, in a paper dating back to 1994, Shahshahani and Landgrebe noted that if the training samples are not very representative of the distribution in the real population then the use of unlabeled data might help to mitigate the resulting inductive bias [4].

Just as an example, on field pollution monitoring, either indoor or outdoor, is an artificial olfaction problem combining both sources of limited knowledge about joint samples and

Manuscript received October 31, 2011; revised March 14, 2012; accepted March 15, 2012. Date of publication April 3, 2012; date of current version October 4, 2012. This work was supported in part by the Project FP7-ENCOMB under Grant 26266. The associate editor coordinating the review of this paper and approving it for publication was Prof. Ricardo Gutierrez-Osuna.

S. De Vito is with the Italian National Agency for New Technologies, Energy and Sustainable Development (ENEA), Portici Research Center, Portici 80055, Italy, and also with the DAEIMI Department, University of Cassino and Lazio Meridionale, Cassino 03043, Italy (e-mail: saverio.devito@enea.it).

G. Fattoruso and G. Di Francia are with the Italian National Agency for New Technologies, Energy and Sustainable Development (ENEA), Portici Research Center, Portici 80055, Italy (e-mail: grazia.fattoruso@enea.it; girolamo.difrancia@enea.it).

M. Pardo is with the Istituto di Matematica Applicata e Tecnologie Informatiche, Consiglio Nazionale delle Ricerche, Genova 16149, Italy (e-mail: matteo.pardo@googlemail.com).

F. Tortorella is with the DAEIMI Department, University of Cassino and Lazio Meridionale, Cassino 03043, Italy (e-mail: tortorella@unicas.it).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSEN.2012.2192425

labels distribution: labeled samples cost and concept/sensor drifts. To the best of our knowledge, SSL techniques have never been applied to Artificial Olfaction problems, although some works contain interesting ideas for drift adaptation by using unlabeled samples. The work from Kim et al., for example, explicitly try to adapt initial knowledge, embedded in a RBF network classifier, using unlabeled samples (e.g. ref. [5]). However the work heavily rely on the strong assumption of non existing concept drift (i.e. $p(y)$ is assumed constant) that is rarely verified in real world applications.

In this paper, we show how to exploit SSL in Artificial Olfaction problems to counteract both limited availability of labeled samples and drift problems. As regards the drift counteraction, we apply SSL techniques to adapt the initial knowledge captured by supervised learning with the use of unsupervised samples. We report the performance gains obtained by applying a semi-supervised boosting algorithm to an artificial olfaction classification problem and a novel SSL-based algorithm to an air pollution monitoring dataset, in which the ground truth was delivered by a conventional gas monitoring station.

II. Related Work

A. Semi-Supervised Learning: A Brief Introduction

The idea of exploiting unlabeled samples to obtain improved pattern recognition capabilities dates back to early 60s but it's only in the mid 90s that a significant amount of efforts began to be spent on the topic. Recently, the most renowned semi-supervised learning techniques application field have become the internet: news filtering, e-mail spam detection, document classification, image classification. They share a peculiar shortage of labeled data to work with, a growing difficulty to achieve labeling and the availability of extremely large repository of unlabeled data.

Given a classification or regression problem and given a n samples labeled set

$$L = \{(x_1^l, y_1^l), \ldots, (x_i^l, y_i^l), \ldots, (x_n^l, y_n^l)\}$$

and a nl samples unlabeled reservoir set

$$U = \{(x_1^u), \ldots, (x_i^u), \ldots, (x_{nl}^u)\}$$

the goal of Semi-Supervised learners is to use samples drawn from U to pursue better insights on the joint y and x probability distribution $p(y, x)$[1] than those obtainable from L alone. Practically, unlabeled samples are exploited for modifying hypotheses obtainable only from labeled data. Generative models do it explicitly while discriminative models can pursue the goal by adding $p(x)$ dependent terms to their objective functions. As an example, Expectation Maximization (EM) with generative mixture models [6], is a generative SSL approach that use unsupervised samples to identify mixture components, at this point few labeled data are enough to completely bound the mixture distributions. Based on the Support vector machines (SVM) paradigm, TDSVMs (Transductive SVM or S^3VMs (Semi-Supervised Support Vector Machines), is a very interesting SSL technique [7, 8]. It was originally designed for addressing transduction problems and represent significant examples for semi-supervised discriminative models. It works estimating unlabeled samples labels and modifying decision boundary accordingly. It can easily be viewed as modifying SVM objective function adding a regularization term depending on unlabeled data.

SSL based algorithm are also often classified as being naturally *inductive* or *transductive*. Transductive algorithms are typically capable to express estimation only on data that they have dealt during semi-supervised training phase, i.e. with unlabeled samples that they have evaluated in the training phase. In these algorithms, training and test phase usually coincide. On the other hand, inductive SSL algorithms are capable to deal with previously unseen samples after the semi-supervised training phase. The learned model is hence defined throughout all feature space rather than on a limited subset. According to Vapnik, transduction represents a simpler task with respect to induction [8], however relationships between transduction and induction are philosophically disputed and the distinction line may eventually become fuzzier.

SSL classifiers are typically based on two main assumptions on data and joint data/ labels distributions:

1) Manifold assumption: it assumes that data classes spans low dimensional manifold in the feature space.
2) Cluster assumption: it assumes that decision boundaries lies on low density regions in data distribution, alternatively it is also stated as 'labeling being smooth on data'. Consequently, similar data must have similar labels.

In case of one or both this assumptions apply, we can expect that the use of unlabeled samples may be beneficial for the overall performance with respect to a supervised learner. However, in general those assumptions may not be verified, and even if they were, the use of improper semi-supervised learner model may cause unlabeled data not leading to any improvement or even to performance worsening.

One of the basic cluster assumption based approach is *self-training*. This can be defined as a boosting strategy that may be applied, to any pattern recognition algorithm like k-NN or Neural networks. It basically relies on the capability of a basic classifier/regressor to assign a label l, actually a pseudo-label, to an unlabeled sample. By incorporating the new (x^u, l^u) couple in the training set, the basic tool is able to train itself again refining its knowledge. In this sense, self-training *updates* the initial hypothesis by using unlabeled samples, step by step. One of the main limitation of this strategy is that the incremental knowledge obtainable by a generic pseudo-labeled sample may be limited: the approach is sensitive to the right choice of unlabeled sample to include in the training set. Actually, the inclusion of poorly pseudo-labeled samples may hurt the classifier/regressor performances. One way to overcome the latter issue is to rely on classifier/regressor reliability on the specific unlabeled sample, but this leads to the choice of high reliability samples that, for classification problems, usually lies far away from current separation hyperplane [9], [10]. In this sense, coupling Self-training with plain SVM classifier in a

[1]Given two random variables (A, B), the **joint probability distribution** for A and B p(A, B) defines the probability of events in which A = α and B = β simultaneously, α and β being two admissible values for A and B, respectively.

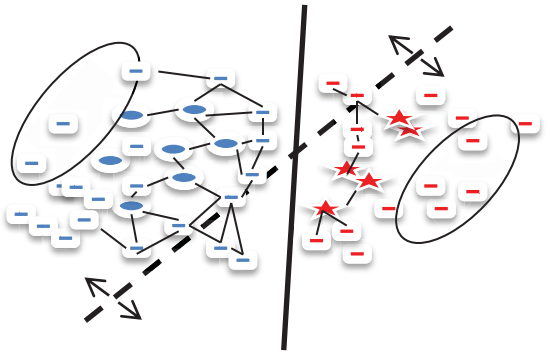

Fig. 1. Self-training criticism. Selecting high classification confidence unlabeled samples (encircled squares) for self training purposes does not provide sufficient incremental knowledge to correct the wrong initial separation hyperplane (dashed line) found by a SVM by using labeled samples (circles and stars). Only unlabeled samples located near the initial separation hyperplane allow for significant novelty inclusion, actually leading to hyperplane tilting and overall classification performance improvement. Exploiting samples *similarity* may help selecting and labeling significant samples.

two class problem can be an interesting example. Here, the choice of *high classification reliability* samples do provide very limited incremental knowledge and so provides very limited improvement in classification performances. Actually it only allows an increase of classification margins in the classifier, if any (see Fig. 1).

Furthermore, in the general case, you can still have samples whose labels are wrongly predicted even though the classifier reports a high reliability. To obtain significant improvements one should rely on different measures of classification/regression confidence. For example, pairwise *similarity* among labeled and unlabeled samples provides a confidence measure that is complementary to the one provided by the classifier and is often computed by geometrical distance-like functions. This is the approach followed by Graph-based SSL algorithms which make use of manifold assumption by exploiting data geometry and representing the data structure with a edge weighted graphs. Weights are obtained by computing a *similarity function* building the so called *similarity matrix* [11], [12]. In this framework, Manifold Regularization [13] is a self-contained algorithm that adds a similarity inconsistency term to SVM objective function so to obtain a large margin classifier that exploits data manifold geometry. SEMI-BOOST [29], described below, is a graph-based wrapper-like SSL strategy that tries to exploit both manifold and cluster assumption by balancing classifier posterior reliability and graph-based similarity in order to choose the right unlabeled samples to boost base classifier performances. In the following we will analyze the performance of SEMI-BOOST in a typical artificial olfaction settings i.e. for solving a coffee blends classification problem by using a very limited number of labeled samples.

B. Drift Counteraction

In the machine learning field, the term concept drift refers to the changes, that occurs over time, in the statistical properties of target variables, i.e. those one try to predict. In one of its simplest phenomenology, target variables area of interest position in the multivariate space changes significantly over time. Of course this is reflected by changes in the statistical distribution of the phenomena descriptive variables (inputs) that cause performance hits in the targets prediction process. Dealing with unpredictable or even recurrent concept drifts is at the basis of the growing interests in the study of pattern recognition algorithms under dynamic environments that is currently developing very rapidly.

In artificial olfaction, attention has instead been traditionally focused on sensors drift that is usually defined as the gradual variation in time of the e-nose readout towards nominally constant gas inputs. Graduality, i.e the relatively slowness in change, distinguishes drift from other (faster) sensors related e-nose perturbations, sometimes simply referred as noise. This is a qualitative and, at times, rather blurry distinction, the sense of which is, either to distinguish between irreversible changes in sensors and measurement system (more below on the drift causes) or to single out the variations which are slow enough to be corrected, e.g. by a periodic calibration. Still, there is some grey area: for example (unmeasured) temperature or humidity variations, which are well known to influence MOX sensors, may be catalogued as both drift and noise, depending on their time scale. Seasonal variations experienced by an e-nose working in field can be considered as drift, daily weather changes could be defined as noise. To complicate matters, daily temperature cycles which a lab e-nose is subject to, could in theory be modeled -if they occur with sufficient regularity- and that would not be defined as drift (this is probably the reason why unpredictability is sometimes included in drift's definition).

Sensor system instability -be it called drift or not- mathematically translates in data points no more being $i.i.d.$, i.e. independent and identically distributed. While the independence assumption may still hold true (though, strictly considered, the changes that occur in the enose depend on the actual gas samples series it measures), the samples are surely not identically distributed anymore. This in turn means that the input-output relationship learned by any pattern recognition method on a batch measured up to a time t, doesn't hold anymore at a later time $t' > t$.

The causes of drift are the chemical and physical interactions between on the one hand the sensors and the e-nose as a whole (e.g. filters, tubings, etc.) and, on the other hand, the gases to be measured and the surrounding environment. Also internal interactions, as the diffusion of contacts on the sensors, can play an important role. Regarding sensor modifications one distinguishes between ageing and poisoning. Ageing is the reorganization of the sensor surface or bulk, normally taking place on long time scales, while poisoning is the irreversible contamination, which may also occur over a short time (and therefore could be regarded as noise). On top of the sensor drift there is the e-nose system drift, which depends to a greater degree on the surrounding environment. It is then no surprise that most of the drift studies in the literature are on controlled lab conditions (as opposed to in-field measurements), which not only minimizes the sensor system drift but also permits to decide the sensor exposure history and to avoid poisoning. Moreover, drift studies are

hampered by the combination of the effort needed to collect a conspicuous dataset and habits to adopt small, time-limited datasets, where either no substantial drift is present or drift is not considered because training and testing sets are not divided along the time axes. Moreover, no specific public datasets exist.

Drift can be counteracted either by improving the robustness of sensors and sensor systems or through drift learning from data. The former direction would cut the problem at the root but is also the most difficult; Korotcenkov and Cho recently surveyed several hardware stability aspects, mostly related to sensor fabrication [14].

Software drift compensation can be taxonomized along few simple directions. The main distinction is whether a calibration gas is utilized, on which drift can be learned. Methods which do not consider regular calibration (also misleadingly called fully unsupervised or adaptive) are clearly easier and more cost-effective. Their main weakness is that they rapidly degrade performance once few patterns have been misclassified, as can happen in the case of overlapping classes or irregular presentation of samples belonging to the different classes. The point is that, in order to follow the drift without having the possibility to resort to an external truth in the form of a known calibration gas, they need to update the training data with the actual predictions, possibly disregarding old patterns. If predictions at some time are wrong, the algorithm could learn a false drift direction. Examples of this class of methods are self organizing maps [15], system identification theory-based learners [16] and local, class-dependent drift estimation [17].

When dealing with methods based on calibration measurements, coming directly from chemometricians' experience with calibrating analytical systems, the main distinction is between univariate and multivariate corrections. The former is normally a straightforward (univariate) multiplicative drift correction (MDC) [18]. Multivariate component corrections (CC) can be unsupervised or supervised [19]. In *unsupervised principal components corrections* (PCA-CC), the first principal components are calculated on the calibration data matrix, without reference to the time information (in this sense the correction is unsupervised) and then subtracted from the drift affected data. In supervised components correction, not the maximum variance direction in the calibration data is determined and subtracted, but the one explicitly related to the time variation, via PLS regression for example (PLS-CC). A further supervised CC method has been recently applied to enose drift correction, termed Orthogonal Signal Correction (OSC)[20]. OSC showed better performance than PCA-CC up to a time horizon of 100 days, at the expense of a bigger number of calibration samples, from all classes and conditions. Ziyatdinov et al. [24], estimate drift principal directions by extracting the common principal directions by analyzing sensor responses to all single target gases. In a performance evaluation, it compared favorably with PCA-CC. In the framework of multivariate drifts correction is also worth to mention approaches based on, Independent component analysis (ICA), Wavelet transforms [21-22] and Canonical Correlation Regression [25].

Romain and Nicolas compared MDC and PCA-CC(PLS), with ethanol as calibration gas [23]. Their drift study includes probably the longest measurement cycle. Contrary to the usual studies using certified gas bottles and mass flow controllers to generate reproducible and non-poisoning samples, they measured odors generated by an urban waste composting facility. Still, they sampled the gas on-field yet performed the measurements in the lab, thereby avoiding environmental stresses to the e-nose. Data collection lasted three years, with a couple of measurement session per year, followed by a further single measurement session more than 4 years afterwards. They find that MDC is best and that PCA-CC correction is even worse than no drift compensation at all. The latter point is quite surprising and would need a double check (data are not publicly available). The authors argue that in real-life measurements, it is difficult to identify a single direction in a multivariate space that is only correlated to sensor drift, and that a individual correction for each sensor is a more flexible alternative.

Summarizing, drift is a temporal process that hampers the initial knowledge about samples distribution. Our idea is to tackle the drift problem in a dynamic pattern recognition framework by adapting a regressor/classifier by means of unlabeled samples. Most of the effort in developing SSL has been targeted to classification problems. However many algorithms can be used for regression problems by simple adaptations. In this sense, COTRAINING is a regression wrapper-like strategy that employs different training classifier/regressors mutually training each other by selecting the right sample to include in the other own training set and estimating its pseudo-label [26]. In the following we will provide an adaptive strategy to exploit unlabeled samples based on COTRAINING algorithm.

III. Experimental and Results

A. Coffee Classification Setup

We employ a coffee dataset [27] collected with the Pico-1 e-nose at the Sensor Lab in Brescia. Measurements were done on *ground coffee*. Six single coffe varieties (SV) and 8 blends including 'Italian Certified Espresso' (ICE) were analyzed. Ten vials for every coffee type of the single group and 12 vials for every blend were prepared. Three successive extractions were performed from the same vial. Further details on the experimental setup and on the coffees can be found in the original paper [27]. We then formulated a two class balanced classification problem by using two groups of coffee ($2x124$ samples) as in the 'difficult' dataset from [27].

This experiment aimed at comparing the performance of a diffused classifier i.e. a Feed Forward Neural Network trained with Automatic Bayesian Regularization (ABR) [28] with a SSL empowered version of the same classifier when using a varying number of training samples.

In this experiment, we have selected a boosting-like approach to semi-supervised learning proposed by Mallapragada et al. [29] and called Semi-Boost. Since the original algorithm was actually designed with a Decision Stump classifier at its core, we adapted it for working with a BPNN (Back Propagation Neural Network) in order to obtain a fairer

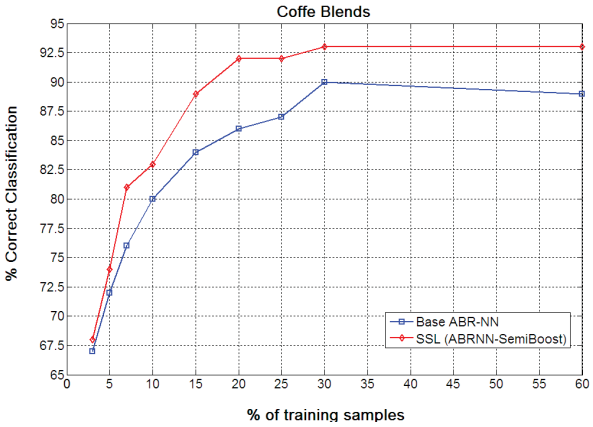

Fig. 2. Coffee blends classification problem. Semi-Boost algorithm, adapted to host a NN classifier at its core (in red/diamonds), can obtain significant performance enhancement with respect to the base classifier (in blue/squares). In both cases neural networks have been trained with automatic Bayesian regularization algorithm.

comparison between the two approaches. The core strategy is aimed at improving a base classifier performance by exploiting iteratively the knowledge associated with unsupervised samples. At each iteration it selects a new unlabeled sample from a dedicated unlabeled samples reservoir in order to include it in a new updated training set assigning a pseudo-label to it. Selection is both based on the "novelty" the sample can carry and on the probability to assign it a correct pseudo-label. A new classifier is hence trained at each iteration, and added to a classifier ensemble; the overall boosted classifier is hence obtained by combining the ensemble decisions.

Sampling (i.e. selecting the optimal unlabeled sample) and pseudo-labeling constitutes the core of the strategy. Actually sampling is performed by ranking a random selection of unlabeled samples by estimating their *relevance* for the current classifier. The latter is obtained by taking into account classification confidence and the presence of labeled samples in its neighbours. Relevance reaches its highest values when an unlabeled sample cannot be classified confidently by the actual classifier and, simultaneously, its closest neighbours is a labeled sample. Alternatively, the sample could be highly similar to an unlabeled sample that has been already confidently classified.

In our experiment, different training-test sets percentage splits have been tested ranging from using only 7 samples (about 3%) up to using 148 samples (about 60% of the dataset) for training purposes. The unlabeled reservoir set have been populated with 20% of the dataset samples. At each split, the percentage of correct classifications on the remaining test samples has been computed as main performance indicator. A 50x cross validation procedure have been carried out to reduce uncertainty on performance estimation and results differences have been checked for statistical significance at 0.95 level.

Results, depicted in Fig. 2, show how the Semi-Boosted version of the BPNN-ABR network outperforms the basic approach at any percentage split. Maximum advantages are obtained within the [7%, 25%] range. It is interesting to note that by using SSL and 37 training samples it is possible to obtain the same performance levels of a basic approach using 74 training samples, practically saving 37 labeled samples and the burden associated with their preparation. At the same time, the SSL approach allows us to reach performance levels in excess of 90% that revealed to be beyond the grasp of the basic approach. In the end, SSL managed to keep an edge over basic classifier obtaining more by using the same number of labeled samples.

B. Pollutants Concentration Estimation Setup

For drift counteraction experiments we considered a regression framework.

The testing dataset was chosen to be the ENEA-Pirelli pollution monitoring dataset, a one year long city air pollution monitoring dataset consisting of 6940 samples, recorded at hourly rate by an on-field deployed, solid-state chemical multisensor device. The device was equipped with 5 Metal Oxides (MOX) sensors plus temperature and Relative Humidity (RH) sensors. Ground truth values for pollutants concentrations were made available for five pollutants (CO, Benzene, NMHC, NO_x, NO_2) by means of a co-located air pollution analyzer; details can be found in [30].

In previous works, we showed that the use of at least 360 initial supervised samples as on-field recorded training set for a neural network (NN) based regressor achieved optimal sample-by-sample pollutants concentration estimation performances over at least six months. The use of more training samples did not allowed better performances and the obtained calibration lost precision over time due to the emergence of sensors and concept drifts. Actually, the city air pollution can be regarded as quasi-*ciclostationary* process subjected to daily, weekly and seasonal influences both due to human behavior (working hours and days) and weather induced phenomena (temperature and RH variations) or both (use of gas or fuel based house heating in winter time). As a result, target variables explore different areas in their uni-variate and multi-variate space (see Fig. 3). These variations are reflected in changes of the value distributions of the response of sensors devised to measure them. In addition, the well known sensor drift set in modifying the sensor response curve, however these modifications are not easily separable by changes inducted by changes in the distributions of interferents. Recalibration could help reducing influences but this would be a costly process requiring the deployment of a conventional mobile station, providing ground truth measurements, for more than 10 days. Hence, this could become unfeasible for a network built up by tens or hundreds of multisensor devices.

In this framework, we aimed at reducing calibration costs with the use of SSL-based techniques that can exploit a significant number of unlabeled samples coupled with a very limited amount of labeled samples. For its efficacy and simplicity, we propose to analyze the co-training framework which is based on the application of two (or more) different regressors which train cooperatively each other by using both labeled and unlabeled samples. For cooperative training, *diversity*

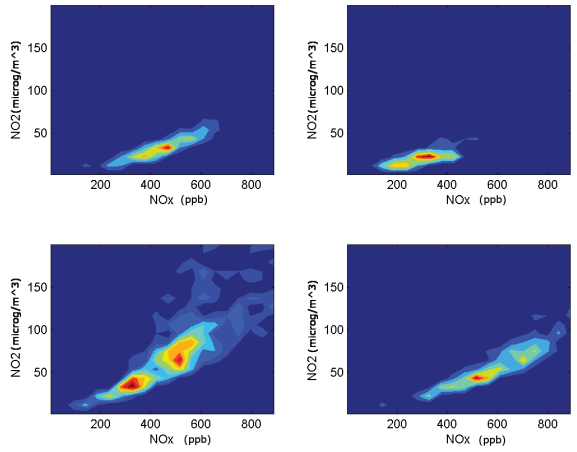

Fig. 3. Example of concept drift example in the air pollution dataset. Figures represent normalized histograms for NO_x and NO_2 joint values distribution as sampled in the first (upper left), sixth (upper right), eighth (down left), and eleventh (down right) month. Joint distribution significantly changes over time-driving regressors to work out of their learnt hypothesis.

is a key issue: it allows strengthening the capabilities of each regressor by exploiting the new knowledge derived from the label estimations of unlabeled samples performed from the other regressor based on a different "view".[2] Practically, at any training cycle, each regressor knowledge is updated by extending its training set with samples whose labels are predicted by the other cooperating regressor. Also in this case, unlabeled sample selection is crucial.

In the framework of cooperative semi-supervised regression we have developed and tested two variants of the cotraining-like COREG algorithm. In this algorithm (see fig. 4), the unlabeled samples are chosen from an unlabeled set reservoir in such a way that their inclusion in the cooperating regressor training set, lowers its empirical error. To this aim, each unlabeled sample is pseudolabeled by one of the current regressor (Fig. 4, row 10) and the attainable empirical error improvement is computed. This is done by computing the difference between the MSE obtained by estimating its kNN training samples labels by including, and not including, the pseudolabeled sample in the training set (Fig. 4, row 13). Pseudo-labeled samples are hence ranked in decreasing order of the computed improvement and the first sample (Fig. 3, row 60) is chosen to enlarge the training set of the cooperating regressor (Fig. 4, row 17). Diversity is ensured by the use of two different p-norm based $k-$NN regressors (we have chosen the classic Euclidean norm and the grade 5 "Minkowski" p-norm respectively) and two different numbers of nearest neighbors (respectively 3 and 5neighbors).

In a first experimental setting, we have hidden the presence of drift by randomly extracting training and test samples from the entire dataset, actually neglecting the time variable. In this way, the drift affected dataset can be considered as drift

[2]When cooperative training was started to be investigated, the main requirements included having at least two complementary but complete feature subsets (often referred as *sufficient and redundant views*), afterwise these requirement were definitely relaxed [31]. Different feature spaces, different regressors or differently biased regressors have been investigated.

Adapted COREG Algorithm:

Let p1= 2, p2 =5, k1=3, k2=5;
Let $\boldsymbol{L_1}$, $\boldsymbol{L_2}$ = $\boldsymbol{L}$ ($\boldsymbol{L}$ being the labeled training set)
Let $\boldsymbol{U}$ be the the unlabeled samples reservoir
Let paramer $\mathbf{T_{limit}}$= max iterations
T=0;
$\boldsymbol{h_1} = kNN(\boldsymbol{L_1}; \boldsymbol{k_1}; \boldsymbol{p_1})$; $\boldsymbol{h_2} = kNN(\boldsymbol{L_2}; \boldsymbol{k_2}; \boldsymbol{p_2})$;
Repeat
for j = 1 to 2 **do**
  for each $\mathbf{x}_u$ in $\boldsymbol{U}$ **do**
    $\boldsymbol{y_u} = h_j(\boldsymbol{x_u})$;
    $\boldsymbol{\Omega} = Neighbors(\mathbf{x}_u; \boldsymbol{k_1}; \boldsymbol{L_j})$;
    $\boldsymbol{h'_j} = kNN(\boldsymbol{L_i} \cup (\boldsymbol{x_u}, \boldsymbol{y_u}), k, \boldsymbol{p_j})$;
    $\Delta \boldsymbol{x_u} = \boldsymbol{MSE}_{\Omega}(\boldsymbol{h_j}) - \boldsymbol{MSE}_{\Omega}(\boldsymbol{h'_j})$;
  end
  if there exists an $\Delta \boldsymbol{x_u} > 0$
  then $\mathbf{X_j}$ = arg max $\Delta \boldsymbol{x_u}$; $\boldsymbol{Y_j} = h_j(\mathbf{X_j})$;
    $\boldsymbol{\pi_j} = (\mathbf{X_j}, \boldsymbol{Y_j})$; $\boldsymbol{U} = \boldsymbol{U} - \boldsymbol{\pi_j}$;
  else $\boldsymbol{\pi_j} = 0$;
end
$\boldsymbol{L_1} = \boldsymbol{L_1} \cup \boldsymbol{\pi_2}$; $\boldsymbol{L_2} = \boldsymbol{L_2} \cup \boldsymbol{\pi_1}$;
if neither of $\boldsymbol{L_1}$ and $\boldsymbol{L_2}$ changes
  then exit
  else $\boldsymbol{h_1} = kNN(\boldsymbol{L_1}; \boldsymbol{k_1}; \boldsymbol{p_1})$; $\boldsymbol{h_2} = kNN(\boldsymbol{L_2}; \boldsymbol{k_2}; \boldsymbol{p_2})$;
T=**T**+1;
until T is equal to $\mathbf{T_{limit}}$
Regressor $\boldsymbol{H}(x) = 0.5 * (\boldsymbol{h_1}(x) + \boldsymbol{h_2}(x))$;

Fig. 4. Adapted COREG algorithm description in pseudo-code. MSE refers to mean squared error.

free but strongly noisy. The Mean Absolute Error (MAE) performance indicator for Carbon Oxide (CO) concentration estimation problem has been computed using different sizes of labeled and unlabeled sets. Obtained results have been eventually compared with the values obtained with the base regressors. At each percentage split a 15x cross-validation procedure has been carried out.

Figure 5 depicts the performance gain obtained by the standard COREG algorithm when using 35% of the dataset as unlabeled set reservoir and using different labeled set percentage splits, ranging from 0.25% to 6%. The COREG algorithm always outperforms the best of the two base regressors even though performance gains are more evident at low training set percentage. A combined version of the two base classifiers basically reports results located in the middle of the two base classifiers performance curves, that curve have been omitted so to avoid picture cluttering. It is interesting to note that performance gain appears to be reduced to a small percentage as the training set fraction becomes more and more representative of the problem sample distribution, as can be expected.

Figure 6 analyzes the relationship among labeled and unlabeled sets sizes. The highest performance improvement can be observed at an unlabeled reservoir size of 10% and using, simultaneously, a very limited fraction of samples for

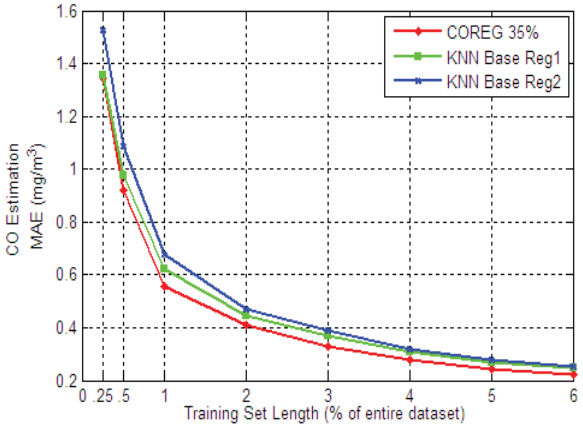

Fig. 5. Drift neutral experiment. A total of 15 runs averaged CO concentration estimation MAE versus training set length expressed in percent of the dimension of the entire dataset. About 35% of the entire database has been used as unlabeled reservoir for standard COREG algorithms. COREG results are compared with both base classifiers results.

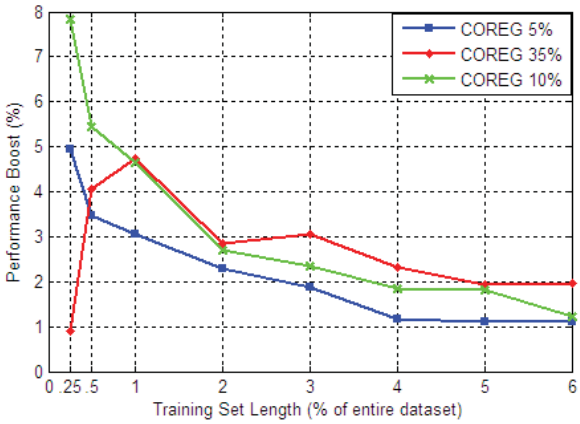

Fig. 6. Drift neutral experiment. MAE performance percentage boost obtained at different training set lengths (0.25–6%) by using different lengths of the unlabeled reservoir (5%, 10%, and 35%).

training (0.25%). Regardless of the sizes of the unlabeled reservoir here analyzed, performance improvement becomes lower as the sizes of labeled training set increase.

A bigger reservoir, seems to guarantee higher improvements although some exceptions have been found. At a training set split of 0.25%, limited performance improvement is observed by using 35% of the samples as unlabeled reservoir. This can be explained by the intrinsic limitations of the COREG algorithm in which selection of relevant unlabeled samples depends strongly on the base classifier performances on training set. In fact, the selected samples are often chosen among the ones that are close to the initial training set and this leads to small generalization improvements (see [29]). Less frequently, limited training sets when coupled with big unlabeled reservoirs could also lead to inclusion of badly labeled samples in the augmented training set. These results suggest that the ratio between sizes of unlabeled and labeled sets plays a key role in determining the obtainable performance improvements.

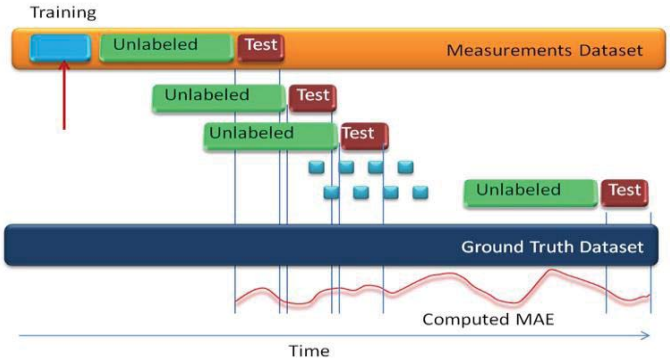

Fig. 7. Sliding window partitioning method in the "drift counteraction" experiment. A fixed length and position training set is used together with a sliding window set of unlabeled (and so unsupervised) samples to estimate gas concentration in a sliding fixed length test set. Estimations are then compared with the ground truth to provide MAE performance. In the proposed test, Training set size l was 24 samples, while Unlabeled reservoir size (u) was set as 100 samples. Test set size (t) was 24.

For the second regression setting, we have tackled the emergence of drift problems by designing an SSL-based adaptive strategy. Let a supervised training set L include l initial samples, while the unsupervised (U) and test (T) sets include the following u and t samples of the dataset, respectively (see Figure 7). The adaptive strategy that we propose is based on sliding the unlabeled reservoir along the time dimension; for any SSL core algorithm invocation, the unsupervised and test sets are t-shifted along all the dataset. Finally, the MAE of the proposed regressor on the whole test set have been compared with the performance obtained by the base classifiers, trained only with the initial 24 samples (see Fig. 7).

As SSL core algorithm, we have designed an original variant of the COREG algorithm that integrates $k-$NNs and BPNNs (Back Propagation Neural Network) regressors. According to this novel algorithm, "best confidence" unlabeled samples are chosen among the ones whose inclusion in the cooperating $k-$NN training set, lowers the $k-$NN empirical error. The latter is computed only on the sample's k nearest neighbors samples in the training set. Selection of "useful" unlabeled samples is hence based on a graph based technique (using 2 different p-norm based similarity matrices). Pseudo-labeling, that is perceived as the most critical aspect, is obtained by using two universal function approximators (Fig. 7 row 13); in this case a three layered BPNN, using 5 hyperbolic tangent as hidden-level units and trained with Automatic Bayesian regularization, has been selected. Practically, each of the two different BPNNs builds the evolving training set for the other one by cross-estimating the labels of unlabeled samples, chosen one at time. As a result they mutually enlarge their training set. For a detailed description the reader can refer to the pseudo-code in fig. 8.

The Figure 9 depicts a comparison of the 10 days moving average MAE between the base BPNN and the SSL based strategy for the CO estimation with $l = t = 24$ and $u = 100$ and maximum iteration number $R_{limit} = 24$. Actually R_{limit} limits the maximum amount of pseudo-labeled samples that will be added to the training set, in this case

CONNIE Algorithm
Let p1= 2, p2 =5, k1=3, k2=5;
Let L_1 , L_2 = L //(L being the labeled training set)
Let U be the the unlabeled samples reservoir
Let parameter $\mathbf{R}_{limit}$= max iterations
R=0;
$h_1(x)= kNN(L_1; k_1; p_1)$; $h_2(x) = kNN(L_2; k_2; p_2)$;
//(h_1 and h_2 are kNN regressors)
Repeat
for j = 1 to 2 **do**
  for each $\mathbf{x}_u$ in U **do**
    // ABR Neural Net Training and estimation of
      // the response on sample x_u
      $NN_j(x)=TrainBPNN(L_j)$; $y_u = NN_j(x_u)$;
      //Obtain the k_j-neighbours of x_u in L_j
      $\Omega = Neighbors(\mathbf{x}_u; k_j; p_j; L_j)$;
      // Build a new regressor by including (x_u,y_u) in Tr. set
      $h'_j=kNN(L_i \cup (x_u,y_u), k_j, p_j)$;
      // Evaluate the improvements on Ω set
      $\Delta Ex_u = MSE_{\Omega}(h_j) - MSE_{\Omega}(h'_j)$;
  end
  if there exists an x'_u so that $\Delta Ex_u > 0$
  then X_j = arg max Δx_u; $Y_j = BPNN(L_j, x_u)$;
    $\pi_j = (X_j, Y_j)$; $U = U - X_j$;
  else $\pi_j = \emptyset$;
end
$L_1 = L_1 \cup \pi_2$; $L_2 = L_2 \cup \pi_1$;
if neither of L_1 and L_2 changes
  then exit
  else $h_1(x) = kNN(L_1; k_1; p_1)$;
    $h_2(x)= kNN(L_2; k_2; p_2)$;
R=R+1;
until R is equal to $\mathbf{R}_{limit}$
$h_1(x) =BPNN(L_1)$; // ABR Train
$h_2(x)= BPNN(L_2)$; // ABR Train
Combined regressor H (x) = 0.5 * (h_1 (x) + h_2 (x));

Fig. 8. Description of the CONNIE algorithm in pseudo-code. This SSL-base algorithm represents the core of the SSL-based drift correction strategy.

training set dimensions will not exceed $24*2 + 24 = 72$ samples.

The use of the integrated $k-$NN-BPNN co-training algorithm obtains an overall performance gain of 11.5% on the MAE when applied along the entire one-year-long data set, with respect to the standard BPNN approach. During first month after the training has occurred, the obtained improvements are quite limited and in some cases the base approach slightly outperforms the SSL based approach (see days 10 to 20). As we go forward during the measurement period, performance boost appears more and more evident and seems to slightly reduce at the end of the year. Close inspection of Fig. 8 reveals how the adaptive strategy, starting from days 50, permits to obtain a performance edge over the basic approach even following the same oscillating pattern.

The results here presented show that the semi-supervised approach scored a performance gain over the basic approach, commonly applied from the artificial olfaction practitioners in both the proposed regression settings. Moreover, the use of moving window SSL-based strategy has allowed the regressor to adapt for concept and sensor drift effects by using unlabeled

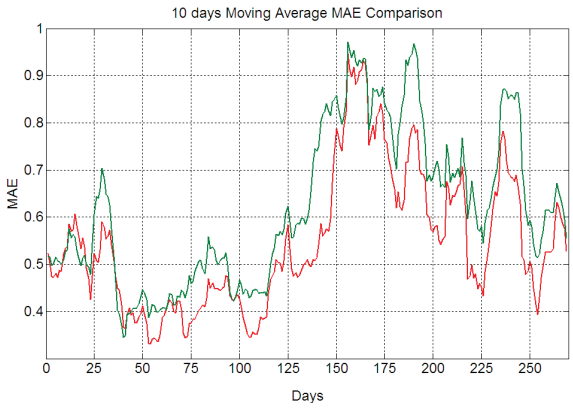

Fig. 9. Drift counter action experiment. CO Estimation comparison with integrated k-NN-BPNN SSL algorithm (red, 100 unlabeled samples reservoir) and standard NN algorithm (green) based on 24 samples. The SSL approach achieved a performance gain of 11.5% with respect to the one-year long averaged MAE score.

samples to modify hypothesis learnt over a very limited initial supervised time segment, reducing the number of samples needed for correction based approaches. This adaptation eventually reduced the drift effects on CO concentration estimation performances.

IV. Conclusion

In this work we proposed and tested the use of SSL techniques in the Artificial Olfaction domain to reduce the need for costly supervised samples and the effects of time dependent drift on state-of-the-art statistical learning approaches. SSL approach allowed for the significant reduction of the number of labeled samples needed to obtain a given performance goal in the coffee dataset. For the last purpose, an SSL based adaptive strategy have been devised and tested on an on-field recorded one-year-long continuous stream-like air pollution dataset. Instead of building hypothesis on drift structure by using significant supervised dataset, this fully adaptive strategy allow to exploit a small number of costly supervised samples profiting by adapting its knowledge to the slowly changing drift effects, by using up-to-date unlabeled samples. Actually, the performance advantage against the base approach was found to reduce by more than 11.5% the Mean Absolute Error computed over one year. Our combined results show that it is reasonable to expect that semi-supervised learning can provide advantages to the performance of data processing subsystems in artificial olfaction. They encourage us to further explore the interesting drift counteraction effect that the devised SSL based methodology has shown.

Further efforts should also be spent in clarifying parameters values dependence in the current setups. Moreover, at the moment, no information on pseudo-labeled samples is retained across different time-windows. For these reasons, we plan to investigate the possibility to retainingg past pseudolabeled samples adding simultaneously an ageing/forgetting process to take into account relevance fading due to combined drift induced $P(y, x)$ changes. Eventually, the availability of further real world data may allow to test more thoroughly the robustness of the approach to on-the-field harshnesses.

REFERENCES

[1] O. Chapelle, A. Zien, and B. Sholkopf, *Semi-Supervised Learning*. Boston, MA: MIT Press, 2006.

[2] X. Zhu, "Semi-supervised learning literature survey," Dept. Comput. Sci., Univ. Wisconsin, Madison, Tech. Rep. 1530, 2008.

[3] V. Castelli and T. Cover, "The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter," *IEEE Trans. Inf. Theory*, vol. 42, no. 6, pp. 2101–2117, Nov. 1996.

[4] B. Shahshahni and D. Landgrebe, "The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon," *IEEE Trans. Geosci. Remote Sensing*, vol. 32, no. 5, pp. 1087–1095, Sep. 1994.

[5] N. Kim, H.-G. Byun, and K. C. Persaud, "Novel signal processing techniques based on PDF information for sensor drift compensation," *Sensor Lett.*, vol. 9, no. 1, pp. 439–443, Feb. 2011.

[6] K. Nigam, A. K. Mc Callum, S. Thrun, and T. Mitchell, "Text classification form labeled and unlabeled documents using EM," *Mach. Learn.*, vol. 39, nos. 2–3, pp. 103–134, May–Jun. 1999.

[7] C. Distante, N. Ancona, and P. Siciliano, "Support vectormachines for olfactory signals recognition," *Sensors Actuat. B, Chem.*, vol. 88, no. 1, pp. 30–39, 2003.

[8] V. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[9] K. P. Bennet, A. Demiriz, and R. Maclin, "Exploiting unlabeled data in ensemble methods," in *Proc. 8th ACM SIGKDD Int. Conf. Knowl. Disc. Data Mining*, 2002, pp. 289–296,

[10] F. d'Alche-Buc, Y. Grandvalet, and C. Ambroise, "Semi-supervised MarginBoost," in *Proc. Neural Inf. Process. Syst. Conf.*, 2002, pp. 553–560.

[11] M. Szummer and T. Jakkola, "Partially labeled classification with Markov random walks," in *Proc. Neural Inf. Process. Syst. Conf.*, 2001, pp. 945–952.

[12] O. Chapelle and A. Zien, "Semi-supervised classification by low density separation," in *Proc. Int. Workshop Artif. Intell. Stat.*, 2005, pp. 57–64.

[13] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from examples," Dept. Comput. Sci., Univ. Chicago, Chicago, IL, Tech. Rep. TR-2004-06, 2004.

[14] G. Korotcenkov and B. K. Cho, "Instability of metal oxide-based conductometric gas sensors and approaches to stability improvement (short survey)," *Sensors Actuat. B, Chem.*, vol. 156, no. 2, pp. 527–538, 2011.

[15] S. Marco, A. Ortega, A. Pardo, and J. Samitier, "Gas identification with tin oxide sensor array and self-organizing maps: Adaptive correction of sensor drifts," *IEEE Trans. Instrum. Meas.*, vol. 47, no. 1, pp. 316–321, Feb. 1998.

[16] M. Holmberg, F. Winquist, I. Lundstrom, F. Davide, C. DiNatale, and A. Damico, "Drift counteraction for an electronic nose," *Sensors Actuat. B, Chem.*, vol. 36, nos. 1–3, pp. 528–535, Oct. 1996.

[17] M. Pardo, G. Niederjaufner, G. Benussi, E. Comini, G. Faglia, G. Sberveglieri, M. Holmberg, and I. Lundstrom, "Data preprocessing enhances the classification of different brands of Espresso coffee with an electronic nose," *Sensors Actuat. B, Chem.*, vol. 69, no. 3, pp. 397–403, 2000.

[18] J. E. Haugen, O. Tomic, and K. Kvaal, "A calibration method for handling the temporal drift of solid state gas-sensors," *Anal. Chim. Acta*, vol. 407, nos. 1–2, pp. 23–39, Feb. 2000.

[19] T. Artursson, T. Eklov, I. Lundstrom, P. Martensson, M. Sjostrom, and M. Holmberg, "Drift correction for gas sensors using multivariate methods," *J. Chemomet.*, vol. 14, nos. 5–6, pp. 711–723, Dec. 2000.

[20] M. Padilla, A. Perera, I. Montoliu, A. Chaudry, K. Persaud, and S. Marco, "Drift compensation of gas sensor array data by orthogonal signal correction," *Chemomet. Intell. Lab. Syst.*, vol. 100, no. 1, pp. 28–35, Jan. 2010.

[21] C. Di Natale, E. Martinelli, and A. D'Amico, "Counteraction of environmental disturbances of electronic nose data by independent component analysis," *Sensors Actuat. B, Chem.*, vol. 82, nos. 2–3, pp. 158–165, Feb. 2002.

[22] M. Zuppa, C. Distante, K. C. Persaud, and P. Siciliano, "Recovery of drifting sensor responses by means of DWT analysis," *Sensors Actuat. B, Chem.*, vol. 120, no. 2, pp. 411–416, 2007.

[23] A. C. Romain and J. Nicolas, "Long term stability of metal oxide-based gas sensors for e-nose environmental applications: An overview," *Sensors Actuat. B, Chem.*, vol. 146, no. 2, pp. 502–506, 2010.

[24] A. Ziyatdinov, S. Marco, A. Chaudry, K. Persaud, P. Caminal, and A. Perera, "Drift compensation of gas sensor array data by common principal component analysis," *Sensors Actuat. B, Chem.*, vol. 146, no. 2, pp. 460–465, 2010.

[25] R. Gutierrez-Osuna, "Drift reduction for metal oxide sensor arrays using canonical correlation regression and partial least squares," in *Proc. 7th Int. Symp. Olfact. Electron. Nose*, 2000, pp. 1–7.

[26] Z.-H. Zhou and M. Li, "Semisupervised regression with cotraining-style algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 11, pp. 1479–1493, Nov. 2007.

[27] M. Pardo and G. Sberveglieri, "Coffee analysis with an electronic nose," *IEEE Trans. Instrum. Meas.*, vol. 51, no. 6, pp. 1334–1339, Dec. 2002.

[28] D. Foresee and M. Hagan, "Gauss–Newton approximation to Bayesian learning," in *Proc. Int. Joint Conf. Neural Netw.*, Jun. 1997, pp. 1930–1935.

[29] P. K. Mallapragada, R. Jin, A. K. Jain, and Y. Liu, "SemiBoost: Boosting for semi-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 2000–2014, Nov. 2009.

[30] S. De Vito, M. Piga, L. Martinotto, and G. Di Francia, "CO, NO2 and NOx urban pollution monitoring with on-field calibrated electronic nose by automatic bayesian regularization," *Sensors Actuat., B, Chem.*, vol. 143, no. 1, pp. 182–191, 2009.

[31] N. V. Chawla and G. Karakoulas, "Learning form labeled and unlabeled data: An empirical study," *J. Artif. Intell. Res.*, vol. 23, pp. 331–366, Mar. 2005.

Saverio De Vito (M'10) received the M.S. degree in computer engineering from the University of Naples "Federico II," Naples, Italy, in 1998.

He was with the Dipartimento di Informatica e Sistemistica, Image Processing and Understanding Group, University of Naples "Federico II," in 1998, working on breast cancer computer-aided diagnosis. From 1999 to 2004, he was with a private ICT Engineering Firm as a Research and Development Team Leader for telemedicine, earth observation, and distance learning projects of ESA and ASI. In 2004, he joined Energy and Sustainable Development, as a Full Researcher. He has been a Contract Professor of applied informatics with the University of Cassino, Cassino, Italy, since 2005. He has coauthored more than 35 research papers. His current research interests include artificial olfaction and electronic noses, statistical pattern recognition, and intelligent wireless sensor networks.

Mr. De Vito has served as a referee for several international journals for the International Joint Conference on Neural Networks, the International Symposium on Olfaction and Electronic Nose, and Nature and Biologically Inspired Computing. He is currently serving as the Vice-Chair of the IEEE Chemometrics Task Force within the IEEE Computational Intelligence Society. He is a member of the International Association for Pattern Recognition-IC.

Grazia Fattoruso received the M.Sc. degree in computer science from the University of Salerno, Salerno, Italy, in 1999, and the Ph.D. degree in computer science of spatially distributed modeling from the Computer Science and Application Department, University of Salerno, in 2009.

She joined the Energy and Sustainable Development (ENEA), Portici Research Center, Portici, Italy, as a full time Researcher in 2000. She is an inventor in several patents. She has coauthored more than 20 scientific papers and participated in several EC projects and Italian government funded projects. Her current research interests include distributed intelligent sensing and modeling, applied geo-statistics, geographical information systems, and decision support systems.

Matteo Pardo received the M.Sc. degree in physics (*summa cum laude*) in 1996 and the Ph.D. degree in computer engineering in 2000.

He has been a Researcher with the Italian National Research Council, first with the Sensor Laboratory, Brescia, Italy, then with the Institute of Applied Mathematics and Information Technology, Genova, Italy, since 2002. From 2008 to 2010, he has been with the Max Planck Institute for Molecular Genetics, Berlin, Germany, with a Von Humboldt Fellowship for experienced researchers. He has authored 28 journal papers. His current research interests include data analysis and pattern recognition for artificial olfaction and genomics.

Dr. Pardo has been the Technical Chair of the International Symposium on Olfaction and Electronic Nose in 2009.

Francesco Tortorella (M'00–SM'09) received the M.S. degree in electronic engineering and the Ph.D. degree in electronic and computer engineering from the University of Naples "Federico II," Naples, Italy.

He was a Research Group Member on image processing and understanding with the Dipartimento di Informatica e Sistemistica, University of Naples "Federico II," from 1991 to 1998. In 1998, he joined the Dipartimento di Automazione, Elettromagnetismo, Ingegneria dell' Informazione e Matematica Industriale, University of Cassino, Cassino, Italy, where he is now a Full Professor of computer architecture and image understanding. He has authored over 80 research papers in international journals and conference proceedings and has served as referee for many international journals. His current research interests include classification techniques, statistical learning, neural networks, medical image analysis and interpretation, and document processing and understanding.

Prof. Tortorella is a member of the International Association for Pattern Recognition.

Girolamo Di Francia received the M.Sc. degree in physics from the University of Naples "Federico II," Naples, Italy.

He started his research activity in the field of fabrication and characterization of semiconductor solar cells (c-Si, GaAs), formerly with Ansaldo Corporation, Genova, Italy, in 1985, and then with the Energy and Sustainable Development (ENEA), Portici Research Center, Portici, Italy. Since 1991, he has been with the ENEA Research Center, Naples, where from 1992, he investigated porous silicon based devices. In 1996, he established there the Gas Sensor Laboratory that investigates sensing materials developing sensing devices and e-nose technologies. In 2009, he was appointed as a Scientific Research Manager. Since 2000, he has been teaching post-graduate courses in condensed matter physics with University of Naples "Federico II."