Semi-Supervised Consensus Clustering for ECG Pathology Classification

Helena Aidos[1(✉)], André Lourenço, Diana Batista[1], Samuel Rota Bulò[3], and Ana Fred[1]

[1] Instituto de Telecomunicações, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal
haidos@lx.it.pt
[2] Instituto Superior de Engenharia de Lisboa, Lisbon, Portugal
[3] FBK-irst, Trento, Italy
[4] CardioID Technologies, Lisbon, Portugal

Abstract. Pervasive technology is changing the paradigm of healthcare, by empowering users and families with the means for self-care and general health management. However, this requires accurate algorithms for information processing and pathology detection. Accordingly, this paper presents a system for electrocardiography (ECG) pathology classification, relying on a novel semi-supervised consensus clustering algorithm, which finds a consensus partition among a set of baseline clusterings that have been collected for the data under consideration. In contrast to typical unsupervised scenarios, our solution allows exploiting partial prior knowledge of a subset of data points. Our method is built upon the evidence accumulation framework to efficaciously sidestep the cluster correspondence problem. Computationally, the consensus partition is sought by exploiting a result known as Baum-Eagon inequality in the probability domain, which allows for a step-size-free optimization. Experiments on standard benchmark datasets show the validity of our method over the state-of-the-art. In the real world problem of ECG pathology classification, the proposed method achieves comparable performance to supervised learning methods using as few as 20% labeled data points.

Keywords: Electrocardiography · ECG · Semi-supervised learning · Consensus clustering · Evidence accumulation clustering

1 Introduction

Heart disease, or more formally cardiovascular disease (CVD), is the first cause of death worldwide. An estimated 17.3 million people died from CVD in 2008, representing 30% of all global deaths. Among these deaths, an estimated 7.3 million were due to coronary heart disease and 6.2 million were due to stroke. In the US, about 0.6 million people die from heart disease every year (25% of the deaths).

A. Bifet et al. (Eds.): ECML PKDD 2015, Part III, LNAI 9286, pp. 150–164, 2015.
DOI: 10.1007/978-3-319-23461-8_10

These statistics trigger our work, which provides a semi-supervised Electrocardiography (ECG) pathology classification system that tries to mitigate the aforementioned serious threats. The system builds upon the pervasive healthcare framework, where devices are becoming more handy, user-friendly and comfortable for the user, focusing on usability and allowing continuous (or quasi-continuous) monitoring of biosignals. The aim is to automatically classify ECG data streams acquired by monitoring devices, giving alerts of abnormal situations. The use of the semi-supervised learning paradigm is motivated by the existence of prior knowledge about classes in this domain, namely pathologies, which can be gathered from annotated records of some patients, but a larger number of records has no annotation, being this an expensive and time consuming process. This large amount of unsupervised data carries important information that a supervised learning approach would neglect, while being exploited by a semi-supervised learning approach.

Several clustering algorithms have been proposed exploiting side-information (e.g., [2,3,8,14,16]), using typically must-link and cannot-link constraints. Basu *et al.* [2] proposed a method for actively picking must-link and cannot-link constraints by selecting the most informative examples from the training set. On the other hand, Li *et al.* [16] presents a framework integrating consensus clustering and semi-supervised learning from a nonnegative matrix factorization perspective, allowing the must-link and cannot-link constraints to be enforced within the clustering algorithm. Gao *et al.* [14] proposed a framework that incorporates the predictive power of multiple supervised and unsupervised models, deriving a consensus label partition for a set of objects.

In this paper we propose a semi-supervised learning algorithm based on consensus clustering, *i.e.* the problem of finding a consensus partition among a set (or ensemble) of baseline clusterings that have been collected for some data under consideration. Our method follows the Evidence Accumulation Clustering (EAC) paradigm [12], which summarizes the information of the clustering ensemble into a pairwise co-association matrix, where each entry corresponds to the number of times a given pair of objects is placed in the same cluster. The advantage of the pairwise voting mechanism is that it subsumes the problem of cluster correspondence among partitions. Several algorithms for consensus clustering have been proposed based on EAC [1,13,17,19,23]. In [23] the problem of extracting a consensus partition is formulated as a matrix factorization problem, in a similar fashion as [16]. In [18,19] the consensus partition is estimated through a probabilistic model for the co-association matrix, while in [17] a generalization of [23] is introduced to cope with partial observations of the co-association matrix. In contrast to the typical unsupervised scenario, which is addressed by the aforementioned works, our method allows to exploit partial knowledge of the cluster assignment of a subset of data points to constrain the solution space of the consensus partition. Computationally, the consensus partition is sought by exploiting a result known as Baum-Eagon inequality [5] in the probability domain, which allows for a step-size-free optimization.

The rest of the paper is organized as follows: in Section 2, we describe the application context of our algorithm, positioning it on the future workflow of pervasive healthcare. In Section 3, the proposed algorithm is described. Section 4 is devoted to the experimental validation on standard benchmark datasets and presents results for the real world problem of classifying pathologies in ECG. Finally, in Section 5, we draw conclusions and outline future works.

2 ECG Pathology Classification

The use of Electrocardiography (ECG) as a diagnostic technique is a well established medical practice rooted in the pioneering work by Einthonven in the end of the 19th century. Clinical practice relies mainly on the widespread short-term ($<$ 1 minute) 12-lead ECG for diagnosis and, in selected cases, on Holter monitors ($\sim$ 24 hour assessment), providing information for the diagnosis and prevention of a wide array of cardiovascular disorders [7,9]. Nevertheless, the outreach of ECG data acquisition and processing can still be significantly improved in the context of a pervasive healthcare framework with the off-the-person ECG paradigm [25]. The goal of off-the-person approaches is not to replace existing data acquisition procedures, but to enhance and complement current practices with a simplified sensor setup that can be transparently brought to the subject, in multiple aspects of his everyday life. This enables a more comprehensive assessment of cardiovascular function, contributing to the development of preventive behaviors and methodologies. Also, it opens the door to many potential applications, such as continuous monitoring, cardiac dysrhythmia detection, and ECG biometrics [20,22], among others.

The commercial exploration of such concepts has already began, and one of the most successful products is AliveCor[1], a Heart monitor for mobile devices. It consists on a 1-lead ECG acquisition system that can be installed on a mobile device, which records the ECG using the hands of the user. The system enables the detection of Atrial Fibrillation (AF), and upload of the recorded information for expert revision.

(a) AliveCor Heart Monitor. (b) CardioID Keyboard.

Fig. 1. Examples of pervasive healthcare devices.

[1] http://www.alivecor.com/

This type of devices fits in the category of pervasive healthcare, where sensors do not need to be with the person, but instead are embedded into everyday use objects. A major advantage of this approach is the fact that the sensor placement does not require a voluntary action from the user, unlike, for example, wearable on-the-person devices, aligned with future medical trends [26]. Figure 1 illustrates two examples, the AliveCor Heart Monitor, and a keyboard with integrated electrodes developed by CardioID[2] that enables continuous ECG monitoring. These approaches produce enormous amounts of data. As an example, in a typical acquisition setup, where the signal is sampled at a frequency of 1 KHz, and with a resolution of 12 bits/sample, a total amount 5 MB of information is acquired per hour, corresponding to 123 MB/day/person or 44 GB/year/person, leading to a scenario where cloud computing is the most desirable approach.

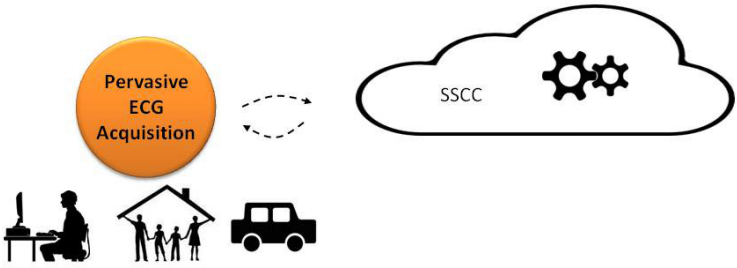


Fig. 2. Global Architecture.

The global architecture of such a system is illustrated in Figure 2, where pervasive healthcare devices stream data to the cloud, and automatic classification algorithms process the data. In this paper we focus on the classification algorithm and propose a novel semi-supervised consensus clustering (SSCC) algorithm to categorize the data.

3 Semi-Supervised Consensus Clustering (SSCC)

Consensus clustering is the problem of organizing a set of n data points $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_\mathsf{n}\}$ into groups, starting from the output of different clustering algorithms[3] that have been run on $\mathcal{X}$, or on sub-sampled versions thereof. This set (*a.k.a. ensemble*) of clusterings is denoted by $\mathcal{E} = \{\phi_1, \ldots, \phi_\mathsf{m}\}$, where each $\phi_u \in \mathcal{J}_u \rightarrow \{1, \ldots, \mathsf{k}_u\}$ is a function encoding a partition of a subset of data points indexed by $\mathcal{J}_u \subseteq \mathcal{I} = \{1, \ldots, \mathsf{n}\}$ into k_u clusters. Partitions not comprising all data points, *i.e.* such that $\mathcal{J}_u \subsetneq \mathcal{I}$, indicate clusterings of sub-sampled versions of $\mathcal{X}$. The use of sub-sampling is motivated, *e.g.* in the presence of large-scale datasets, or to promote diversity in the ensemble [10].

[2] http://www.cardio-id.pt/

[3] Or different parametrizations of the same algorithm.

As the name suggests, consensus clustering tries to find a good representative for all clusterings in the ensemble $\mathcal{E}$. Formally, we call *consensus partition* a partition having minimum divergence from the other partitions in the ensemble:

$$\phi^\star \in \underset{\hat{\phi}\in\mathcal{I}\to\{1,\ldots,\mathsf{k}\}}{\arg\min} \sum_{u=1}^{\mathsf{m}} d(\hat{\phi}, \phi_u)\,, \tag{1}$$

where $d(\cdot,\cdot)$ is a divergence measure between partitions.

In this paper, we depart from a purely unsupervised approach in favor of a *semi-supervised* perspective, by assuming partial knowledge of the cluster assignments (*a.k.a.* labels) of a subset of data points. Accordingly, we denote by $\mathcal{L} \subset \mathcal{I}$ the indices of data points that are labeled, and by $\ell_i \in \{1,\ldots,\mathsf{k}\}$ the label given to the ith data point, $i \in \mathcal{L}$. We can then use this a prior knowledge to constrain the solution space of (1) to obtain a semi-supervised consensus clustering formulation, *i.e.*

$$\begin{aligned} \phi^\star \in \underset{\hat{\phi}\in\mathcal{I}\to\{1,\ldots,\mathsf{k}\}}{\arg\min} & \sum_{u=1}^{\mathsf{m}} d(\hat{\phi}, \phi_u) \\ \text{s.t. } & \hat{\phi}(i) = \ell_i \text{ for all } i \in \mathcal{L}\,. \end{aligned} \tag{2}$$

The same knowledge could in principle be exploited at the ensemble construction phase. However, constraining the clusterings will lead to a drop of the ensemble's diversity, thus loosing one of the most desirable properties of an ensemble [15]. Moreover, there is a vast number of unsupervised clustering algorithms available for the ensemble construction, and only a limited number of algorithms that have been extended to include constraints.

In order to compare two clusterings, we face the so-called cluster correspondence problem, *i.e.* two partitions are the same if we can turn one into the other by a proper re-labeling of the clusters, and if two partitions are different, we would like to measure their divergence under the best possible re-labeling. There is however a way to sidestep the cluster correspondence problem, by adopting a pairwise divergence measure like the following one:

$$d(\hat{\phi}, \phi_u) = \sum_{i,j\in\mathcal{J}_u} \left[\mathbb{1}_{\hat{\phi}(i)=\hat{\phi}(j)} - \mathbb{1}_{\phi_u(i)=\phi_u(j)}\right]^2, \tag{3}$$

which counts the number of times two data points are clustered together in $\hat{\phi}$, but not in ϕ_u, and vice versa. In (3), $\mathbb{1}_P$ denotes the indicator function for the truth value of proposition P.

The objective function in (2) is related to the evidence accumulation framework [12]. Indeed, it can be re-written in terms of the so-called *co-association* matrix, which is defined as

$$\mathtt{C}_{ij} = \begin{cases} \frac{1}{\mathtt{N}_{ij}} \sum_{u\in\mathcal{U}_{ij}} \mathbb{1}_{\phi_u(i)=\phi_u(j)} & \mathtt{N}_{ij} > 0\,, \\ 0 & \text{otherwise}\,, \end{cases} \tag{4}$$

where $\mathcal{U}_{ij} \subseteq \{1,\ldots,\mathsf{m}\}$ denotes the indices of those clusterings, where both data points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ have been clustered, *i.e.* $\mathcal{U}_{ij} = \{u \in \{1,\ldots,\mathsf{m}\} : i,j \in \mathcal{J}_u\}$,

and $\mathtt{N}$ is a matrix with entries $\mathtt{N}_{ij} = |\mathcal{U}_{ij}|$ if $i \neq j$, and 0 otherwise. The relation with the co-association matrix is established as follows:

$$\begin{aligned}
\sum_{u=1}^{\mathsf{m}} d(\hat{\phi}, \phi_u) &= \sum_{u=1}^{\mathsf{m}} \sum_{i,j \in \mathcal{J}_u} \left[\mathbb{1}_{\hat{\phi}(i)=\hat{\phi}(j)} - \mathbb{1}_{\phi_u(i)=\phi_u(j)} \right]^2 \\
&= \sum_{i,j \in \mathcal{I}} \sum_{u \in \mathcal{U}_{ij}} \left[\mathbb{1}_{\hat{\phi}(i)=\hat{\phi}(j)} + \mathbb{1}_{\phi_u(i)=\phi_u(j)} - 2\mathbb{1}_{\hat{\phi}(i)=\hat{\phi}(j)} \mathbb{1}_{\phi_u(i)=\phi_u(j)} \right] \\
&= \sum_{i,j \in \mathcal{I}} \left[\mathtt{N}_{ij} \mathbb{1}_{\hat{\phi}(i)=\hat{\phi}(j)} + \sum_{u \in \mathcal{U}_{ij}} \mathbb{1}_{\phi_u(i)=\phi_u(j)} - 2\mathbb{1}_{\hat{\phi}(i)=\hat{\phi}(j)} \sum_{u \in \mathcal{U}_{ij}} \mathbb{1}_{\phi_u(i)=\phi_u(j)} \right] \\
&= \sum_{i,j \in \mathcal{I}} \mathtt{N}_{ij} \left[\mathbb{1}_{\hat{\phi}(i)=\hat{\phi}(j)} + \mathtt{C}_{ij} - 2\mathbb{1}_{\hat{\phi}(i)=\hat{\phi}(j)} \mathtt{C}_{ij} \right] \\
&= \sum_{i,j \in \mathcal{I}} \mathtt{N}_{ij} \left[\mathbb{1}_{\hat{\phi}(i)=\hat{\phi}(j)} - \mathtt{C}_{ij} \right]^2 + \sum_{i,j \in \mathcal{I}} \mathtt{N}_{ij} \mathtt{C}_{ij} (1 - \mathtt{C}_{ij}) \,. \qquad (5)
\end{aligned}$$

Note that the right-most term in (5) is regarded as a constant for the optimization in (2), thus not affecting the minimizers.

With the objective of re-writing (2) in matrix form, we introduce a different, but equivalent, representation of the consensus partition in terms of a matrix $\mathtt{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{\mathsf{n}}] \in \mathcal{S}_{01}^{\mathsf{k} \times \mathsf{n}}$, where $\mathcal{S}_{01}^{\mathsf{k} \times \mathsf{n}}$ denotes the set of binary, left-stochastic matrices. The equivalence follows from the fact that any $\phi \in \mathcal{I} \to \{1, \ldots, \mathsf{m}\}$ has a one-to-one corresponding matrix $\mathtt{Z} \in \mathcal{S}_{01}^{\mathsf{k} \times \mathsf{n}}$ with $(\mathtt{Z}_{ki} = 1) \iff (\phi(i) = k)$. Under this variable change, the term $\mathbb{1}_{\phi(i)=\phi(j)}$ becomes $\boldsymbol{z}_i^\top \boldsymbol{z}_j$.

By exploiting the matrix representation and the relation in (5), the semi-supervised consensus clustering formulation in (2) can be cast into the following equivalent one (with omitted constant terms):

$$\begin{aligned}
\mathtt{Z}^\star \in \arg\min_{\mathtt{Z} \in \mathcal{S}_{01}^{\mathsf{k} \times \mathsf{n}}} & \|\mathtt{C} - \mathtt{Z}^\top \mathtt{Z}\|_{\mathtt{N}}^2 \\
\text{s.t. } & \mathtt{Z}_{\ell_i i} = 1 \text{ for all } i \in \mathcal{L} \,,
\end{aligned} \qquad (6)$$

where $\|\cdot\|_{\mathtt{N}}$ is the Frobenious matrix norm weighted by $\mathtt{N}$, *i.e.* $\|\mathtt{A}\|_{\mathtt{N}} = \sqrt{\sum_{ij} \mathtt{N}_{ij} \mathtt{A}_{ij}^2}$. The optimization problem in (6) is non-convex and finding a global solution is hard. For this reason we opt for a relaxed version of it with the binary-valued matrix variable $\mathtt{Z} \in \mathcal{S}_{01}^{\mathsf{k} \times \mathsf{n}}$ being replaced with real-valued one $\mathtt{Y} \in \mathcal{S}^{\mathsf{k} \times \mathsf{n}}$, where $\mathcal{S}^{\mathsf{k} \times \mathsf{n}}$ denotes the set of real, left-stochastic, matrices, *i.e.* nonnegative matrices with columns summing up to 1. The relaxed optimization problem becomes

$$\begin{aligned}
\mathtt{Y}^\star \in \arg\min_{\mathtt{Y} \in \mathcal{S}^{\mathsf{k} \times \mathsf{n}}} & \|\mathtt{C} - \mathtt{Y}^\top \mathtt{Y}\|_{\mathtt{N}}^2 \\
\text{s.t. } & \mathtt{Y}_{ki} = \mathbb{1}_{k=\ell_i} \text{ for all } (k, i) \in \{1, \ldots, \mathsf{k}\} \times \mathcal{L} \,.
\end{aligned} \qquad (7)$$

Given a solution $\mathtt{Y}^*$ we recover a putative solution $\phi^\star$ to (2) by taking $\phi^\star(i) \in \arg\max_{k \in \{1, \ldots, \mathsf{k}\}} \mathtt{Y}_{ki}^\star$.

In order to optimize (7) we follow an approach similar to [23,24] by making use of a result known as *Baum-Eagon inequality*:

Theorem 1 (Baum-Eagon [5]). *Let* $\mathtt{Y} \in \mathcal{S}^{\mathtt{k}\times\mathtt{n}}$ *and let* $f(\mathtt{Y})$ *be a homogeneous* [4] *polynomial in the variables* $\mathtt{Y}_{ki}$ *with nonnegative coefficients. Define the mapping* $\hat{\mathtt{Y}} = M(\mathtt{Y})$ *as follows:*

$$\hat{\mathtt{Y}}_{ki} = \mathtt{Y}_{ki}\frac{\partial}{\partial \mathtt{Y}_{ki}} f(\mathtt{Y}) \Big/ \sum_{h=1}^{\mathtt{k}} \mathtt{Y}_{hi}\frac{\partial}{\partial \mathtt{Y}_{hi}} f(\mathtt{Y}) \tag{8}$$

for all $i \in \{1,\ldots,\mathtt{n}\}$ *and* $k \in \{1,\ldots,\mathtt{k}\}$. *Then* $f(M(\mathtt{Y})) > f(\mathtt{Y})$ *unless* $M(\mathtt{Y}) = \mathtt{Y}$. *In other words,* M *is a growth transformation for the polynomial* f.

The Baum-Eagon inequality is an effective tool for the maximization of polynomial functions in probability domain. The idea is to rewrite (7) into a maximization of a polynomial with nonnegative coefficients in a way to preserve $\mathtt{Y}^\star$ as the optimal solution. By doing so, we can use the Baum-Eagon inequality to obtain a step-size-free optimizer by re-iterating the update rule $\mathtt{Y}^{t+1} = M(\mathtt{Y}^t)$ starting from a matrix $\mathtt{Y}^0 \in \mathcal{S}^{\mathtt{k}\times\mathtt{n}}$ with positive entries. The theorem indeed guarantees a strict increase of the objective at each step until a fixed-point is reached.

We can turn (7) into a maximization problem by changing the sign of the objective function. However, the resulting function in $\mathtt{Y}$ will not be a polynomial with nonnegative coefficients. Nevertheless, there is a trick that can be exploited to transform the problem in the desired form. We will use the fact that $\mathtt{E}_\mathtt{n} = \mathtt{Y}^\top\mathtt{E}_\mathtt{k}\mathtt{Y}$ for any $\mathtt{Y} \in \mathcal{S}^{\mathtt{k}\times\mathtt{n}}$, where $\mathtt{E}_\mathtt{n}$ denotes a $\mathtt{n}\times\mathtt{n}$ matrix of ones:

$$\begin{aligned}
-\|\mathtt{C} - \mathtt{Y}^\top\mathtt{Y}\|_\mathtt{N}^2 &= -\|\mathtt{Y}^\top\mathtt{Y}\|_\mathtt{N}^2 + 2\langle \mathtt{C}, \mathtt{Y}^\top\mathtt{Y}\rangle_\mathtt{N} + \text{const} \\
&= \underbrace{\|\mathtt{Y}^\top\mathtt{E}_\mathtt{k}\mathtt{Y}\|_\mathtt{N}^2 - \|\mathtt{E}_\mathtt{n}\|_\mathtt{N}^2}_{=0} - \|\mathtt{Y}^\top\mathtt{Y}\|_\mathtt{N}^2 + 2\langle \mathtt{C}, \mathtt{Y}^\top\mathtt{Y}\rangle_\mathtt{N} + \text{const} \\
&= \|\mathtt{Y}^\top(\mathtt{E}_\mathtt{k} - \mathtt{I})\mathtt{Y}\|_\mathtt{N}^2 + 2\langle \mathtt{C}, \mathtt{Y}^\top\mathtt{Y}\rangle_\mathtt{N} + \text{const}\,,
\end{aligned} \tag{9}$$

where $\langle \mathtt{A}, \mathtt{B}\rangle_\mathtt{N} = \sum_{ij} \mathtt{N}_{ij}\mathtt{A}_{ij}\mathtt{B}_{ij}$ is a weighted matrix dot product, and $\mathtt{I}$ is a properly-sized identity matrix. Note that in the derivation "const" represents additive terms not depending on the variable $\mathtt{Y}$, thus not affecting the optimization results. We can now take the quantity in (9) (with constant terms omitted) as the polynomial f with nonnegative coefficients to be maximized, *i.e.*

$$f(\mathtt{Y}) = \|\mathtt{Y}^\top(\mathtt{E}_\mathtt{k} - \mathtt{I})\mathtt{Y}\|_\mathtt{N}^2 + 2\langle \mathtt{C}, \mathtt{Y}^\top\mathtt{Y}\rangle_\mathtt{N}\,.$$

and maximizers of $f(\mathtt{Y})$ on the feasible set of (7) will correspond to minimizers of (7) as required.

The last thing to care about is that Theorem 1 assumes $\mathtt{Y} \in \mathcal{S}^{\mathtt{k}\times\mathtt{n}}$, but our feasible domain is a convex subset thereof, due to the integration of the supervisions. Hence, it is not clear whether the theorem applies also to the constrained setting we have. To show that the theorem actually does apply,

[4] The same result was proven to hold also in the case of non-homogeneous polynomials [6].

assume without loss of generality that the labeled data points are the last ones, such that we can write $\mathtt{Y} = [\mathtt{Y}^u, \mathtt{Y}^l]$, where $\mathtt{Y}^l$ is entirely specified with the label information as $\mathtt{Y}^l_{ki} = \mathbb{1}_{k=\ell_i}$ for all $i \in \mathcal{L}$, while $\mathtt{Y}^u \in \mathcal{S}^{\mathtt{k}\times(\mathtt{n}-|\mathcal{L}|)}$ are variables to be optimized. Since $g(\mathtt{Y}^u) = f(\mathtt{Y})$ is a polynomial in $\mathtt{Y}^u$ with nonnegative coefficients, we can apply Theorem 1 to obtain a growth transformation for g, and thus find a solution also to the constrained optimization problem in (7). In practice, it is not necessary to compute g explicitly for the optimization, because it is sufficient to avoid updating the labeled entries of $\mathtt{Y}$ during the computation of the update rule based on f.

4 Experiments

In this section we validate the proposed algorithm (SSCC) both on standard benchmark datasets and on a real world problem, namely the pathology classification of ECG data. In the architecture described in Figure 2, the raw ECG is acquired and preprocessed to extract relevant features, which are then fed to the algorithms to classify the pathologies. We adopt the feature extraction process proposed in [4], which is described in Section 4.2.

4.1 Data Description

We evaluated the proposed algorithm in two different scenarios: using some benchmark datasets from the UCI Machine Learning repository[5], and using the MIT-BIH (Massachusetts Institute of Technology - Beth Israel Hospital) arrhythmia database [21].

In the first scenario, we used four benchmark datasets to validate our algorithm, namely breast cancer, iris, wine and std yeast cell. The *Breast cancer* dataset consists of 683 patterns having nine features belonging to two classes. The *Iris* dataset consists of three species of Iris plants, characterized by four features and 50 samples in each class. The *Wine* dataset consists of the results of a chemical analysis of wines grown from the same region in Italy divided into three classes with 178 patterns, and described by 13 features. The *Std yeast cell* is composed by 384 genes over two cell cycles of yeast cell data, and is characterized by 17 features and it has five classes.

The second scenario consists in classifying pathologies in ECGs from the MIT-BIH arrhythmia database. Each record is approximately 30 minutes long and has in total 48 two-channel Holter records. The upper signal is usually a modified limb lead II and the lower signal is most often a modified lead V1. All signals were digitized at a sample rate of 360 Hz. The database includes different sets of annotations verified by more than one cardiologist: beats are identified and labeled, and the beginning of all rhythms is indicated.

[5] http://archive.ics.uci.edu/ml/

4.2 Feature Extraction

For the classification of pathologies in ECGs, we will focus on the discrimination between normal sinus and the most common arrhythmia, atrial fibrillation (AF). Only modified limb II records are used and each record is split according to rhythm annotations (note that lead II and lead I, introduced in section 2, contain information about the frontal plane, and can be used to detect atrial fibrillation). Each record is segmented in windows of 60 seconds, leading to 98 AF segments and 911 normal sinus rhythm segments. These segments are the objects that will be classified. Two types of features are obtained: spectral features extracted using the wavelet transform, and time domain features used to provide information about heart rate characteristics.

The spectral features were obtained by the power spectral density (PSD) of the wavelet decomposition of the signals. The decomposition of the signals are performed up to the sixth level using the redundant discrete wavelet transform [11], obtaining six detailed and one approximated set of coefficients. Afterwards, the PSD of each set of wavelets coefficients was estimated using Welch's method [27], and the integral over the range $[0, 55]$ Hz was computed, leading in total to seven features per pattern.

Besides the spectral features we considered two additional time-domain features: average of RR interval [7] and standard deviation of RR intervals.

4.3 Setup

We constructed the ensemble for the proposed methodology by performing 200 runs of k-means, with k uniformly chosen between $\sqrt{N}/2$ and $\sqrt{N}$ (N is the number of samples of the dataset), for the benchmark datasets, and between 2 and 20, for the MIT-BIH database. Since the labels of the points are only used to extract the consensus partition, we present also the results of an unsupervised consensus clustering method (PPC) [23].

We compared SSCC also against a semi-supervised consensus clustering approach called Bipartite Graph-based Consensus Maximization (BGCM) [14], which can be seen as a consensus method, where the ensemble is constructed with supervised and unsupervised methods. In this paper we constructed in total six partitions in the ensemble: three from supervised methods, namely k-nearest neighbor, with $k \in \{1, 3, 5\}$, and three from an unsupervised method, the k-means, with k equal to the true number of classes. Moreover, this algorithm has two parameters, α and β, corresponding to the price paid from deviating from the estimated labels of groups and observed labels of objects. We set those parameters to 2 and 8, respectively.

In the MIT-BIH database, the classes are unbalanced, so we randomly selected 98 out of the 911 normal sinus rhythm segments. Also, in order to test the influence of the percentage of labeled points in each algorithm, we randomly selected 5% of labeled points in each class and the remaining are unlabeled points to be classified by the algorithms. This procedure was repeated 50 times, and was also run with minimum 10% and maximum 60% labeled data points.

Thus, we assessed the performance of each algorithm as an average error rate of 50 runs. The same scheme was applied to the benchmark datasets from the UCI Machine Learning repository, creating datasets with balanced classes.

4.4 Validation of the Algorithm: Benchmark Datasets

Figure 3 presents the average error rates for the four benchmark datasets considered. For SSCC and BGCM, we only present the results for 10% and 20% of labeled points, since as we increase the percentage of labeled points, there is a decrease in error rates. PPC is an unsupervised approach, which means it does not require any labeled points, thus the error remains constant when we increased the percentage of labeled points.

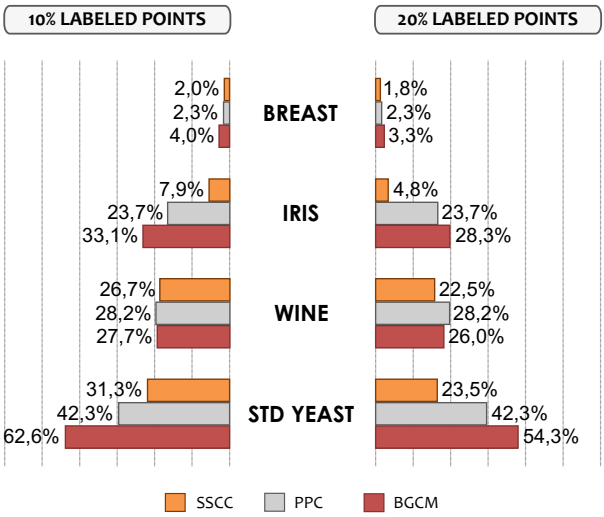


Fig. 3. Average error rates of 50 runs.

Notice that, for 10% of labeled points, SSCC achieves approximately 8% of error rate for the Iris dataset, performing three or four times better than the unsupervised version (which does not use any labeled points) and BGCM. That difference is even more accentuated when we have 20% of labeled points, where SSCC achieves less than 5% of error rate. The higher difference in error rates between SSCC and the other two algorithms is also visible in the Std yeast dataset, although a little less accentuated. In the remaining datasets, SSCC is still the best algorithm, however it is only better 1% or 2% than PPC and BGCM. Overall, the proposed methodology performs better than the other two algorithms considered in these experiments.

4.5 Application on Real Datasets: ECG MIT-BIH Database

Figure 4 presents the results of applying each algorithm to the MIT-BIH arrhythmia dataset, when the percentage of labeled points are varying. As can be observed by analyzing the figure, when the percentage of labeled points is of only 5%, SSCC and PPC have similar performances (SSCC shows slightly better results). On the other hand, BGCM has an error rate of approximately 17%, which is almost twice the error rate of SSCC. Moreover, as it could be expected, when the percentage of labeled points increases, both semi-supervised approaches show a significant improvement in the error rates, achieving around 3% of misclassified points when 60% of the dataset is labeled. In particular, when 30% or more points are labeled, the two semi-supervised approaches are quite similar, on average. However, by carefully analyzing the standard deviation, it is possible to conclude that BGCM has a more unstable behavior, since it presents much higher standard deviation values in comparison with SSCC.

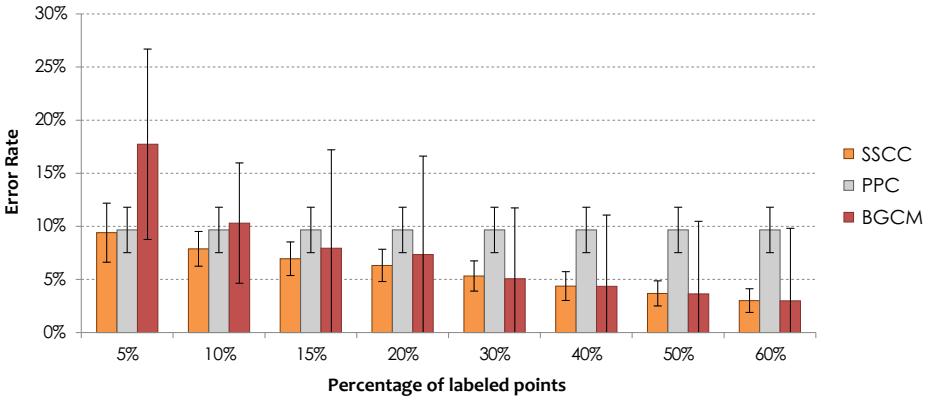
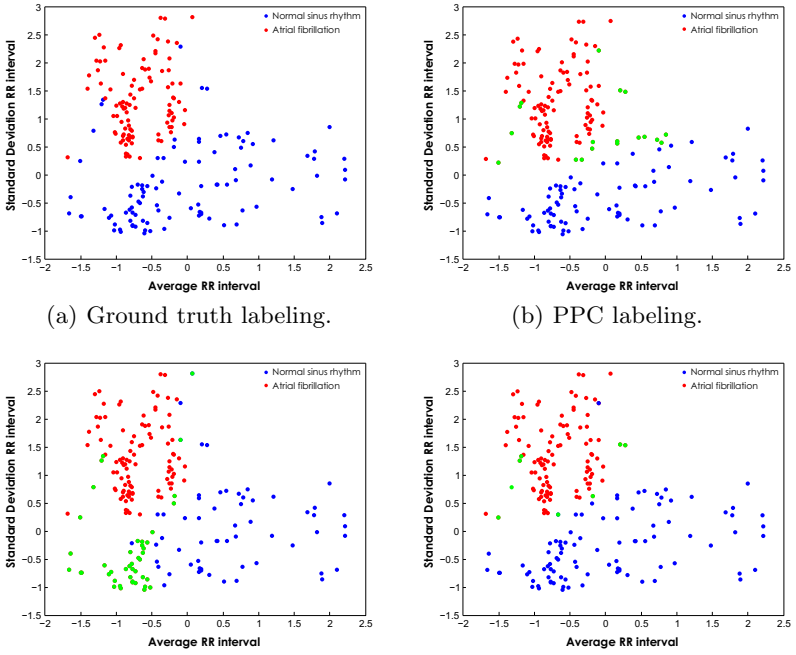


Fig. 4. Average error rates and standard deviation of 50 runs on the MIT-BIH arrhythmia database.

In [4], a supervised study was conducted using this database, and using 75% of data to train the classifier and 25% to test, the 1-nearest neighbor has an error rate of $2.6\% \pm 1.1\%$, and an artificial neural network with 14 hidden neurons has $3.0\% \pm 1.2\%$. The results presented in this study are quite comparable with the supervised study, since with only 20% of labeled patterns, SSCC has an error rate of $6.3\% \pm 1.5\%$, and with 60% of labeled patterns, SSCC has achieved $3.0\% \pm 1.1\%$.

Unlike normal sinus rhythm, atrial fibrillation has an inherent irregular RR interval, allowing for an easy visualization of the dataset. Accordingly, figure 5 presents the two time domain features considered in this study. The labeling produced by each algorithm corresponds to one run out of the 50 runs, and for the BGCM and SSCC, we are fixing the amount of labeled patterns at 10%. Notice

that PPC incorrectly labeled the patterns in the frontier of both classes and the normal patterns that are mixed in the atrial fibrillation class. In fact, that couple of patterns that are mixed with the atrial fibrillation are incorrectly classified by SSCC and, some of them, by BGCM. On the other hand, SSCC correctly classified almost all the patterns in the frontier of both classes. The BGCM algorithm incorrectly classified a large amount of normal sinus rhythm in the lower left corner, due to k-means initialization. Notice that, for this algorithm, the number of components k in k-means must be set to the true number of classes, so that cloud of green points was assigned to atrial fibrillation. The blue dot in the middle corresponds to a pattern with the true labeled known.



(a) Ground truth labeling. (b) PPC labeling.

(c) BGCM labeling. (d) SSCC labeling.

Fig. 5. Scatterplot of two features (average and standard deviation of RR intervals) from the ECG MIT-BIH dataset, when we have 10% of labeled patterns. The blue dots correspond to the normal sinus, the red ones to the atrial fibrillation, and the green dots are the points incorrectly classified by each algorithm.

Moreover, the SSCC algorithm incorrectly classified a few patterns that are closer to the atrial fibrillation class, the same patterns were identified by PPC. Since those patterns are so close to the atrial fibrillation, it may be worth to re-analyze those segments to ensure that they are in fact normal rhythms, and do not correspond to abnormal patterns closely resembling normal ones.

5 Conclusions

In real world problems it is unfeasible to ask experts to annotate all the available data. This is particularly true for the real-world scenario addressed in this paper, namely the automatic pathology classification of electrocardiographic (ECG) signals, where the amount of annotated data is very small compared to the amount of available data.

In this paper we proposed a semi-supervised consensus clustering algorithm, which automatically allows to label the unknown objects using only a small subset of known information. Our approach is based on the evidence accumulation clustering, a consensus clustering paradigm that summarizes the ensemble information into a co-association matrix. In contrast to the typical algorithms in this context, which are unsupervised, we allowed the partial inclusion of labeled information. Algorithmically, we have provided a simple iterative scheme based on the Baum-Eagon inequality for the computation of the consensus partition.

We validated our approach on benchmark datasets from UCI Machine Learning repository, showing superior performance against another state-of-the-art semi-supervised consensus clustering approach (BGCM), in every dataset that we considered under different shares of supervision (10 and 20 %). We also considered the real-world application of automatic pathology classification of ECG signals. Specifically, we considered the detection of atrial fibrillation. We have used the MIT-BIH arrhytmia dataset, varying the percentage of labeled data. The performance of the proposed algorithm was always better than BGCM and achieved comparable performance with respect to supervised learning methods using as few as 20% of labeled objects.

The proposed approach will be deployed by our industrial partner in a cloud-based implementation, allowing ECG data acquired using pervasive healthcare devices, such as the keyboard, to be automatically processed.

As future work, we are developing a scalable version of the algorithm, allowing to tackle the large amount of data that is being produced in this context.

Acknowledgments. This work was supported by the Portuguese Foundation for Science and Technology, scholarship number SFRH/BPD/103127/2014, and grants PTDC/EEI-SII/2312/2012 and PEst-OE/EEI/LA0008/2013.

References

1. Aidos, H., Fred, A.: Consensus of clusterings based on high-order dissimilarities. In: Celebi, M.E. (ed.) Partitional Clustering Algorithms, pp. 313–351. Springer International Publishing (2015)
2. Basu, S., Banerjee, A., Mooney, R.J.: Active semi-supervision for pairwise constrained clustering. SDM **4**, 333–344 (2004)
3. Basu, S., Davidson, I., Wagstaff, K.: Constrained Clustering: Advances in Algorithms, Theory, and Applications. Chapman and Hall/CRC (2008)
4. Batista, D., Fred, A.: Spectral and time domain parameters for the classification of atrial fibrillation. In: International Conference on Bio-inspired Systems and Signal Processing (BIOSIGNALS 2015), pp. 329–337 (2015)

5. Baum, L.E., Eagon, J.A.: An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. Bull. American Math. Society **73**, 360–363 (1967)
6. Baum, L.E., Sell, G.R.: Growth transformations for functions on manifolds. Pacific J. Math. **27**, 221–227 (1968)
7. Chung, E.K.: Pocket guide to ECG diagnosis. Blackwell Science (1996)
8. Covoes, T.F., Hruschka, E., Ghosh, J.: A study of k-means-based algorithms for constrained clustering. Intelligent Data Analysis **17**, 485–505 (2013)
9. Drew, B.J., Califf, R.M., Funk, M., Kaufman, E.S., Krucoff, M.W., Laks, M.M., Macfarlane, P.W., Sommargren, C., Swiryn, S., Van Hare, G.F.: Practice standards for electrocardiographic monitoring in hospital settings an american heart association scientific statement from the councils on cardiovascular nursing, clinical cardiology, and cardiovascular disease in the young. Circulation **110**(17), 2721–2746 (2004)
10. Fern, X.Z., Brodley, C.E.: Solving cluster ensemble problems by bipartite graph partitioning. In: Proceedings of the Twenty-first International Conference on Machine Learning, ICML 2004, p. 36. ACM, New York (2004). http://doi.acm.org/10.1145/1015330.1015414
11. Fowler, J.E.: The redundant discrete wavelet transform and additive noise. Signal Processing Letters, IEEE **12**(9), 629–632 (2005)
12. Fred, A., Jain, A.: Combining multiple clusterings using evidence accumulation. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(6), 835–850 (2005)
13. Fred, A.L., Lourenço, A., Aidos, H., Rota Bulò, S., Rebagliati, N., Figueiredo, M., Pelillo, M.: Learning similarities from examples under the evidence accumulation clustering paradigm. In: Pelillo, M. (ed.) Similarity-Based Pattern Analysis and Recognition. Advances in Computer Vision and Pattern Recognition, pp. 85–117. Springer London (2013)
14. Gao, J., Liang, F., Fan, W., Sun, Y., Han, J.: Graph-based consensus maximization among multiple supervised and unsupervised models. In: Advances in Neural Information Processing Systems, pp. 585–593 (2009)
15. Kuncheva, L., Vetrov, D.: Evaluation of stability of k-means cluster ensembles with respect to random initialization. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(11), 1798–1808 (2006)
16. Li, T., Ding, C., Jordan, M.: Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization. In: dSeventh IEEE International Conference on Data Mining, ICDM 2007, pp. 577–582, October 2007
17. Lourenço, André, Bulò, Samuel Rota, Rebagliati, Nicola, Fred, Ana, Figueiredo, Mário, Pelillo, Marcello: Consensus clustering using partial evidence accumulation. In: Sanches, João M., Micó, Luisa, Cardoso, Jaime S. (eds.) IbPRIA 2013. LNCS, vol. 7887, pp. 69–78. Springer, Heidelberg (2013)
18. Lourenço, A., Rota Bulò, S., Rebagliati, N., Fred, A.L.N., Figueiredo, M.A.T., Pelillo, M.: Probabilistic evidence accumulation for clustering ensembles. In: Marsico, M.D., Fred, A.L.N. (eds.) ICPRAM 2013 - Proceedings of the 2nd International Conference on Pattern Recognition Applications and Methods, pp. 58–67. SciTePress (2013)
19. Lourenço, A., Rota Bulò, S., Rebagliati, N., Fred, A.L., Figueiredo, M.A., Pelillo, M.: Probabilistic consensus clustering using evidence accumulation. Machine Learning **98**(1–2), 331–357 (2015)
20. Lourenço, A., Silva, H., Fred, A.: Unveiling the biometric potential of Finger-Based ECG signals. Computational Intelligence and Neuroscience 2011 (2011)

21. Moody, G.B., Mark, R.G.: The impact of the MIT-BIH arrhythmia database. Engineering in Medicine and Biology Magazine, IEEE **20**(3), 45–50 (2001)
22. Odinaka, I., Lai, P.H., Kaplan, A., O'Sullivan, J., Sirevaag, E., Rohrbaugh, J.: ECG biometric recognition: A comparative analysis. IEEE Trans. Inf. Forensics Security **7**(6), 1812–1824 (2012)
23. Rota Bulò, S., Lourenço, A., Fred, A., Pelillo, M.: Pairwise probabilistic clustering using evidence accumulation. In: Hancock, E., Wilson, R., Windeatt, T., Ulusoy, I., Escolano, F. (eds.) SPR and SPR 2010. LNCS, vol. 6218, pp. 395–404. Springer, Heidelberg (2010)
24. Rota Bulò, S., Pelillo, M.: Probabilistic clustering using the baum-eagon inequality. In: ICPR, pp. 1429–1432 (2010)
25. da Silva, H.P., Carreiras, C., Lourenço, A., Fred, A., das Neves, R.C., Ferreira, R.: Off-the-person electrocardiography: performance assessment and clinical correlation. Health and Technology, 1–10 (2015)
26. Topol, E.: The Creative Destruction of Medicine. Basic Books (2012)
27. Welch, P.: The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Transactions on audio and electroacoustics, 70–73 (1967)