A Clustering Algorithm Based on an Ensemble of Dissimilarities: An Application in the Bioinformatics Domain

Manuel Martín Merino*, Alfonso José López Rivero*, Vidal Alonso, Marcelo Vallejo, Antonio Ferreras

Computer Science School, Universidad Pontificia de Salamanca, Salamanca (Spain)

Received 7 May 2022 | Accepted 4 July 2022 | Early Access 19 September 2022

Abstract

Clustering algorithms such as k-means depend heavily on choosing an appropriate distance metric that reflect accurately the object proximities. A wide range of dissimilarities may be defined that often lead to different clustering results. Choosing the best dissimilarity is an ill-posed problem and learning a general distance from the data is a complex task, particularly for high dimensional problems. Therefore, an appealing approach is to learn an ensemble of dissimilarities. In this paper, we have developed a semi-supervised clustering algorithm that learns a linear combination of dissimilarities considering incomplete knowledge in the form of pairwise constraints. The minimization of the loss function is based on a robust and efficient quadratic optimization algorithm. Besides, a regularization term is considered that controls the complexity of the distance metric learned avoiding overfitting. The algorithm has been applied to the identification of tumor samples using the gene expression profiles, where domain experts provide often incomplete knowledge in the form of pairwise constraints. We report that the algorithm proposed outperforms a standard semi-supervised clustering technique available in the literature and clustering results based on a single dissimilarity. The improvement is particularly relevant for applications with high level of noise.

Keywords

Bioinformatics, Clustering, Kernel Methods, Machine Learning, Metric Learning.

DOI: 10.9781/ijimai.2022.09.007

I. Introduction

CLUSTERING algorithms such as k-means depend heavily on finding an appropriate dissimilarity that reflects accurately the object proximities [1]. This depends on the nature of the data and project requirements [2]. In practice, a wide range of dissimilarities may be defined based for instance on different features of the objects [3], [4]. Different dissimilarities lead often to significant changes in clustering results. Some researchers have addressed this problem learning a general distance from the data [5] but this is a challenging task for high dimensional applications [6]. Therefore, instead of considering a single distance metric an appealing approach is to learn a combination of dissimilarities from the data.

Several authors have developed learning algorithms for multiview clustering that are able to integrate a set of dissimilarities obtained from different features of the objects [7]. Following the same approach [8], Hu et al. [9] have proposed multiple kernel k-means clustering algorithms that might consider a set of dissimilarities using the kernel trick. However, these learning algorithms are unsupervised and may not provide metrics that help to increase the cluster separability [6].

* Corresponding author.

E-mail addresses: mmartinmac@upsa.es (M. Martín Merino), ajlopezri@upsa.es (A. López).

For certain Bioinformatics applications, weak supervised information is available in the form of which pairs of proteins or genes are related [10]. This incomplete supervision may be incorporated into semi-supervised clustering algorithms formulated as pair-wise constraints [11], [12]. Must-link constraints when x_i and x_j belong to the same cluster and cannot-link constraints when x_i and x_j belong to different clusters.

Some researchers have proposed algorithms to learn the metric from a set of equivalence constraints based on the Mahalanobis distance [1], [6], [13]. However, they are based on a single metric that may not be appropriate for certain applications and do not perform well with high dimensional data with noise. Besides, they are prone to overfitting and are computationally intensive due to the large number of parameters involved. Other non-linear metric learning approaches have been developed based on kernel methods [14], [15]. Again they are based on a single dissimilarity and suffer from similar drawbacks.

In this paper, we follow the approach of multiple kernel clustering algorithms [16], [17], that learn a combination of kernels to improve the clustering results. However, this kind of researches relies on complex optimization algorithms and often are not designed to incorporate supervised information in the form of pairwise constraints. The main contribution of this paper is to propose a novel semi-supervised clustering algorithm that learns an ensemble of dissimilarities from incomplete knowledge in the form of pairwise constraints. The problem is formulated as learning the combination of

multiple kernels (similarities) that maximizes the separability among the clusters considering the pairwise constraints. The loss function is convex and quadratic without local minima and it is optimized in dual space efficiently. Besides, it incorporates a penalty term to control the complexity of the family of distances avoiding the overfitting.

The algorithm has been evaluated using several benchmark UCI data sets and two problems of cancer samples identification based on the gene expression profiles. The empirical results suggest that the method proposed improves the clustering results obtained considering a single dissimilarity and a standard supervised clustering method proposed by Xing et al. [13] that learns the metrics from pairwise constraints.

This paper is organized as follows: Section II presents the clustering algorithm proposed that learns a combination of dissimilarities using pairwise constraints. Section III illustrates the performance of the algorithm using several benchmark and two complex cancer samples identification datasets. Section IV discusses the contributions of this paper in the context of related work. Finally, Section V gets conclusions and outlines future research trends.

II. Material and Methods

In this section we present the semi-supervised clustering algorithm developed based on an ensemble of dissimilarities and the experimental datasets considered. First, sections A and B introduce the kernel version of k-means clustering algorithm and the empirical kernel map, that allow us to extend a kernel clustering algorithm to work with a given dissimilarity. Thus, the problem of learning a linear combination of dissimilarities may be formulated as learning a linear combination of kernels. Next, in section C an idealized kernel is defined for clustering applications that helps to reduce the intra-cluster distances while increasing the inter-cluster separability considering the available pairwise constraints. Section D presents the learning algorithm for the linear combination of kernels that best approximate the idealized kernel, subject to a set of pairwise constraints. Section E comments the meaning of the non-null Lagrange multipliers in the dual space as support vectors. Finally, section F describes the features of the benchmark and cancer datasets considered.

A. Kernel K-means Clustering

Let $X = \{x_1, x_2, \ldots, x_n\} \in \Re^d$ be the training set, Z_{ki} the clustering indicator matrix defined as 1 if x_i belong to cluster k and 0 otherwise. k-means clustering looks for a set of representatives $\{c_k\}_{k=1}^{C}$ and a partition of the objects into C groups that minimize the sum of square distances to the cluster representatives:

$$\min_{Z\in[0,1]} \sum_{k=1}^{C}\sum_{i=1}^{n} Z_{ki}\|\mathbf{x}_i - \mathbf{c}_k\|^2 \tag{1}$$

$$\text{s. t} \quad \sum_{k=1}^{C} Z_{ki} = 1 \tag{2}$$

This error function is optimized by an iterative algorithm in two steps. First the centroids for each cluster are computed, next each object is assigned to the group corresponding to the nearest centroid according to the euclidean distance. The use of the euclidean distance induces a bias towards spherical groups. K-means clustering has been extended to more general dissimilarities by mapping non-linearly the original samples to a high dimensional reproducing kernel Hilbert space $\mathcal{F}$ [9]. Let Φ be the non-linear mapping to feature space $\mathcal{H}$. Kernel k-means optimizes the following sum of square errors in the reproducing kernel Hilbert space:

$$\min_{Z\in[0,1]} \sum_{k=1}^{C}\sum_{i=1}^{n} Z_{ki}\|\Phi(\mathbf{x}_i) - \mathbf{c}_k\|_{\mathcal{H}}^2 \tag{3}$$

$$\text{s. t} \quad \sum_{k=1}^{C} Z_{ki} = 1 \tag{4}$$

where $c_k = \frac{1}{n_k}\sum Z_{ki}\,\Phi(x_i)$ is the centroid for cluster k in the kernel feature space. Considering that in this feature space $\Phi^T(x_i)\Phi(x_j) = K(x_i, x_j)$, the L_2 norm can be written exclusively in terms of kernels evaluations as:

$$\|\Phi(\mathbf{x}_i) - \mathbf{c}_k\|_{\mathcal{H}}^2 = K(\mathbf{x}_i,\mathbf{x}_i) - \frac{2}{n_k}\sum_{j=1}^{n} Z_{kj}K(\mathbf{x}_i,\mathbf{x}_j) + \frac{1}{n_k^2}\sum_{j=1}^{n}\sum_{l=1}^{n} Z_{kj}Z_{kl}K(\mathbf{x}_j,\mathbf{x}_l) \tag{5}$$

The optimization of the square error function (3) in the feature space can be solved by algorithm 1.

Algorithm 1. Kernel k-means algorithm

1: **Inputs** K: kernel matrix, C: number of clusters

2: ***Initialize***: The C clusters $C_1^{(0)}, \ldots, C_C^{(0)}$

3: Set t = 0

4: For each x_i compute the cluster with the nearest centroid: $k^*(\mathbf{x}_i) = \text{argmin}_K \|\Phi(\mathbf{x}_i) - \mathbf{c}_k\|^2$ using (5)

5: Update the clusters $C_k^{(t+1)} = \{\mathbf{x}_i \mid k^*(\mathbf{x}_i) = k\}$

6: Go to step 3 and update t = t + 1 if not converged

7: **Return**: $C_1, \ldots, C_C$ partitioning of the objects

B. The Empirical Kernel Map

We have mentioned earlier that the learning algorithm for the kernel k-means clustering can be written exclusively in terms of kernel evaluations. For certain applications only a dissimilarity matrix is available and it is often difficult to obtain a vectorial representation for the data. Therefore, the dissimilarity should be incorporated into the algorithm directly through the kernel definition. To this aim, we first map the dissimilarity to a feature space where the dot product defines a Mercer kernel [18]. Depending on the kernel definition, the map may transform linearly or non-linearly the original distance given rise to a wider family of dissimilarities. Next, we introduce the empirical kernel map proposed by [19].

Let $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a dissimilarity and $R = \{x_1, x_2, \ldots, x_n\}$ a subset of representatives drawn from the training set. The mapping to embed a given dissimilarity to a feature space is defined as:

$$\begin{aligned} \Phi: \mathcal{X} &\to \mathcal{F} \\ \mathbf{z} &\to \Phi(\mathbf{z}) = [\phi_1(\mathbf{z}), \phi_2(\mathbf{z}), \ldots, \phi_n(\mathbf{z})] \end{aligned} \tag{6}$$

where

$$\Phi(\mathbf{z}) = D(\mathbf{z}, R) = [d(\mathbf{z},\mathbf{x}_1), d(\mathbf{z},\mathbf{x}_2), \ldots, d(\mathbf{z},\mathbf{x}_n)] \tag{7}$$

This mapping Φ embeds the dissimilarity into a functional Hilbert space where feature j is given by $d(\,.\,, x_j)$. The number of representatives considered determines the dimensionality of the feature space. Now, the dot product in feature space defines the kernel for a given dissimilarity:

$$k(\mathbf{z},\mathbf{z}') = \langle\phi(\mathbf{z}),\phi(\mathbf{z}')\rangle = \sum_{j=1}^{n} d(\mathbf{z},x_j)d(\mathbf{z}',x_j)\ \forall \mathbf{z},\mathbf{z}' \in \mathcal{X} \tag{8}$$

An interesting property of the kernel matrix is that it is symmetric and positive semi-definite [18]. This characteristic will help to define a convex quadratic loss function for the clustering algorithm that can be optimized efficiently. Obviously, a clustering based on kernels can be extended easily to work with a given dissimilarity just considering the definition (8) for the kernel.

C. The Idealized Kernel of Dissimilarities

Let $\{x_i\}_{i=1}^n \in \Re^d$ be a set of objects. We are given weak supervised information to learn the distance metric in the form of similarity/ dissimilarity constraints. Must link constraints provide pairs of objects that are considered similar and cannot link constraints identify dissimilar ones. Let S and D be the subset of object pairs known to be similar/dissimilar. Mathematically they are defined as:

$$\mathcal{S} = \{(x_i, x_j) : x_i \text{ is similar to } x_j\} \tag{9}$$

$$\mathcal{D} = \{(x_i, x_j) : x_i \text{ is dissimilar to } x_j\} \tag{10}$$

Next, the idealized kernel is defined with the aim of maximizing the separability among different clusters. First, notice that the kernel function is a dot product in feature space $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ [18]. Therefore, it can be considered a similarity measure defined in the reproducing kernel Hilbert space [20]. For clustering applications, the ideal similarity (kernel) should be large for pairs of similar objects and small for dissimilar ones. Mathematically, the idealized kernel is defined for a family of kernels $\{k^l\}_{l=1}^M$ and a set of pairwise constraints (S, D) as follows:

$$K_{i,j}^* = k^*(x_i, x_j) = \begin{cases} max_l\{K_{ij}^l\} & \text{If } (x_i, x_j) \in \mathcal{S} \\ min_l\{K_{ij}^l\} & \text{If } (x_i, x_j) \in \mathcal{D} \end{cases} \tag{11}$$

Now, the idealized dissimilarity between two objects (x_i, x_j) is the euclidean distance in the kernel feature space induced by $\mathbf{k}^*$. Substituting the dot products by the idealized kernel, it can be written as:

$$\begin{aligned} d^2(x_i, x_j) &= \|\phi(x_i) - \phi(x_j)\|^2 \\ &= k^*(x_i, x_i) + k^*(x_j, x_j) - 2k^*(x_i, x_j) \end{aligned} \tag{12}$$

For pairs of similar objects the idealized dissimilarity takes the smallest value of the family of dissimilarities while for dissimilar ones takes the largest value. This measure will increase the cluster separability reducing the intra-cluster variance.

To illustrate the performance of the idealized dissimilarity let consider the breast cancer data set employed in the experimental section. We have applied a classical multidimensional scaling algorithm (MDS) [21] to project the data over a two dimensional subspace preserving approximately the original dissimilarities.

Fig. 1 shows the representation when the euclidean distance is considered and no supervisory information is available. The two classes (red-blue) overlap significantly and a clustering algorithm will fail to identify the two groups. Fig. 2 shows the projection for the MDS algorithm based on the idealized dissimilarity obtained from a family of 9 distances and a small set of randomly chosen pairwise constraints. We have considered 20% of all possible similarity/dissimilarity constraints. Similarity constraints are generated selecting pairs of patients that belong to the same class while dissimilarity constraints are retrieved from pairs of patients assigned to different classes. Fig. 2 shows that considering the idealized similarity both clusters become separable. Obviously, this measure may increase the overfitting. Therefore, the algorithm proposed to learn this dissimilarity should take care of this problem.

The idealized kernel (11) defined here for weak supervised clustering problems is related to the one proposed by [22] for classification:

$$k(x_i, x_j) = \begin{cases} 1 & if\ y_i = y_j \\ 0 & otherwise, \end{cases} \tag{13}$$

where y_i denotes the class label for x_i. However, the definition (13) takes into account only the class labels missing relevant information about the probability distribution for the objects. By contrast, the idealized kernel presented here takes into account a set of dissimilarity measures and hence, considers the probability distribution for the data. Besides, the kernel definition (13) is only valid for supervised problems in which class labels are available for the training set. It cannot be considered to incorporate incomplete knowledge in the form of equivalence constraints.

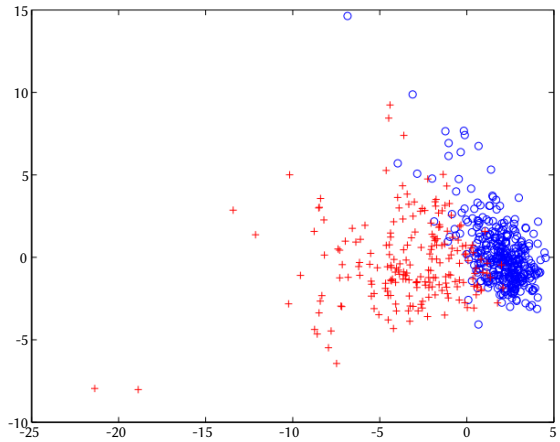

Fig. 1. Multidimensional scaling algorithm for a breast cancer dataset based on the euclidean distance. Both clusters (control and cancer) are quite overlapped in the projection.

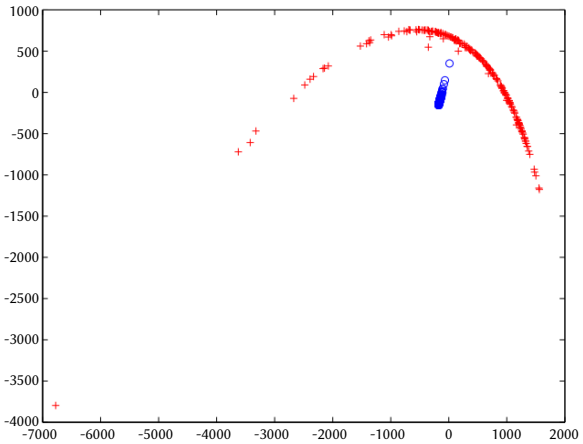

Fig. 2. Multidimensional scaling algorithm for a breast cancer dataset based on the idealized dissimilarity. Now, the two groups (control and cancer) can be easily identified by a clustering algorithm.

D. Multiple Kernel Learning for Clustering Algorithms Using Pairwise Constraints

In this section, we present the algorithm to learn the linear combination of similarities (kernels) that maximizes the cluster separability considering a set of pairwise constraints.

Let $\{d_{ij}^l\}_{l=1}^M$ be a set of M dissimilarity matrices that may come from different definitions or considering different features of the data. The dissimilarities are introduced into the clustering algorithm using the empirical kernel map (8). Let $\{k^l\}_{l=1}^M$ be the family of kernels obtained. Considering non-linear kernels will extend the original family of dissimilarities by non-linear mapping to a feature space. The problem can now be formulated as learning an optimal combination of kernels that maximizes the separability among the clusters using the pairwise constraints.

The linear combination of kernels is defined as:

$$k(x_i, x_j) = \sum_{l=1}^{M} \beta_l \, k^l(x_i, x_j) \tag{14}$$

where the β_l coefficients are constrained to be ≥ 0. Therefore, provided that each kernel is symmetric and positive semi-definite, the linear combination of kernels with $\beta_l \geq 0$ will be convex and positive semi-definite [23]. This property will help to define a convex quadratic loss function for the distance learning algorithm that may be optimized efficiently. Linear combination of kernels are preferred in this research over non-linear ones [4] because they are more robust to overfitting and the estimation of the parameters is more efficient computationally. The β_l coefficients are learned considering that he linear combination of kernels (14) should approximate the idealized kernel (11) with minimum error subject to the similarity/dissimilarity constraints. This optimization problem can be formulated in the primal as follows:

$$\min_{\beta,\xi} \Omega(\beta) + \frac{C_S}{N_S} \sum_{(x_i,x_j) \in S} \xi_{ij} + \frac{C_D}{N_D} \sum_{(x_i,x_j) \in D} \xi_{ij} \tag{15}$$

$$\text{s.t.} \quad \beta^T \mathbf{K}_{ij} \geq K^*_{ij} - \xi_{ij} \quad \forall (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S} \tag{16}$$

$$\beta^T \mathbf{K}_{ij} \leq K^*_{ij} - \xi_{ij} \quad \forall (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D} \tag{17}$$

$$\beta_l \geq 0 \quad \xi_{ij} \geq 0 \quad \forall l = 1, \ldots, M \tag{18}$$

C_S and C_D are regularization parameters that penalize training errors in the estimation of the idealized kernel. Particularly, they penalize similarity/dissimilarity constraint violations respectively. Both parameters may be determined by a grid search strategy using ten fold-crossvalidation. N_S, N_D are the number of pairwise constraints in S and D. $\Omega(\beta)$ is a regularization function that penalizes the complexity of the linear combination of kernels learned. Increasing the values of the regularization parameters C_S and C_D will minimize training errors in the constraints satisfaction but will increase the complexity of the similarity/kernel learned and the overfitting of the data. K^*_{ij} is the idealized kernel matrix introduced in section C and ξ_{ij} are the slack variables which are greater than zero for errors in the constraints satisfaction. Finally, K_{ij} is a matrix defined as $K_{ij} = \left[K^1_{ij}, K^2_{ij}, \ldots, K^M_{ij}\right]^T$, where K^l_{ij} is the idealized kernel matrix for similarity l.

The equations (16)-(17) model the constraints and ensure that the combination of similarities/ kernels learned are $\geq K^*_{ij}$ for similar objects and $\leq K^*_{ij}$ for dissimilar ones. The choice of the functional regularization term $\Omega(\beta)$ will determine the properties of the solution obtained. The L_1 norm is frequently considered in the Multiple Kernel Learning (MKL) literature [16], [24]. In this case, the solution will become sparse [25] and only a small set of similarities/kernels correlated with the idealized similarity will have non-null coefficient. However, in the bioinformatics applications considered in this paper, we are given frequently a small set of curated dissimilarities coming from different sources or distance metric definitions. Sparse solutions may lose relevant information and worsen the clustering results obtained [25].

Another choice for the regularization function $\Omega(\beta)$ is the L_2 norm. This penalization term distributes the weights more evenly reducing the value of the coefficients for less relevant kernels without removing them. Some authors have suggested in the literature that the L_2 norm gives better results in biomedical applications [25], [26]. Therefore, in this paper we will consider the L_2 norm regularization function.

Substituting $\Omega(\beta)$ by the L_2 norm the optimization problem in the primal is now formulated as follows:

$$\begin{aligned} \min_{\beta,\xi} \quad & \frac{1}{2} \| \beta \|^2 + \frac{C_S}{N_S} \sum_{(x_i,x_j) \in S} \xi_{ij} + \frac{C_D}{N_D} \sum_{(x_i,x_j) \in D} \xi_{ij} \\ \text{s.t.} \quad & \beta^T \mathbf{K}_{ij} \geq K^*_{ij} - \xi_{ij} \; \forall (\mathbf{x}_i \; \mathbf{x}_j) \in \mathcal{S} \\ & \beta^T \mathbf{K}_{ij} \leq K^*_{ij} + \xi_{ij} \; \forall (\mathbf{x}_i \; \mathbf{x}_j) \in \mathcal{D} \\ & \beta_l \geq 0 \; \xi_{ij} \geq 0 \; \forall l = 1, \ldots, M \end{aligned} \tag{19}$$

The previous constrained optimization problem can be solved using the method of Lagrange multipliers. Next, the problem can be written in the dual as follows

$$\begin{aligned} \max_{\alpha_{ij},\gamma} \quad & -\frac{1}{2} \sum_{\substack{(x_i,x_j) \in \mathcal{S} \\ (x_k,x_l) \in \mathcal{S}}} \alpha_{ij}\alpha_{kl}\mathbf{K}^T_{ij}\mathbf{K}_{kl} - \frac{1}{2} \sum_{\substack{(x_i,x_j) \in \mathcal{D} \\ (x_k,x_l) \in \mathcal{D}}} \alpha_{ij}\alpha_{kl}\mathbf{K}^T_{ij}\mathbf{K}_{kl} \\ & + \sum_{\substack{(x_i,x_j) \in \mathcal{S}, \\ (x_k,x_l) \in \mathcal{D}}} \alpha_{ij}\alpha_{kl}\mathbf{K}^T_{ij}\mathbf{K}_{kl} - \sum_{(x_i,x_j) \in \mathcal{S}} \alpha_{ij}\gamma^T\mathbf{K}_{ij} \\ & -\frac{1}{2}\gamma^T\gamma + \sum_{(x_i,x_j) \in \mathcal{D}} \alpha_{ij}\gamma^T\mathbf{K}_{ij} + \sum_{(x_i,x_j) \in \mathcal{S}} \alpha_{ij}K^*_{ij} \\ & - \sum_{(x_i,x_j) \in \mathcal{D}} \alpha_{ij}K^*_{ij} \end{aligned}$$

subject to:

$$0 \leq \alpha_{ij} \leq \begin{cases} \frac{C_S}{N_S} & \text{for } (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{S} \\ \frac{C_D}{N_D} & \text{for } (\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{D} \end{cases} \tag{20}$$

$$\gamma_l \geq 0 \quad \forall l = 1, \ldots, M \tag{21}$$

where α_{ij} and γ_l are the Lagrange multipliers. The optimization problem in the dual is convex and quadratic without local minima [27]. Besides, the computational burden depends on the number of active constraints, that is those with $\xi_{ij} \geq 0$. This is more efficient than solving the problem in the primal where the computational complexity is proportional to the number of variables.

Once the α_{ij} and γ_l are estimated in the dual, the coefficients β_l for the linear combination of kernels can be obtained from $\frac{\partial L}{\partial \beta} = 0$. The vector of coefficients can be written as:

$$\beta = \sum_{(x_i,x_j) \in S} \alpha_{ij} \mathbf{K}_{ij} - \sum_{(x_i,x_j) \in D} \alpha_{ij} \mathbf{K}_{ij} + \gamma \tag{22}$$

Substituting in equation (14) we obtain the optimal combination of kernels learned from a set of equivalence constraints. Then, any clustering algorithm that works directly from a kernel matrix may be extended to incorporate a linear combination of dissimilarities. This will help to identify clusters that are non-separable using a single metric.

E. Support Vectors and KKT Complementary Conditions

In this section we study the relation between the value of the Lagrange multipliers and the constraints satisfaction. We also comment the meaning of the support vectors in the context of Multiple Kernel Learning.

The value of the Lagrange Multipliers α_{ij} determines if the linear combination of kernels complies with the constraints (16)-(17). To study this relation more in depth, let consider the Karush-Kuhn-Tucker (KKT) complementary conditions [27] for the optimization problem (19). They can be written in the primal as follows:

$$\begin{aligned} & \alpha_{ij} \left(\beta^T \mathbf{K}_{ij} - K^*_{ij} + \xi_{ij}\right) = 0 && (x_i, x_j) \in \mathcal{S} \\ & \alpha_{ij} \left(\beta^T \mathbf{K}_{ij} - K^*_{ij} - \xi_{ij}\right) = 0 && (x_i, x_j) \in \mathcal{D} \\ & \eta_{ij}\xi_{ij} = 0 && (x_i, x_j) \in \mathcal{S}, \mathcal{D} \\ & \gamma_l\beta_l = 0 && \forall_l = 1, \ldots, M \end{aligned} \tag{23}$$

From the previous KKT complementary conditions the following properties can be derived:

For similarity constraints, that is pairs of $(x_i, x_j) \in S$

$$\beta^T \mathbf{K}_{ij} = \begin{cases} = K^*_{ij} & 0 < \alpha_{ij} < \frac{C_S}{N_S} \\ > K^*_{ij} & \alpha_{ij} = 0 \\ < K^*_{ij} & \alpha_{ij} = \frac{C_S}{N_S} \end{cases}$$

For dissimilarity constraints, that is pairs of $(x_i, x_j) \in D$

$$\beta^T \mathbf{K}_{ij} = \begin{cases} = K^*_{ij} & 0 < \alpha_{ij} < \frac{C_D}{N_D} \\ < K^*_{ij} & \alpha_{ij} = 0 \\ > K^*_{ij} & \alpha_{ij} = \frac{C_D}{N_D} \end{cases}$$

The above properties show that when the similarity/dissimilarity constraints are satisfied with a margin larger than zero, the Lagrange multipliers α_{ij} are null and the corresponding similarity for the pair of objects will not appear in the solution. On the other hand, when the linear combination of kernels fails to satisfy the constraints or they are satisfied with margin exactly equal to zero the Lagrange multipliers are non-null and the similarity for the corresponding pair of objects will be considered in the solution. They are the support vectors and the optimization problem can be formulated exclusively in terms of them. Therefore, the complexity of the optimization algorithm will not depend on the size of the training set but on the number of the support vectors.

F. Datasets Description

We have considered a wide range of data sets to check the performance of the clustering algorithm proposed. Table I shows the different datasets considered and their features. The first three rows correspond to benchmark datasets retrieved from the UCI machine learning database (http://archive.ics.uci.edu/ml/datasets/). The last two rows are complex bioinformatics problems aimed to identify human cancer samples using the gene expression profiles. Both datasets can be recovered from a public webpage (http://bioinformatics2.pitt.edu). We have selected applications with wide range of signal to noise ratio (Var./Samp). In particular the cancer datasets (last two rows) have a high signal to noise ratio with large number of variables and small number of samples. Therefore, they are problems that favor the overfitting of the data and will serve to check the generalization ability of the algorithm proposed. Moreover, as the number of samples is small, the supervisory information available is also quite limited and learning the metric becomes a challenging task. For all the datasets the class label is available. This will help to evaluate rigorously the clustering results considering objective measures. Finally, the variables have been normalized subtracting the median and dividing by the inter-quantile range.

TABLE I. Properties of the Different Data Sets Considered

Data sets	Samples	Variables	Var./Samp	Classes
Wine (UCI)	177	13	0.17	3
Ionosphere (UCI)	351	35	0.01	2
Breast Cancer (UCI)	569	32	0.05	2
Lymphoma	96	4026	41.9	2
Colon Cancer	62	2000	32	2

III. Results

In this section we first comment the preprocessing of the datasets and how the supervisory information is generated. Next, the set of dissimilarities considered by the learning algorithm are introduced as well as the method to estimate the parameters. Finally, we describe the objective measures to evaluate the clustering algorithms and the experimental results are discussed.

Cancer samples using the gene expression profiles are represented in high dimensional spaces with high level of noise to signal ratio. Noisy features may deteriorate the clustering performance. Therefore, feature selection to remove redundant variables is recommended to improve the clustering results [10]. To this aim, genes (features) are ranked by the interquantile range (IQR). Those genes with small variability are considered irrelevant to discriminate between different disease states. We have considered five subsets with the 280; 146; 101; 56 and 34 genes ranked higher considering the IQR. Supervised feature selection algorithms are not considered because in clustering problems class labels are not available. For clustering algorithms based on a single dissimilarity we have chosen the subset of genes that gives rise to the smallest error. Clustering methods based on multiple kernels consider all the dissimilarity matrices obtained from different subset of features. It is expected that the learning algorithm will help to remove dissimilarities based on noisy features.

Regarding the set of dissimilarities integrated into the clustering algorithm we have considered nine measures widely used in bioinformatics applications. Euclidean, Manhattan, Chevichev, Mahalanobis, Cosine, Correlation, Spearman, Kendall-τ and χ^2. In order to build an ensemble of dissimilarities we have considered for each distance different subsets of features and non-linear transformations using kernel methods. After that, we obtained an ensemble of 45 dissimilarity matrices for each type of kernel.

To generate the set of pairwise constraints we have followed the approach of [13]. The similarity constraints S are obtained by sampling randomly all the object pairs that belong to the same class. The size of S is chosen such that the number of connected components is approximately the 20% of the number of objects. The dissimilarity constraints D are chosen sampling randomly the object pairs that belong to different classes. Twenty independent random sets for S and D are generated and the average error is reported.

The optimal values for the regularization parameters C_S and C_D are estimated using a grid search strategy and the errors are computed by ten-fold cross-validation over the set of constraints. The number of clusters for each problem has been set up to the number of classes. As kernel k-means algorithm is sensitive to the initialization we have reported the average error over 20 independent trials with random initialization.

The clustering algorithms have been evaluated by two error functions widely used in the literature [13]. The first one is the accuracy. It determines the probability that two objects that belong to the same or different classes are grouped in the same way by the clustering algorithm. Mathematically it can be defined as:

$$\text{accuracy} = \sum_{i>j} \frac{1\{1\{y_i = y_j\} = 1\{\hat{c}_i = \hat{c}_j\}\}}{0.5\, N\,(N-1)} \tag{24}$$

where y_i is the reference class label for object i and c_i is the group assigned to object i by the clustering algorithm. N is the number of objects in the dataset. The accuracy may lead often to wrong conclusions because the average value for two random partitions is greater than zero. To overcome this problem, it has been proposed in the literature the adjusted randindex [28].

Table II and Table III compare the different clustering algorithms according to accuracy and the adjusted randindex. First row provides the results for the semi-supervised learning algorithm proposed in this paper and based on an ensemble of dissimilarities. Polynomial kernels allow to increase the number of dissimilarities incorporating non-linear transformations of the original ones. We have compared in the second row with a standard clustering method that learns the metric from pairwise constraints [13]. Third row provides the performance of kernel k-means based on the best measure for the whole family of dissimilarities considered. Each column reports the best distance regarding the data set analyzed. Again, the original dissimilarities may be non-linearly transformed to obtain more general measures using polynomial kernels. Finally, last row shows the results for k-means standard clustering algorithm based on the euclidean distance. Polynomial kernels allow us to transform non-linearly this metric to consider more general dissimilarities and non-spherical groups.

TABLE II. Accuracy for the Semi-supervised Clustering Algorithm Proposed Versus Other Approaches. The Results Are Averaged Over Twenty Independent Random Subsets for S and D

Technique	Kernel	Wine	Ionosphere	Breast	Colon	Lymphoma
Clustering proposed	Linear	0.94	0.90	0.92	0.89	0.95
	Pol. 3	0.96	0.89	0.92	0.90	0.92
Metric learning (Xing)	Linear	0.87	0.74	0.85	0.87	0.90
	Pol. 3	0.51	0.74	0.86	0.88	0.90
Kernel K-means (Best dissimilarity)	Linear	0.94	0.88	0.90	0.88	0.94
	Pol. 3	0.94	0.88	0.90	0.88	0.93
		$\chi 2$	*Mahalanobis*	*Manhattan*	*Correlation*	$\chi 2$
K-means (Euclidean)	Linear	0.92	0.72	0.88	0.87	0.90
	Pol. 3	0.87	0.73	0.88	0.88	0.90

TABLE III. Adjusted RandIndex for the Semi-Supervised Clustering Algorithm Proposed Versus Other Approaches. The Results Are Averaged Over Twenty Independent Random Subsets for S and D

Technique	Kernel	Wine	Ionosphere	Breast	Colon	Lymphoma
Clustering proposed	Linear	0.82	0.63	0.69	0.60	0.79
	Pol. 3	0.85	0.60	0.69	0.63	0.73
Metric learning (Xing)	Linear	0.68	0.23	0.50	0.54	0.66
	Pol. 3	0.50	0.23	0.52	0.58	0.65
Kernel K-means (Best dissimilarity)	Linear	0.82	0.58	0.66	0.59	0.77
	Pol. 3	0.81	0.58	0.66	0.59	0.76
		$\chi 2$	*Mahalanobis*	*Manhattan*	*Correlation*	$\chi 2$
K-means (Euclidean)	Linear	0.79	0.20	0.59	0.59	0.65
	Pol. 3	0.67	0.21	0.60	0.59	0.65

From the analysis of Table II and Table III, we report three relevant conclusions:

First, the semi-supervised clustering algorithm proposed in this paper improves significantly the performance of a benchmark clustering algorithm developed by Xing [13] that learns the metric from pairwise constraints. The accuracy and the adjusted randindex are significantly improved even for the cancer datasets, with high level of noise and large number of variables. This result can be explained because the clustering proposed here has smaller number of parameters and the regularization term helps to reduce the overfitting. Notice also that our model integrates dissimilarities based on different sets of features removing the problem of choosing the optimal set of variables, which is a complex task in clustering problems.

To determine if the differences between our clustering algorithm and the one proposed by Xing are statistically significant we have computed the boxplots for both techniques. To this aim, we have generated 20 independent random sets of constraints for S and D and we have estimated the accuracy and the adjusted randindex. Fig. 3 shows the boxplots for the accuracy and Fig. 4 for the adjusted randindex. Odd numbers in the x-axis correspond to the boxplots for our clustering algorithm and the different datasets considered in the same order as in Table II. Similarly, even numbers correspond to the supervised clustering algorithm proposed by Xing. The boxplots show that the differences between both algorithms are statistically significant at 95% confidence level for all the datasets considered in this paper.

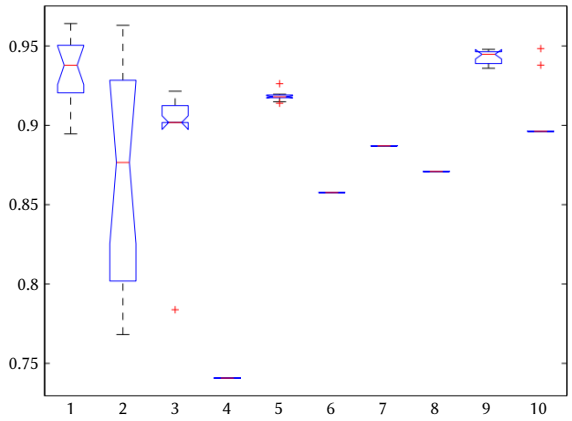

Fig. 3. Accuracy boxplots that compare the multiple kernel learning clustering proposed with the metric learning algorithm developed by Xing. 20 independent trials have been recorded considering 20 sets of constraints generated randomly.

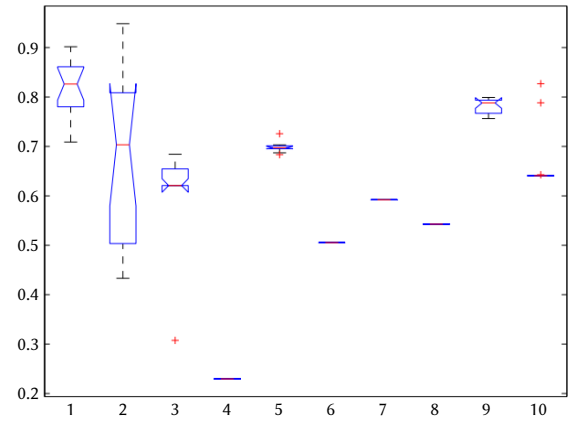

Fig. 4. Adjusted RandIndex boxplots. They compare the multiple kernel learning clustering proposed with the metric learning algorithm developed by Xing. 20 independent trials have been recorded considering 20 sets of constraints generated randomly.

The clustering algorithm proposed that integrates an ensemble of dissimilarities improves the accuracy and adjusted randindex of kernel k-means based on the best similarity. The combination of dissimilarities provides more information that a single measure. Besides, Table II and Table III show that the best dissimilarity depends on the particular problem considered. Moreover, for unsupervised applications choosing the best measure is an ill-posed problem, because no supervised index error can be defined to guide the selection of an appropriate metric. Our algorithm helps to overcome the problem of choosing an optimal dissimilarity, the best kernel and even the optimal subset of features. This is frequently a challenging task, for instance in complex bioinformatics applications.

Finally, we remark that the learning algorithm proposed improves significantly the standard k-means clustering algorithm based on the euclidean distance for all the datasets considered. The results are similar for a non-linear transformation of the euclidean distance considering polynomial kernels of degree 3.

IV. Discussion

Several algorithms developed in the literature to learn the distance metric are related to the one proposed here. The first approach tries to learn a full or diagonal Mahalanobis distance considering pairwise constraints [1], [12], [13]. Some authors have extended the previous techniques to more general dissimilarities using kernel methods [14], [15], [29]. However, they are based on a single distance metric that may fail to reflect accurately the objects proximities. Besides, as the number of parameters grows with the space dimensionality they are prone to overfitting and the computational complexity is high. Although new algorithms have been proposed to improve the computational efficiency and to reduce the overfitting [6] they suffer from similar drawbacks. Several differences are worth to mention with the approach proposed here. First our algorithm is able to integrate a set of dissimilarities that may exhibit different properties from a set of pairwise constraints. Second, the loss function incorporates a penalty term and has a small number of parameters which helps to reduce the overfitting. Finally, the optimization problem is quadratic, the complexity depends on the number of support vectors and it is efficient computationally.

Our approach is more related to multiple kernel clustering methods [7]–[9], [16] that are able to integrate different dissimilarities that come from different features or representations of the objects using kernel methods. However, these algorithms differ from our approach because they integrate the dissimilarities in an unsupervised way and the resulting metric may not help to improve the clustering results. In this way, some researchers have mentioned that learning the metric without any supervised information may be an ill-defined problem [15].

Finally, few authors have addressed the problem of multiple kernel learning from a set of pairwise constraints for clustering applications [17], [30]. However, they rely on complex optimization problems that are more difficult to solve than the one proposed in this research.

V. Conclusion

In this paper we have developed a semi-supervised learning algorithm to integrate an ensemble of kernels (similarities) into a clustering algorithm using weak supervision in the form of pairwise constraints. Our method offers three advantages over previous metric learning algorithms. First, it learns a combination of dissimilarities that may come from different features of the objects or different kernels. This strategy avoids the problem of choosing the right kernel (similarity), the best subset of features or the optimal value for the kernel parameters that may be a challenging task for certain type of applications. Second, the loss function is convex and quadratic and it may be efficiently optimized. Finally, the learning algorithm is robust to overfitting.

The clustering algorithm proposed has been applied to three benchmark datasets and to complex cancer identification problems based on the gene expression profiles. The experimental results suggest that learning a combination of similarities (kernels) improves the performance of a clustering algorithm based on the best similarity (kernel). Besides, the algorithm developed outperforms a standard semi-supervised clustering proposed in the literature that learns the metric from the data. In particular, our method performs significantly better for cancer problems with high level of noise to signal ratio which suggests that it is robust to overfitting.

Future research trends will focus on the application of this formalism to other bioinformatics problems such as gene function prediction.

References

[1] C. Shen, J. Kim, L. Wang, "Scalable large-margin mahalanobis distance metric learning," IEEE transactions on Neural Networks, vol. 21, no. 9, pp. 1524–1530, 2010.

[2] H. Fyad, F. Barigou, K. Bouamrane, "An Experimental Study on Microarray Expression Data from Plants under Salt Stress by using Clustering Methods," International Journal of Interactive Multimedia and Artificial Intelligence, vol. 6, no. 2, pp. 38-47, 2020.

[3] M. Martín-Merino, Á. Blanco, "A local semi-supervised sammon algorithm for textual data visualization," Journal of Intelligent Information Systems, vol. 33, no. 1, pp. 23–40, 2009.

[4] A. Woznica, A. Kalousis, M. Hilario, "Learning to combine distances for complex representations," in Proceedings of the 24th International Conference on Machine Learning, 2007, pp. 1031–1038.

[5] A. Seal, A. Karlekar, O. Krejcar, E. Herrera-Viedma, "Performance and convergence analysis of modified C-means using jeffreys-divergence for clustering", International Journal of Interactive Multimedia and Artificial Intelligence, vol. 7, no. 2, pp. 141-149, 2021.

[6] B. Nguyen, B. De Baets, "Kernel-based distance metric learning for supervised k-means clustering," IEEE Transactions on Neural Networks and Learning Systems, vol. 30, no. 10, pp. 3084–3095, 2019.

[7] A. Huang, T. Zhao, C.-W. Lin, "Multi-view data fusion oriented clustering via nuclear norm minimization," IEEE Transactions on Image Processing, vol. 29, pp. 9600–9613, 2020.

[8] B. Zhao, J. T. Kwok, C. Zhang, "Multiple kernel clustering," in Proceedings of the 2009 SIAM International Conference on Data Mining, 2009, pp. 638–649, SIAM.

[9] J. Hu, M. Li, E. Zhu, S. Wang, X. Liu, Y. Zhai, "Consensus multiple kernel k-means clustering with late fusion alignment and matrix-induced regularization," IEEE Access, vol. 7, pp. 136322–136331, 2019.

[10] D. Huang, W. Pan, "Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data," Bioinformatics, vol. 22, no. 10, pp. 1259–1268, 2006.

[11] H. Zeng, Y.-m. Cheung, "Semi-supervised maximum margin clustering with pairwise constraints," IEEE Transactions on Knowledge and Data Engineering, vol. 24, no. 5, pp. 926–939, 2011.

[12] A. Bar-Hillel, T. Hertz, N. Shental, D. Weinshall, G. Ridgeway, "Learning a mahalanobis metric from equivalence constraints," Journal of Machine Learning Research, vol. 6, no. 6, 2005.

[13] E. Xing, A. Y. Ng, M. I. Jordan, S. J. Russell, "Distance metric learning with application to clustering with side-information," in Advances in Neural Information Processing Systems, vol. 15, 2002, MIT Press.

[14] B. Nguyen, B. De Baets, "Kernel distance metric learning using pairwise constraints for person reidentification," IEEE Transactions on Image Processing, vol. 28, no. 2, pp. 589–600, 2018.

[15] D.-Y. Yeung, H. Chang, "A kernel approach for semisupervised metric learning," IEEE Transactions on Neural Networks, vol. 18, no. 1, pp. 141–149, 2007.

[16] Y. Yao, Y. Li, B. Jiang, H. Chen, "Multiple kernel kmeans clustering by selecting representative kernels," IEEE Transactions on Neural Networks and Learning Systems, vol. 32, no. 11, pp. 4983–4996, 2020.

[17] T. Yang, R. Jin, A. K. Jain, "Learning kernel combination from noisy pairwise constraints," in 2012 IEEE Statistical Signal Processing Workshop (SSP), 2012, pp. 752–755, IEEE.

[18] V. Vapnik, Statistical Learning Theory. New York: John Wiley & Sons, 1998.

[19] E. Pekalska, P. Paclick, R. Duin, "A generalized kernel approach to dissimilarity-based classification," Journal of Machine Learning Research, vol. 2, pp. 175–211, 2001.

[20] G. Wu, E. Y. Chang, N. Panda, "Formulating distance functions via the kernel trick," in Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, 2005, pp. 703–709.

[21] M. Martín-Merino, A. Blanco, J. De Las Rivas, "Combining dissimilarities in a hyper reproducing kernel hilbert space for complex human cancer prediction," Journal of Biomedicine and Biotechnology, vol. 2009, 2009.
[22] N. Cristianini, J. Kandola, J. Elisseeff, A. Shawe-Taylor, "On the kernel target alignment," Journal of Machine Learning Research, vol. 1, pp. 1–31, 2002.
[23] C. Soon Ong, A. Smola, R. Williamson, "Learning the kernel with hyperkernels," Journal of Machine Learning Research, vol. 6, pp. 1043–1071, 2005.
[24] A. Rakotomamonjy, F. Bach, S. Canu, Y. Grandvalet, "Simple multiple kernel learning," Journal of Machine Learning Research, vol. 9, pp. 2491–2521, 2008.
[25] S. Yu, T. Falck, A. Daemen, L.-C. Tranchevent, J. A. Suykens, B. De Moor, Y. Moreau, "L2-norm multiple kernel learning and its application to biomedical data fusion," BMC bioinformatics, vol. 11, no. 1, pp. 1–24, 2010.
[26] M. Kloft, U. Brefeld, P. Laskov, K.-R. Müller, A. Zien, S. Sonnenburg, "Efficient and accurate lp-norm multiple kernel learning," in Advances in Neural Information Processing Systems, vol. 22, 2009, Curran Associates, Inc.
[27] W. Kaplan, Maxima and minima with applications: practical optimization and duality, vol. 51. John Wiley & Sons, 1998.
[28] L. Hubert, P. Arabie, "Comparing partitions," journal of classification, vol. 2, pp. 193–228, 1985.
[29] M. S. Baghshah, S. B. Shouraki, "Non-linear metric learning using pairwise similarity and dissimilarity constraints and the geometrical structure of data," Pattern Recognition, vol. 43, no. 8, pp. 2982–2992, 2010.
[30] B. Yan, C. Domeniconi, "An adaptive kernel method for semi-supervised clustering," in European Conference on Machine Learning, 2006, pp. 521–532, Springer.

Manuel Martín Merino

Manuel Martín Merino received the B. S. degree in physics from the University of Salamanca (Spain) in 1996 and the PhD. degree in applied physics from the same University in 2003. He is currently professor of Artificial Intelligence and supervisor of the Telefonica Chair in the Computer Science school at the University Pontificia of Salamanca. His research interests include Machine Learning, Artificial Neural Networks, Kernel methods, Support Vector Machines, clustering algorithms and their applications to Bioinformatics and text mining problems.

Alfonso José López Rivero

Alfonso José López Rivero. PhD in Computer Science. He is a professor, since 1996, and member of the research group GESTICON (Ethical and Technological Management of Knowledge) at the Pontifical University of Salamanca, UPSA, (Spain). He is a member of the organizing and scientific committee of several international symposia and co-author of articles published in several recognized journals, workshops and symposia. He has been Dean of the Faculty of Computer Science and Director of the Office for the Transfer of Research Results at UPSA.

Vidal Alonso Secades

Vidal Alonso was born in Luanco, Spain, in 1966. He received the Computer Science Degree in 1992 from the Polytechnic University of Madrid, and the Ph. D. degree, in 2004 from the Pontifical University of Salamanca, Spain. He was a Full Professor of Computer Science at the Pontifical University of Salamanca since 1994. He has occupied the position of Vice rector at his University for five years, until 2015. He also was the Director of the Computer Science School for nine years (2000-2009.) He works in data structures, knowledge discovery and data quality. Dr. Alonso is a member of ALI (Computer Science Spanish Association) and he won the Castilla y Leon Digital Award in 2007.

Marcelo Vallejo García

PhD in Computer Science. Bachelor in Business Administration. Currently Professor of Financial Economics and Accounting, in the Faculty of Computer Science of the Pontifical University of Salamanca. Among his former positions, we can highlight the following: Director of the Doctoral Program in Insurance, Legal and Business Sciences by Pontifical University of the Pontifical University of Salamanca, Vice Dean of the UPSA Faculty of Science Computer during the period 2015-2021 and Responsible of the module "Union Economic and Monetary", financed by the European Commission, as a part of Jean Monnet actions. With more than 30 years of teaching experience, he is the current coordinator of Bachelor's Degree in Technology Business of the Pontifical University of Salamanca. He has participated as a collaborating researcher and principal investigator in several competitive projects related to their areas of research and teaching. He is the author and co-author of scientific publications and has participated as speaker at several national and international conferences, mainly in topics relatives to corporate reputation and management and reporting of financial and no-financial business information.

Antonio Ferreras García

Antonio Ferreras, received a PhD in Telecommunications Engineering, in 1995 from the Polytechnic University of Madrid for his studies in Optoelectronics. He also has degrees in Economics, Psychology and Law: from 2018 he is professor at the Pontificial University of Salamanca. Currently, his research area is focused on cybersecurity and data science.