GHOST: Graph Higher-Order Similarity Transformation for Classification

Enzo Battistella, Maria Vakalopoulou, Nikos Paragios, Eric Deutsch

GHOST: Graph Higher-Order Similarity Transformation for Classification

Enzo Battistella, Maria Vakalopoulou, *Member,* Nikos Paragios, *Fellow Member*, and Éric Deutsch

Abstract—Exploring and identifying a good feature representation to describe large-scale datasets is one of the main problems of machine learning algorithms. However, plenty of feature selection techniques and distance metrics with very different properties exist, which entails an intricacy for identifying the proper method. This paper provides a general algorithm to design a high-order distance metric over a sparse selection of features dedicated to semi-supervised clustering and classification. We extend usual learning methods to design a metric accounting for patterns over sets of objects. Our approach is based on Conditional Random Field (CRF) energy minimization and Dual Decomposition, which allow efficiency and great flexibility in the features to consider. In particular, it enables to leverage the higher-order graph structures information efficiently. The optimization technique employed ensures the tractability of very high dimensionality problems using hundreds of features and samples. On a challenging task of Covid-19 patients stratification, we compare the classification results between state of the art baselines and our proposed classifier, relying on the higher-order distance learned to prove this metric formulation's relevance.

Index Terms—Distance Learning, Conditional Random Field, Higher-order, Feature Selection

1 Introduction

In machine learning, the choice of a suitable feature space is of prime importance. Not only, dimensionality curse might affect the data, but some variables might be noisy or non-relevant. Many techniques have been investigated to select the most relevant features using some statistical metrics as correlation [1] or the importance weights an algorithm, as elastic-net, grants to the variables [2]. Also, many approaches as clustering algorithms require a relevant metric to define a similarity notion between the samples. However, depending on the algorithm's mathematical properties and the metric, very different results are obtained [3]. To tackle the difficult task of designing a dedicated similarity function, some studies investigated semi-supervised clustering or distance learning [4]–[6]. Such approaches present the significant advantage to perform both a selection of the most relevant dimensions and a warping of the feature space favoring the spatial proximity of samples presenting the same labels. Despite the tremendous amount of work carried on this topic, another aspect has been poorly considered, leveraging higher-order patterns in the data.

Graph structures are a standard model for data representation and allow great expressiveness. They have been adopted in various fields from spotting weak points in physical [7] or internet [8] networks to modeling opinion diffusion processes [9], graph approaches' versatility is remarkable. Graphs attractiveness is explained by the valuable information they carry in many application fields. They have already significantly spread for biological and medical data as they allow to better express the intricate interactions at stake [10]. Some of the prominent investigation areas are genomics and proteomics [11]. But, their relevance for imaging data has been demonstrated in, for instance, graph matching problems [12], [13]. And, lately, they have been gaining increasing importance with Graph Convolutional Networks [14].

However, the information graphs carry is complex to leverage from both a time and a memory point of view. Most of the studies exploiting these structures are limited to second-order properties [15] i.e., the edges in the graph. Some studies propose various ways to better estimate higher-order properties. However, they mainly focus on local properties [16] or small, specific and predefined patterns to orient their analysis [17]. Even if some studies attempt to design methods to avoid resorting to higher-order structure surrogates involving simplification information loss [18], those approaches have not been translated to distance learning. Nevertheless, in exploratory studies as [19], many different and complex higher-order graph properties are leveraged as the connectivity or the centrality of the nodes or their clique order. This work and the ever growing influence of graph approaches will motivate our Ghost approach.

This study proposes a new approach to bridge the gap between feature selection, distance learning, and leveraging higher-order patterns. The rest of the article is structured as follows. In Section 2 we position our work compared to existing studies. In Section 3 we present in details the method we based our general framework on, our problem formulation and resolution algorithm, how we leverage higher-order graph information and perform classification using the dedicated distance we learn. In Section 5, we present the datasets we considered and the results obtained compared to state of the art classification baselines.

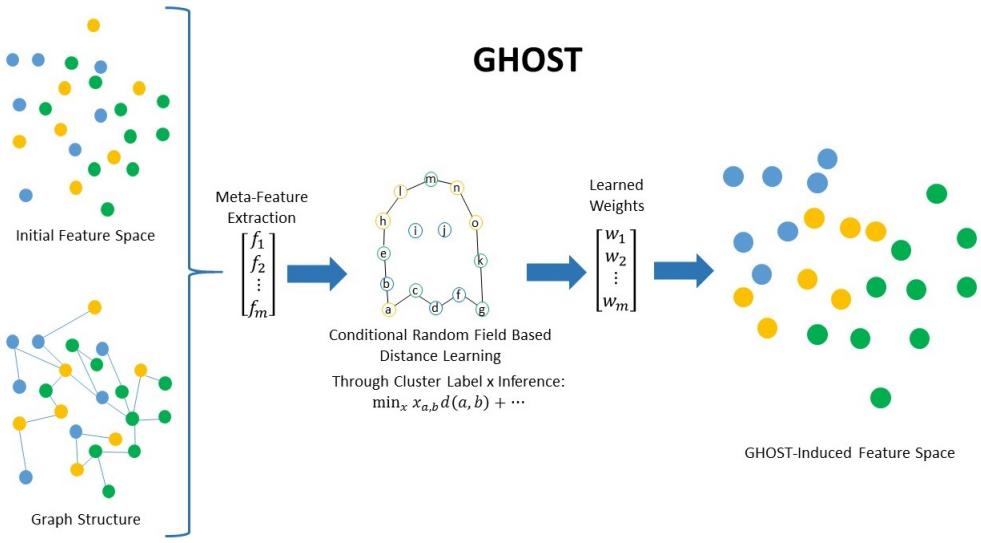

Figure 1: **Proposed Ghost approach.** A general overview of the different steps of our process. Our proposed Ghost approach aims at identifying the most relevant meta-features for a classification task. Meta-features might include any kind of distance or higher-order metric between any number of the natural features of the data. The approach relies on Conditional Random Field energy minimization for distance learning. It combines cluster label inference and optimization of meta-feature weights to induce a new metric space where samples with a same label are close.

2 Related Work

Our study is based on the work of [6] which has the tremendous advantage of allowing a very general definition of distance, allowing both feature and metric selection, and weighting. Besides, it is designed for a center-based clustering algorithm, which presents the advantage of defining a relevant cluster representative which exciting properties have been demonstrated for feature selection in [3], [20]. This clustering paradigm is the critical point in our approach for exploiting higher-order graph structures in a reasonable computational time and space.

The idea of investigating clustering guided by field experts properties has been poorly investigated in the literature. Notwithstanding, we can observe two main trends. First, the semi-supervised approach considers partial annotations guiding the clustering process. Following this paradigm, the methods from [21] or [22] account for information of samples which must or cannot be clustered together to influence the clustering using domain-related knowledge. This approach might be further generalized to constraints on conjunctions and disjunctions of instances [23]. Second, the most prominent approach, metric learning, aims to learn a measure to discover specific properties thanks to a supervised framework. It offers the ability to identify structures similar to a given ground-truth. This paradigm involves a completely annotated dataset at the difference of the semi-supervised approach. However, once the metric learned, its strength dwells in its ability to be applied for clustering without the need of any additional label. Besides, many studies demonstrated the essential role of the distance measure considered for clustering [5], even more prominent than the choice of a correct clustering algorithm. The distance notion has to capture the required information to enable any algorithm to achieve a relevant clustering of the data. Following this precept, several studies considered the arduous task of metric learning from different perspectives. In [24], the authors leverage a deep learning architecture to define a space representation allowing to define a better similarity notion between instances while authors from [25] resort to a Support Vector Machine algorithm. In [4] constraints are defined to formulate the metric-learning task as a convex optimization problem. Here, we will consider more in detail the formulation proposed in [6] which relies on the Conditional Random Field (CRF) energy minimization principle to specify a metric in a center-based clustering context. This model is all the more interesting that the relevance of center-based techniques as been demonstrated in [3] for feature selection towards classification purposes.

Despite the excellent expressiveness of CRF and their ability to capture higher-order relations, the toll for fully leveraging this higher-order information might be heavy on memory and time consumption aspects. To cope with this issue, authors from [26], [27] proposed to exploit the binary nature of the CRF labels to optimize the resolution. Finally, in [28], dual decomposition is exploited to divide the initial energy to minimize in several easier-to-solve sub-problems.

To naturally leverage higher-order information for clustering, some first attempts investigated the presence of simple patterns in a graph [16]. Still, it is generally performed with minimal patterns as a clique of order 3 or only for a local higher-order clustering [17]. However, resorting to those surrogates expose to an information loss in the complex higher-order relations available [18].

In this study, we adapt and extend the formulation of metric learning presented in [6]. The main contribution is to bring the center-based clustering and the notion of pairwise metrics to higher-order settings through the inclusion of graph properties. In addition, we modify the error function that was considered to better account for unbalanced

classes. Finally, we leverage graph structural information from our data.

3 METHODOLOGY

Without loss of generality, let us define a metric assessing the similarity of a set of h objects as a h^{th}-order metric. For instance, a usual distance is referred to as a 2^{nd}-order metric or pairwise metric. For the simple case of pairwise metrics, we based our study on [6] which provides a general, flexible and efficient approach to solve the learning problem in a clustering context. This approach relies on CRF energy minimization. In this study, we extend our formulation introducing 3^{rd}-order to h^{th}-order metrics for any $h > 2$. The case of 3^{rd}-order metrics will be more detailed for clarity's sake. However, the same approach is applied to obtain the results for any order. This section first details the notations and potential functions formulations used to define the problem's energy. Then, we introduce the optimization problem, discussing dual decomposition and its application to the task. Finally, we present the process used to extract the higher-order information we leverage. The proposed Ghost approach is represented in Figure 1.

3.1 Center-Based Clustering

Our approach is based on center-based clustering. Considering a set of objects to cluster $\mathcal{V}$, we define a set of binary variables $\{x_{p,q}\}_{p,q\in V}$ indicating whether p is assigned to the cluster of center q, $x_{p,q} = 1$, or not, $x_{p,q} = 0$. We consider a distance $d_{p,q}$ between objects p and q.

$$\begin{aligned} \min_{x} \sum_{p,q\in V} d_{p,q}x_{p,q} \quad s.t. \sum_{q\in V} x_{p,q} = 1,\ \forall p \\ x_{p,q} \leq x_{q,q}, \quad x_{p,q} \in \{0,1\},\ \ \forall p,q. \end{aligned} \tag{1}$$

The above system minimizes the distance between a point and the center of the cluster it is assigned to with respect to three constraints. First, each point has to belong to one and only one cluster. Second, if a point is assigned to a cluster center, this center must be assigned to itself. Third, assignment variables are binary. We can cast the previous optimization problem as an equivalent energy minimization task:

$$E(x,d) = \sum_{p,q} u_{p,q}(x_{p,q},d) + \sum_{p,q} \phi_{p,q}(x_{p,q},x_{q,q}) + \sum_{p} \phi_p(x_p) \tag{2}$$

$u_{p,q}$ being the second-order potentials of the CRF standing for the distance to the cluster center and ϕ_p, $\phi_{p,q}$ the constraints. More precisely:

$$\begin{aligned} u_{p,q}(x_{p,q},d) &= d_{p,q}x_{p,q} \\ \phi_{p,q}(x_{p,q},x_{q,q}) &= \delta(x_{p,q} \leq x_{q,q}) \\ \phi_p(x_p) &= \delta(\sum_{q} x_{p,q} = 1) \end{aligned} \tag{3}$$

with $x_p = \{x_{p,q} \mid q \in V\}$ and $\delta(e) = 0$ if e True and ∞ otherwise.

In this study, we propose a generalization of this energy formulation for clustering. Here an example in a third-order setting. In addition to the previous notations, we consider a third-order distance $d_{p,p',q}$ for triplet p, p' and q.

$$\begin{aligned} E(x,d) = \sum_{p,q} u_{p,q}(x_{p,q},d) + \sum_{p,p',q} u_{p,p',q}(x_{p,q}x_{p',q},d) + \\ \sum_{p,q} \phi_{p,q}(x_{p,q},d_{q,q}) + \sum_{p} \phi_p(x_p) \end{aligned} \tag{4}$$

the new function $u_{p,p',q}$ being the third-order potentials of the CRF everything else remaining unchanged. More precisely:

$$u_{p,p',q}(x_{p,q}x_{p',q},d) = d_{p,p',q}x_{p,q}x_{p',q} \tag{5}$$

3.2 Metric Learning Formulation

Our framework learns a distance between objects using a set of K training subjects $\{V^k, \mathcal{C}^k, y^k\}$ for each set $k \in K$, V^k is the set of objects to be clustered according to ground truth $\mathcal{C}^k$ and knowing input data y^k. We are also assuming that we can get from the input data a positive feature function for each pair of objects p, q as $f_{p,q}(y^k)$ and for each triplet of objects p, p', q as $f_{p,p',q}(y^k)$. The codomain of the feature's functions will be called meta-feature space as it is obtained from the actual features of the task and is the input space of the framework. We consider a meta-feature space of size d. One should notice that even though there is a ground truth cluster for each set, the cluster centers remain unknown. A feasible solution x^k of $\mathcal{C}^k$ denoted as $x^k \in \mathcal{X}(\mathcal{C}^k)$, will consist in a set of assignment such as for each ground truth cluster $C \in \mathcal{C}^k$ all the objects $p \in C$ are assigned to the same center $q \in C$. Besides, we are looking for a distance over a set S of cardinal $2 \leq \mid S \mid \leq 3$ expressed as:

$$d^k_S = \begin{cases} d^k_{p,q}\ if\ S = \{p,q\} \\ d^k_{p,p',q}\ otherwise \end{cases} \tag{6}$$

where

$$d^k_{p,q} = w^T f_{p,q}(y^k), \quad d^k_{p,p',q} = w^T f_{p,p',q}(y^k)$$

For conciseness sake's, we will denote $E^k(x,d) = E(x,d^k)$ and $u^k(x,d) = u(x,d^k)$.

w being the weight vector we want to estimate. At the difference of the formulation proposed in [6], we impose $w_i \geq 0, \forall i \leq d$. This specificity aims to enforce the positivity of the distance obtained. Also, as it has been presented in [6], a projection of the weights onto $\mathbb{R}_+$ ensures better performance in the second-order settings. We impose this constraint to improve the tractability of the higher-order distance learning resolution, as is highlighted in the proofs (in Supplementary Materials).

Notice that our framework is very robust and flexible as we use the same weight vector w for both the second-order and the third-order distances, which means that a component i of vectors $f_{p,q}(y^k)$ and $f_{p,p',q}(y^k)$ have to relate to the same property. The i-th component of $f_{p,p',q}(y^k)$ can be a generalization of $f_{p,q}(y^k)$ one, but it can also stand alone, and in this case, $f_{p,q}(y^k)$ will be null. Similarly, a component $f_{p,q}(y^k)$ might not possess any relevant generalization, and in this case, $f_{p,p',q}(y^k)$ will be null. For instance, a suitable third-order function $f_{p,p',q}$ could have for component the perimeter or surface of the triangle $\{p,p',q\}$. With this particular example, in addition to the initial second-order warping of the space, such as we have a small distance between

each object of the cluster and the center, we will also have a small distance between pairs of objects. However, much more intricate properties can be introduced. For instance, we can consider statistical distances as the Mahalanobis distance between the set of observations $\{p, p'\}$ and q, which is designed to estimate if the object q is a natural center for the set $\{p, p'\}$ regarding mean and variance considerations. We will present in Section 3.9, possible higher-order feature functions definition on graph structures.

A Max-Margin approach is considered to approximate w. We are looking for $x^k \in \mathcal{X}(\mathcal{C}^k)$ whose energy $E^k(x^k, d)$ is smaller than the energy of any other solution x by an error function $\Delta(x, d)$ to be defined, i.e.

$$\exists x^k \in \mathcal{X}(\mathcal{C}^k),\ E^k(x^k, d) \leq E^k(x, d) - \Delta(x, d) + \xi_k \quad (7)$$

where slack variable ξ_k is considered in case of infeasible training sets. Adding this constraint to the previous energy minimization problem gives the regularized loss:

$$\min_{\{x^k \in \mathcal{X}(\mathcal{C}^k)\}} \tau J(w) + \sum_k \mathcal{L}_{E^k} \quad (8)$$

where $J(w)$ is a regularization term penalizing w complexity, while the hinge loss $\mathcal{L}_{E^k}$ includes ξ_k and is expressed as:

$$\mathcal{L}_{E^k}(x^k, w) = E^k(x^k, w) - \min_x (E^k(x, w) - \Delta(x, \mathcal{C}^k)) \quad (9)$$

It favors feasible solutions with energy close to the minimal energy for any possible assignment penalized by the error function according to the violated constraints.

3.3 Max-Margin Energy

The good choice of $\Delta(x, \mathcal{C}^k)$ is essential to obtain a relevant w. In particular, we need this error function to be 0 if $x \in \mathcal{X}(\mathcal{C}^k)$ and to have a value representing on what extent x violates the constraints imposed by the ground truth $\mathcal{X}(\mathcal{C}^k)$. We adapt the error function proposed in [6] to better account for unbalanced classes. We consider the training set k with cluster ground truth $\mathcal{C}^k$ the function:

$$\begin{aligned}\Delta(x, C^k) = \alpha \sum_{C \in \mathcal{C}^k} W(1 - \sum_{q \in C} x_{q,q}) + \\ \beta \sum_{C \in \mathcal{C}^k} \frac{1}{\mid C \mid} \sum_{p \in C} (1 - \sum_{q \in C} x_{p,q})\end{aligned} \quad (10)$$

with $W(z) = |z|([z < 0].(\mid V^k \mid - \mid C \mid) + [z > 0]. \mid C \mid)$, $[.]$ being the indicator function. The first term penalizes solutions x^k presenting no or several exemplars for a ground truth cluster $C \in \mathcal{C}^k$. We put an additional penalty on the sizable clusters presenting no exemplar or the small clusters presenting several exemplars. The second term penalizes the solutions that do not assign for an object of a ground truth cluster $C \in \mathcal{C}^k$ an exemplar from C. We considered a weight inversely proportional to the cluster's size to balance the importance of small clusters in the learning process. The learning constants α and β are characterizing the relative importance of the two terms. The regularized loss defined in equation 8 can be expressed as a new CRF energy $\bar{E}^k = E^k - \Delta$:

$$\begin{aligned}\bar{E}^k(x, w) = \sum_{p,q} \bar{u}^k_{p,q}(x_{p,q}) + \sum_{p,p',q} \bar{u}^k_{p,p',q}(x_{p,q}x_{p',q}) + \\ \sum_{p,q} \bar{\phi}_{p,q}(x_{p,q}) + \sum_p \bar{\phi}_p(x_p) + \sum_{C \in \mathcal{C}^k} \bar{\phi}_C(x_C) - \beta|\mathcal{C}^k|\end{aligned} \quad (11)$$

with:

$$\begin{aligned}&\bar{u}^k_{p,q}(x_{p,q}) = u^k_{p,q}(x_{p,q}, d) + \beta[\exists C \in \mathcal{C}^k,\ p, q \in C]\frac{x^k_{p,q}}{\mid C \mid} \\ &\bar{u}^k_{p,p',q}(x_{p,q}x_{p',q}) = u^k_{p,p',q}(x_{p,q}x_{p',q}, d) \\ &\bar{\phi}_{p,q}(x_{p,q}) = \phi_{p,q}(x_{p,q}, x_{q,q}) \\ &\bar{\phi}_p(x_p) = \phi_p(x_p) \\ &\bar{\phi}_C(x_C) = -\alpha W(1 - \sum_{q \in C} x_{q,q})\end{aligned} \quad (12)$$

It is interesting to notice that thanks to the property of Δ for feasible solutions, $\forall x^k \in \mathcal{X}(\mathcal{C}^k)$, $\bar{E}^k(x^k, w) = E^k(x^k, w)$.

3.4 Optimizing over $\{x^k\}$

For a fixed w, minimizing $\bar{E}^k(x^k, w)$ requires the constraints to be satisfied, $\bar{\phi}_p(x^k_p) = 0$ and $\bar{\phi}_{p,q}(x^k_{p,q}) = 0$, which entails $x^k \in \mathcal{X}(C^k)$. And, in this case, $\Delta(x, C^k) = 0$. Thus,

$$x^k = \arg\min_{x \in \mathcal{X}(C^k)} (\sum_{p,p',q} d_{p,p',q}x_{p,q}x_{p',q} + \sum_{p,q} d_{p,q}x_{p,q}) \quad (13)$$

To minimize this problem, we only need to find the set Q^k of exemplars q minimizing the above function per cluster in $\mathcal{C}^k$ and then assign each point of the cluster to its exemplar. In this case, the constraints will be satisfied as we ensure each cluster to have one and only one center and assign all cluster samples to this center.

3.5 Dual Decomposition

The dual decomposition approach is a widespread approach in optimization. Its efficient resolution by projected subgradients has been introduced for MRF and CRF in [29]. It presents the crucial property of optimally solving the dual linear programming problem. Besides, its versatility allows the generalization to higher-order CRF as performed in [28].

Dual decomposition principle aims at isolating several much easier subproblems tailored to be equivalent to the original problem after summation. A simple and popular decomposition is to consider each node independently. Each subproblem might be solved by very efficient inference techniques as graph-cuts approaches without venturing into a scaling issue. The global resolution leads to a projected subgradient scheme, provably offering an optimal solution. Formally, dual decomposition expression relies on simple subproblems also called slave problems and on a master problem enacting as a coordinator.

Here, for each $k < K$, we define a slave problem per datapoint $p \in V^k$, $\bar{E}_p^k$, and one per cluster $C \in \mathcal{C}^k$, $\bar{E}_C^k$.

$$\bar{E}_p^k(x,w) = \sum_{q \neq p} \bar{u}_{p,q}^k(x_{p,q}) + \sum_{p',q \neq p} \bar{u}_{p,p',q}^k(x_{p,q}x_{p',q}) + \sum_q \bar{\phi}_{p,q}(x_{p,q}) + \bar{\phi}_p(x_p) - \frac{\beta}{\mid V^k \mid} + \sum_q (\frac{1}{\mid V^k \mid +1}(\bar{u}_{p,q}^k(x_{q,q}^k) + \sum_{p'} \bar{u}_{q,p',q}^k(x_{q,q}^k x_{p',q}^k)) + \lambda_{p,q}x_{q,q})$$

$$\bar{E}_C^k(x,w) = \bar{\phi}_C(x_C) + \sum_q (\frac{1}{\mid V^k \mid +1}(\bar{u}_{p,q}^k(x_{q,q}^k) + \sum_{p'} \bar{u}_{q,p',q}^k(x_{q,q}^k x_{p',q}^k)) + \lambda_{C_q}x_{q,q}) \tag{14}$$

where the Lagrangian variables $\lambda = \{\{\lambda_{p,q}\}, \{\lambda_{C_q}\}\}$ are used to ensure the consistency of the solution. We impose the satisfaction of: $\lambda \in \Lambda^k = \{\lambda : \sum_{p \in S^k} \lambda_{p,q} + \lambda_{C_q} = 0, \forall C \in \mathcal{C}^k, q \in C\}$. Therefore, by design, $\bar{E}^k(x^k, w) = \sum_p \bar{E}_p^k(x,w) + \sum_C \bar{E}_C^k(x,w)$. Thus, finally, the loss function to be minimized is:

$$\min_{\{x^k \in \mathcal{X}(\mathcal{C}^k)\}, w, \{\lambda^k \in \Lambda^k\}} \tau J(w) + \sum_k \sum_{p \in V^k} \mathcal{L}_{\bar{E}_p^k} + \sum_k \sum_{C \in \mathcal{C}^k} \mathcal{L}_{\bar{E}_C^k} \tag{15}$$

3.6 Slave Problems Optimization

To optimize w, we first need to solve the slave problems by leveraging their specific structures. An essential characteristic to notice is that, for fixed $\{x^k\}$, the slaves energy can be related to CRF energies. Details of all the proofs and computation steps are provided in Supplementary Materials.

3.6.1 Optimizing over $\{\hat{x}^{k,p}\}$

Regarding the point-wise subproblems, we proceed as follows. The solution in pairwise settings has been demonstrated in [6]. In the following lemma, we generalize the solution that was proposed to a third-order context as:

Lemma 1. For fixed $p \in V^k$, let $\theta_q^k = \frac{\bar{u}_{q,q}^k(1) + \bar{u}_{q,q,q}^k(1)}{|V|^k + 1} + \lambda_{p,q}$ and $\bar{\theta}_q^k = [\theta_q^k]_+ + \bar{u}_{p,q}^k(1) + \bar{u}_{p,q,q}^k(1) + \bar{u}_{p,p,q}^k(1)$ where $[z]_+ = max(0,z)$. minimizer $\hat{x}^p$ of $\bar{E}_p^k(x,w,\lambda^k)$ is given by

$$\begin{aligned} \hat{x}_{q,q}^p &= [\theta_q^k < 0] \\ \hat{x}_{p,q}^p &= [q = \bar{q}] \; where \; \bar{q} = \arg\min_q(\bar{\theta}_q^k) \end{aligned} \tag{16}$$

3.6.2 Optimizing over $\{\hat{x}^{k,C}\}$

Regarding the cluster-wise subproblems, we proceed as follows. The solution in pairwise settings has been demonstrated in [6]. We can notice that our formulation of the cluster-wise subproblem presents a high similarity with the original formulation, and the only difference for the optimization is in θ_q^k expression:

Lemma 2. For fixed $C \in \mathcal{C}^k$, let $\theta_q^k = \frac{\bar{u}_{q,q}^k(1) + \bar{u}_{q,q,q}^k(1)}{\mid V^k \mid +1} + \lambda_{C_q}$, $\forall q \in C$. A minimizer $\hat{x}^C$ of $\bar{E}_C^k(x,w,\lambda^k)$ is given by

$$\forall q \in C, \; \hat{x}_{q,q}^C = \begin{cases} [\theta_q^k < \alpha(\mid V^k \mid - \mid C^k \mid)], \\ \quad if \; \sum_{q' \in C} [\theta_q^k - \alpha(\mid V^k \mid - \mid C^k \mid)]_- + \\ \quad \alpha \mid V^k \mid < 0 \\ 0 \; otherwise \end{cases} \tag{17}$$

with $[z]_- = min(0,z)$.

3.6.3 Optimizing over λ and w

To optimize over λ and w, we perform an iterative projected subgradient approach:

$$w \longleftarrow w - s_t\delta_w, \;\; \lambda^k \longleftarrow proj_{\Lambda^k}(\lambda^k - s_t\delta_{\lambda_k}) \tag{18}$$

with $\{\delta_w\}$ and $\{\delta_{\lambda_k}\}$ subgradient functions and $proj_{\Lambda^k}$ the projection onto Λ^k. Then, the following lemma gives the updates to be applied iteratively to efficiently obtain the approximation of w and $\{\lambda_{p,q}^k, \lambda_{C_q}^k\}$. The updates are obtained by summing the respective updates of each subproblem according to the formula provided in equation 15.

Lemma 3. Let s_t be the weight granted to the optimization at step t. We define $\hat{X}_q^k = \hat{x}_{q,q}^{k,C} + \sum_p \hat{x}_{q,q}^{k,p}$ and $\hat{X}_{p,q}^k = \hat{x}_{q,q}^{k,p}\hat{x}_{p,q}^{k,p} + \hat{x}_{q,q}^{k,C}[p = q, q \in C]$. Then, update reduces to:

$$\begin{aligned} w \; -&= s_t(\tau \nabla J(w) + \sum_k \delta_w^k) \\ \lambda_{p,q}^k \; -&= s_t(\frac{\hat{X}_q^k}{\mid V^k \mid +1} - \hat{x}_{q,q}^{k,p}) \\ \lambda_{p,q}^k \; -&= s_t(\frac{\hat{X}_q^k}{\mid V^k \mid +1} - \hat{x}_{q,q}^{k,C}) \end{aligned} \tag{19}$$

where

$$\delta_w^k = \sum_{p,p',q \in V^k} x_{p,q}^k x_{p'q}^k f_{p,p',q}^k + \sum_{p,q \in V^k} x_{p,q}^k f_{p,q}^k - (\sum_{p,q \neq p \in V^k} (\hat{x}_{p,q}^{k,p}\hat{x}_{q,q}^{k,p} f_{p,q,q}^k + \hat{x}_{p,q}^{k,p} f_{p,p,q}^k) + \sum_{p,q \neq p \in V^k} \hat{x}_{p,q}^{k,p} f_{p,q}^k + \sum_{q \in V^k} \frac{1}{\mid V^k \mid +1}(\hat{X}_q^k f_{q,q}^k + \sum_{p \in V^k} \hat{X}_{p,q}^k f_{q,p,q}^k)) \tag{20}$$

Note that $\nabla J(w)$ has to refer to a subgradient if J is non-differentiable. Besides, a constraint can be imposed over w by applying a projection during w update.

$$w \longleftarrow proj_W(w - s_t\delta_w) \tag{21}$$

where W can be any convex set of constraints included in $\mathbb{R}_+$. For instance, we mimic a true distance and satisfy the positivity constraint over w with $proj_W(z) = max(z,0)$. It will also enforce an additional sparsity on the weight vector. Notice that at least a positivity constraint has to be imposed for the resolution of the higher-order distance learning. We summarize the complete learning process in Algorithm 1.

To improve the tractability of the approach we leveraged a stochastic gradient descent (sgd) framework. It consists in randomly selecting a subset of the training samples at each iteration and performing the updates by relying only on those samples.

Algorithm 1: Learning Process

Data: training cohorts $\{V^k, \mathcal{C}^k, y^k\}$, features functions $\{f^2_{p,q}(y^k), f^3_{p,p',q}(y^k)\}$

$\lambda^k \leftarrow 0,\ \forall k$

do

  Optimize x^k: $\forall C \in \mathcal{C}^k$, $q_c = arg\min_{q \in C}(\sum_{p,p' \in C} d_{p,p',q} x_{p,q} x_{p'q} + \sum_{p \in C} d_{p,q} x_{p,q})$;

  $x^k_{p,q} = 1,\ p \in C \iff q = q_C$;

  Iterate T subgradient updates:

  repeat

    Solve slaves $\bar{E}^k_p$, $\bar{E}^k_C$ via lemmas 1, 2;

    Update w, λ^k via lemma 3

  until *T times*;

  Project w over $W \subset \mathbb{R}_+$

while *Not Convergence*;

3.7 Generalization to Higher-Order Distances

Our approach's strength is its ability to be efficiently generalized to any order. Let $h \geq 2$ be the order of the distance we are looking for. h corresponds to the maximal set size we will consider in our metric definition. Our target metric considers any set S of size $\mid S \mid \leq h$ and is defined as:

$$d_S = w^T f_{\{p\}_{p \in S}}(y) \tag{22}$$

where $f_{\{p\}_{p \in S}}$ is a positive feature function providing a closeness score on set S. This metric aims to establish a characterization of the meaningfulness to group samples in S altogether. In this case, we consider the energy defined as:

$$E(x,d)^k = \sum_{l \in [0,h-2]} \sum_{p,p_1,\dots,p_l,q} u^k_{p,p_1,\dots,p_l,q}(x_{p,q} \prod_{i \in [1,l]} x_{p_i,q}, d) + \sum_{p,q} \phi_{p,q}(x_{p,q}, x_{q,q}) + \sum_p \phi_p(x_p) + \sum_C \phi_C(x_C) - \beta|\mathcal{C}^k| \tag{23}$$

We now consider the higher order potential of order $l < h-2$, $u_{p,p_1,\dots,p_l,q}(\prod_{i \in [1,l]} x_{p_i,q}, d)$ focusing on establishing the cost of assigning $p, p_1, \dots, p_l$ to q. The potentials definitions are:

$$\begin{aligned}
&u^k_{p,p_1,\dots,p_l,q}(x_{p,q} \prod_{i \in [1,l]} x_{p_i,q}, d) = d^k_{p,p_1,\dots,p_l,q} x_{p,q} \prod_{i \in [1,l]} x_{p_i,q} \\
&d^k_{p,\prod_{i \in [1,l]} p_i} x_{p,q} \prod_{i \in [1,l]} x_{p_i,q} = w^T f_{p,p_1,\dots,p_l,q}(y^k) \\
&\phi_{p,q}(x_{p,q}, x_{q,q}) = \delta(x_{p,q} \leq x_{q,q}) \\
&\phi_p(x_p) = \delta(\sum_q x_{p,q} = 1)
\end{aligned} \tag{24}$$

First, regarding the optimization over $\{x^k\}$ for a fixed vector w. As previously, the satisfaction of the constraints induces:

$$x^k = \arg\min_{x \in \mathcal{X}(C^k)} (\sum_{l \in [0,h-2]} \sum_{p,p_1,\dots,p_l,q} d^k_{p,p_1,\dots,p_l,q} x_{p,q} \prod_{i \in [1,l]} x_{p_i,q}) \tag{25}$$

And, again, this problem's minimization only requires the set Q^k of exemplars q minimizing the above function per cluster in $\mathcal{C}^k$. Then, we assign each point of the cluster to its exemplar. In this case, the constraints will be satisfied as we ensure each cluster to have one and only one center and assign all cluster samples to this center.

As before, for each $k < K$, we define a slave problem per datapoint $p \in V^k$, $\bar{E}^k_p$, and one per cluster $C \in \mathcal{C}^k$, $\bar{E}^k_C$.

$$\begin{aligned}
&\bar{E}^k_p(x,w) = \sum_{l \in [0,h-2]} \sum_{p_1,\dots,p_l,q \neq p} u^k_{p,p_1,\dots,p_l,q}(x_{p,q} \prod_{i \in [1,l]} x_{p_i,q}, d) \\
&+ \sum_q \bar{\phi}_{p,q}(x_{p,q}) + \bar{\phi}_p(x_p) - \frac{\beta}{\mid V^k \mid} + \\
&\sum_q (\frac{1}{\mid V^k \mid +1} (\sum_{l \in [0,h-2]} \sum_{p_1,\dots,p_l} u^k_{q,p_1,\dots,p_l,q}(x_{q,q} \prod_{i \in [1,l]} x_{p_i,q}, d)) \\
&+ \lambda_{p,q} x_{q,q}) \\
&\bar{E}^k_C(x,w) = \bar{\phi}_C(x_C) + \\
&\sum_q (\frac{1}{\mid V^k \mid +1} (\sum_{l \in [0,h-2]} \sum_{p_1,\dots,p_l} u^k_{q,p_1,\dots,p_l,q}(x_{q,q} \prod_{i \in [1,l]} x_{p_i,q}, d) \\
&+ \lambda_{C_q} x_{q,q})
\end{aligned} \tag{26}$$

where the Lagrangian variables $\lambda = \{\{\lambda_{p,q}\}, \{\lambda_{C_q}\}\}$ are used to ensure the consistency of the solution. We impose the satisfaction of: $\lambda \in \Lambda^k = \{\lambda : \sum_{p \in S^k} \lambda_{p,q} + \lambda_{C_q} = 0, \forall C \in \mathcal{C}^k, q \in C\}$. Therefore, by design, $\bar{E}^k(x^k, w) = \sum_p \bar{E}^k_p(x,w) + \sum_C \bar{E}^k_C(x,w)$. Thus, finally, the lost function to be minimized is:

$$\min_{\{x^k \in \mathcal{X}(\mathcal{C}^k)\}, w, \{\lambda^k \in \Lambda^k\}} \tau J(w) + \sum_k \sum_{p \in V^k} \mathcal{L}_{\bar{E}^k_p} + \sum_k \sum_{C \in \mathcal{C}^k} \mathcal{L}_{\bar{E}^k_C} \tag{27}$$

3.7.1 *Optimizing over* $\{\hat{x}^{k,p}\}$

Regarding the point-wise subproblems, we proceed as follows. In the following lemma, we generalize the solution that was proposed in Lemma 1 to a general order setting as:

Lemma 4. For fixed $p \in V^k$, let $\theta^k_q = \frac{\sum_{l \in [0,h-2]} u^k_{q,\dots,q}(1)}{|V|^k + 1} + \lambda^k_{p,q}$ and $\bar{\theta}^k_q = [\theta^k_q]_+ + \sum_{l \in [0,h-2]} \sum_{p_1,\dots,p_l \in \{p,q\}} u^k_{p,p_1,\dots,p_l,q}$ where $[z]_+ = max(0,z)$. minimizer $\hat{x}^p$ of $\bar{E}^k_p(x,w,\lambda^k)$ is given by

$$\begin{aligned}
\hat{x}^{k,p}_{q,q} &= [\theta^k_q < 0] \\
\hat{x}^{k,p}_{p,q} &= [q = \bar{q}]\ where\ \bar{q} = \arg\min_q(\bar{\theta}^k_q)
\end{aligned} \tag{28}$$

3.7.2 *Optimizing over* $\{\hat{x}^{k,C}\}$

Regarding the cluster-wise subproblems, we generalize the Lemma 2 with:

Lemma 5. For fixed $C \in \mathcal{C}^k$, let $\theta^k_q = \frac{\sum_{l \in [0,h-2]} u^k_{q,\dots,q}(1)}{\mid V^k \mid +1} + \lambda^k_{C_q}$, $\forall q \in C$. A minimizer $\hat{x}^{k,C}$ of $\bar{E}^k_C(x,w,\lambda^k)$ is given by

$$\forall q \in C,\ \hat{x}^{k,C}_{q,q} = \begin{cases} [\theta^k_q < \alpha(\mid V^k \mid - |C^k|)], \\ \quad if \sum_{q' \in C} [\theta^k_q - \alpha(\mid V^k \mid - |C^k|)]_- + \\ \quad\quad \alpha \mid V^k \mid < 0 \\ 0\ otherwise \end{cases} \tag{29}$$

with $[z]_- = min(0,z)$.

3.7.3 Optimizing over λ and w

Lemma 6. Let s_t be the weight granted to the optimization at step t. We define $\hat{X}^k_q = \hat{x}^{k,C}_{q,q} + \sum_p \hat{x}^{k,p}_{q,q}$ and $\hat{X}^k_{p,\prod_{i\in[1,l]} p_i,q} = \hat{x}^{k,p}_{q,q} \prod_{i\in[1,l]} \hat{x}^{k,p}_{p_i q} + \hat{x}^{k,C_q}_{q,q}[p_i = q, \forall i \in [1,l]]$. Then, the updates reduce to:

$$\begin{aligned} w \; -&= s_t(\tau \nabla J(w) + \sum_k \delta^k_w) \\ \lambda^k_{p,q} \; -&= s_t(\frac{\hat{X}^k_q}{\mid V^k \mid +1} - \hat{x}^{k,p}_{q,q}) \\ \lambda^k_{C_q} \; -&= s_t(\frac{\hat{X}^k_q}{\mid V^k \mid +1} - \hat{x}^{k,C}_{q,q}) \end{aligned} \tag{30}$$

where

$$\begin{aligned} \delta^k_w = &\sum_{l\in[0,h-2]} \sum_{p,p_1,\ldots,p_l,q\in V^k} f_{p,p_1,\ldots,p_l,q}(x^k_{p,q} \prod_{i\in[1,l]} x^k_{p_i,q}, d) - \\ &\sum_{q\in V^k} (\hat{X}^k_q f^k_{q,q} + \\ &\sum_{l\in[0,h-2]} (\sum_{p\neq q\in V^k} \sum_{p_i\in\{p,q\}\forall i\in[1,l]} \hat{x}^{k,p}_{p,q}\hat{x}^{k,p}_{q,q} f^k_{p,p_1,\ldots,p_l,q} + \\ &\frac{1}{\mid V^k \mid +1} \sum_{p\in V^k} \sum_{p_i\in\{p,q\}\forall i\in[1,l]} \hat{X}^k_{p,\prod_{i\in[1,l]} p_i,q} f^k_{q,\prod_{i\in[1,l]} p_i,q})) \end{aligned} \tag{31}$$

3.8 Extension to Cluster Metrics

A final interesting addition we can bring to our higher-order distance learning framework is to consider a metric between a sample and a ground truth cluster. The difference between this particular setting and the previous higher-order metrics is that here we will consider a metric able to tackle sets of objects of different sizes (the size of the clusters) and thus will not have a defined order. The interest for such a distance is to benefit from a structural metric characterizing the closeness between a sample and a given cluster. It will be especially valuable during inference to identify the fittest cluster for a sample.

We formulate this new problem by adding to the higher-order distance defined in the previous section a term $w^T f_{p,C}(y^k)$ for any $p \in V^k$ and any $C \in \mathcal{C}^k$. Then, from the previously defined energy $E(x,d)^k$ we will define our new energy

$$E^{k*}(x,d) = E^k(x,d) + \sum_{p\in V^k} \sum_{C\in\mathcal{C}^k} w^T f_{p,C}(y^k) \sum_{q\in C} x_{p,q}$$

where we penalize the assignment of a sample p to a center q by the distance between the sample and the center's cluster according to given cluster-wise feature functions. First, regarding the optimization over $\{x^k\}$ for a fixed vector w. As previously, the satisfaction of the constraints induces to find the cluster centers q minimizing for its cluster C:

$$\begin{aligned} q = \arg\min_{q\in C} (&\sum_{l\in[0,h-2]} \sum_{p,p_1,\ldots,p_l,p\in C,p_i\in C\forall i\in[1,l]} d^k_{p,p_1,\ldots,p_l,q} x_{p,q} \\ &\prod_{i\in[1,l]} x_{p_i,q} + \sum_{p\in C} w^T f_{p,C}(y^k) x_{p,q}) \end{aligned} \tag{32}$$

x^k is inferred by assigning each sample of a cluster to the cluster center. Then, we modify the cluster-wise slave problems of the dual decomposition as follows:

$$\begin{aligned} \bar{E}^{k*}_C(x,w) = &\bar{E}^k_C(x,w) + \sum_{p\in V^k} w^T f_{p,C}(y^k) \sum_{q\in C} x_{p,q} + \\ &\frac{1}{2(\mid V^k \mid +1)} w^T f_{q,C}(y^k) \sum_{q\in C} x_{q,q} \end{aligned} \tag{33}$$

with $\bar{E}^k_C(x,w)$ the energy defined as in equation 26. Therefore, the cluster-wise slave resolution is now:

Lemma 7. For fixed $C \in \mathcal{C}^k$. Let $\theta^{k*}_q = \frac{\sum_{l\in[0,h-2]} u^k_{q,\ldots,q}(1) + f_{q,C}(y^k)}{\mid V^k \mid +1 + \lambda^{k*}_{C_q}}$, $\forall q \in V^k$. A minimizer $\hat{x}^{k,C*}$ of $\bar{E}^k_{k,C*}(x,w,\lambda^{k*})$ is given by:

$$\forall q \in C,\; \hat{x}^{k,C*}_{q,q} = \begin{cases} [\theta^{k*}_q < \alpha(\mid V^k \mid - \mid C^k \mid)], \\ \quad if \; \sum_{q'\in C} [\theta^{k*}_q - \alpha(\mid V^k \mid - \mid C^k \mid)]_- + \\ \quad \alpha \mid V^k \mid < 0 \\ 0 \; otherwise \end{cases} \tag{34}$$

Then, the updates are defined as:

Lemma 8. Let s_t be the weight granted to the optimization at step t. We consider $\hat{X}^k_q$ and $\hat{X}^k_{p,\prod_{i\in[1,l]} p_i,q}$ as defined in lemma 6.

$$\begin{aligned} w^* \; -&= s_t(\tau \nabla J(w) + \sum_k \delta^{k*}_w) \\ \lambda^{k*}_{p,q} \; -&= s_t(\frac{\hat{X}^k_q}{\mid V^k \mid +1} - \hat{x}^{k,p}_{q,q}) \\ \lambda^{k*}_{C_q} \; -&= s_t(\frac{\hat{X}^k_q}{\mid V^k \mid +1} - \hat{x}^{k,C*}_{q,q}) \end{aligned} \tag{35}$$

where

$$\begin{aligned} \delta^{k*}_w = &\delta^k_w + \sum_{p\in V^k} \sum_{C\in\mathcal{C}^k} f_{p,C}(y^k) \sum_{q\in C} x^{k*}_{p,q} - \\ &\sum_{p\in V^k} \sum_{C\in\mathcal{C}^k} f_{p,C}(y^k) \sum_{q\in C} (x^{k,C*}_{p,q} + x^{k,p}_{p,q}) + \\ &(\frac{\hat{X}^k_q}{\mid V^k \mid +1} \sum_{q\neq p\in C} (x^{k,C*}_{p,q} + x^{k,p}_{p,q})) \end{aligned} \tag{36}$$

3.9 Extracting and Leveraging Structural Information from Data

Several approaches exist in order to design a graph structure on data set with no natural graph representation. Here, we relied on a distance matrix between objects computed as the sum over all the different feature functions used in the distance learning. Then, a k-nearest neighbors approach was computed, meaning that there is an edge between two objects p and q iff p (resp. q) is in the k objects the closest of q (resp. p).

Once a graph structure obtained, we studied different ways of leveraging their properties. Our first, most simple, approach is considering the shortest path $S_{p,q}$ between objects p and q in the graph. The distance between those objects will then be the weighted length $L_{p,q}$ of such a

No balanced error function, No path	Balanced Accuracy			Weighted Precision			Weighted Sensitivity			Weighted Specificity		
	Training	Validation	Test	Training	Validation	Test	Training	Validation	Test	Training	Validation	Test
Average	1	0.6	0.42	1	0.76	0.35	1	0.62	0.4	1	0.58	0.43
Minimum	1	0.52	0.4	1	0.75	0.27	1	0.54	0.38	1	0.49	0.42
Maximum	1	0.59	0.38	1	0.68	0.32	1	0.63	0.37	1	0.6	0.39
Min center	1	0.62	0.34	1	0.75	0.26	1	0.6	0.32	1	0.57	0.35
KNN	1	0.62	0.41	1	0.69	0.29	1	0.64	0.4	1	0.61	0.43
Balanced error function, No path	**Balanced Accuracy**			**Weighted Precision**			**Weighted Sensitivity**			**Weighted Specificity**		
	Training	Validation	Test	Training	Validation	Test	Training	Validation	Test	Training	Validation	Test
Average	1	0.71	0.61	1	0.72	0.62	1	0.72	0.6	1	0.71	0.62
Min	1	0.66	0.62	1	0.68	0.62	1	0.67	0.62	1	0.65	0.63
Min Max	1	0.65	0.58	1	0.65	0.59	1	0.65	0.56	1	0.65	0.59
Min center	1	0.74	0.62	1	0.75	0.64	1	0.75	0.61	1	0.74	0.63
KNN	1	0.72	0.6	1	0.73	0.62	1	0.72	0.58	1	0.73	0.61
Balanced error function, Path	**Balanced Accuracy**			**Weighted Precision**			**Weighted Sensitivity**			**Weighted Specificity**		
	Training	Validation	Test	Training	Validation	Test	Training	Validation	Test	Training	Validation	Test
Average	1	1	0.77	1	1	0.84	1	1	0.73	1	1	0.82
Min	1	1	0.76	1	1	0.83	1	1	0.71	1	1	0.81
Min Max	1	1	0.79	1	1	0.85	1	1	0.75	1	1	0.83
Min center	1	1	0.78	1	1	0.8	1	1	0.73	1	1	0.8
KNN	1	1	0.69	1	1	0.81	1	1	0.63	1	1	0.75

Table 1: Results with the synthetic dataset of the different experiments in the second-order settings for the various inference strategies.

path. The generalization of this method for a set of objects $\{p_1, ..., p_l\}$ and a potential center q is defined as follows:

$$SP_l(p_1, ..., p_l) = \sum_{i \in \{1,...,l\}} \sum_{j \in \{i+1,...,l\}} \frac{L_{p_i,q} + L_{p_j,q}}{L_{p_i,p_j}}$$

The interpretation of this graph metric is that the center of a cluster q has to be a hub for the objects of its cluster i.e. the ratio between the shortest path and the shortest path passing by q has to be small for any pair of objects in the cluster.

Similarly, we considered the eccentricity of a set of objects $\{p_1, ..., p_l\}$ as a l-order graph metric. We deem a set of objects has to have a small maximal weighted graph diameter to belong to the same cluster.

Then, we considered two connectivity related metrics. The first one is based on the clique order of a set of objects $CO(p_1, ..., p_l)$ and is defined as $max\_degree(G) - CO(p_1, ..., p_l)$ with $max\_degree(G)$ the maximal degree of a node in the whole graph. By doing so we consider that the bigger the clique order in the set of objects the more relevant their association in a cluster.

The second metric is based on the connectivity resilience $CR(p_1, ..., p_l)$ which is the minimal number of nodes to remove to disconnect the set of objects. The metric is defined as $l - CR(p_1, ..., p_l)$.

3.10 Leveraging a Task Dedicated Distance for Classification

In order to perform the classification, we relied on a K-Nearest Neighbors framework. Once the distance learnt, we labeled a new sample as the ground truth cluster from which it is the closest. We experimented and compared different strategies to determine the closest cluster:

- Average distance to the points of the cluster.
- Minimum distance to the points of the cluster.
- Maximum distance to the points of the cluster.
- Distance to the center of the cluster.
- Majoritarian cluster of the k-nearest neighbors.

The distance between the new sample p and objects of the cluster C is computed using the learnt dedicated distance. For $l > 2$-order distances, we compute the distance on the set $\{p, p_1, ..., p_{l-2}, q\}$ where we iterate over all possible sets $\{p_1, ..., p_{l-2}\} \in C^{l-2}$ and q is the cluster center discovered during the learning step.

4 Implementation Details

We implemented the algorithm proposed in [6] and we used it as a baseline. It is available at https://github.com/ebattistella/Second-order-Distance-learning. Besides, the adaptation to general higher-order distances we propose in this study has been implemented and is available at https://github.com/ebattistella/Higher-order-Distance-Learning-GHOST-.

To prove the relevance of our higher-order formulation, we leveraged two datasets of a very different nature. First, we synthesized a dataset with samples in dimension 100 with 60 noisy dimensions. Clusters are designed by considering 100 samples generated from Gaussian distributions with different variances and means between the two clusters on the non-noisy dimensions. The noise is simulated by taking a much larger variance. Ideal graphs were generated on this dataset as one clique per cluster with no connection between cliques. Then, we added noise to the graphs using a rewiring method [30]. For each pair of nodes, we added or removed an edge with a probability of p. We considered values of $p \in [0, 0.5]$ with an increment of 0.1. We generated a training, a validation, and a test sets considering different variances and means. For each set, we considered base variances randomly chosen for each feature between 0 and 200 shifted by respectively 10, 30 and 50 for the non-noisy dimensions and 1000, 2000 and 10000 for the noisy ones. Regarding the means, we considered base means randomly chosen for each feature between -50 and 50 shifted by respectively 0, 10, and 50. The aim was here to visualize the generalizability of the learned weights and their resilience to increasing noise. We then leveraged a Covid-19 dataset introduced in [31]. We used the same training and testing sets as the authors and compared their results to the classification performance of our proposed Ghost approach over the Severe/Non-severe staging task. A graph on the data

was obtained through the method proposed in Section 3.9 by considering the 5 closest neighbors.

The second-order feature functions we based all our experiments on are feature-wise euclidean distances. In addition, to assess the influence of different metrics, we considered as meta-features Euclidean, Minkowski, City-block, Cosine, Correlation, Hamming, Jaccard, Chebyshev, Matching, Yule, Braycurtis, Dice, Kulsinski, Russellrao, Pearson-correlation based, Spearman-correlation based, Kendall-correlation based distances on the full feature space.

For those datasets, we performed a thorough set of experiments to highlight the relevance of our higher-order distance formulation and the consideration of graph structures. We first used the simple pairwise distance defined using the basic formulation from [6]. Then, we complemented this with our balanced error function to better account for the cluster size. We finally added a shortest-paths-based metric SP_2 to assess the value of graph information even in second-order settings and compare it with the higher-order. We performed the higher-order distance learning using the combination of the second-order meta-features with the different higher-order metrics defined in Section 3.9. We considered the third-order approach, the cluster metric approach, and the combination of both approaches which constitutes our Ghost approach.

5 Results and Discussion

5.1 Synthesized Dataset

In this subsection, we considered the two synthesized clusters. This experiment aimed to assess the capacity of our higher-order framework to leverage information from a graph according to its level of noise and combine it with usual second-order metrics to perform classification. We used as a baseline the second-order framework performances with and without considering path length information in a graph. First, we reported the results in the second-order without graph information nor balanced error function in Table 1. Here, we reported the performance of the different strategies to infer the label of a new sample defined in Section 3.10. We observed superior results of the distance to cluster center approach on all metrics. This trend was consistent in the different experiments we performed. Thus, for concision sake's in the following, we only reported results from this strategy.

Table 2 presents the comparison of the performance using the different frameworks defined in this study. The second-order results with path length information have been obtained with the ideal graph without noise. The foremost point to notice is the greater performance and lesser overfitting of the methods leveraging graph information when resorting to graph with a rewiring probability below 0.4. Although, it is worth mentioning that the second-order non-graph-based meta-features do not compensate for the noise brought by the graph-based meta-features for the frameworks relying on graph information with p above 0.4. Then, notice that whereas the cluster and the third-order frameworks alone performed similarly, their combination reported higher results.

5.2 Covid-19 Dataset

Table 3 presents the results of our Ghost framework over the Covid-19 dataset introduced in [31] and compares them to the results obtained by the authors using an Ensemble approach for both feature selection and classification. We also compared our results with an approach investigating the use of deep learning-based representations to integrate the imaging information. By leveraging the 3D contouring model proposed in [31], we extracted a deep feature space of dimension 128 of the lung. The same classification ensemble approach used in [31] with classical imaging features is used. In addition, we report the results of the automated machine learning tool TPOT [32] which relies on genetic algorithms to find the best machine learning pipeline, including feature selection and classification models. This experiment highlights the interest of considering graph information as it improves the results over the basic framework. Also, it not only enables to outperforms an ensemble method over several standard classifiers leveraging carefully chosen features obtained with both field experts knowledge and an ensemble data-driven framework but also the results obtained using deep learning representation-based features and a classification pipeline defined by a state of the art genetic algorithm.

6 Conclusion

This paper proposed a novel distance learning framework to leverage higher-order information, including graph patterns and cluster-based metrics, towards a dedicated to the task metric definition. Moreover, we demonstrated the value of leveraging graph-based information for classification. In particular, we have highlighted the interest in designing a graphical representation of data to extract structural information. Also, we studied the relevance of higher-order metrics and experimented several directions to better account for structures in the data. We have proved the relevance of our approach on a challenging Covid-19 patient stratification task and present results outperforming ensemble approaches for feature selection and classification on common classical and deep learning representations-based imaging features as well as a dedicated pipeline automatically designed by a state of the art genetic algorithm. In the future, we aim at studying other kinds of data with known graphical representations as PPI networks. We also want to study more intricate metrics to better exploit higher-order information as for instance Mahalanobis distance or graph-density based information.

Acknowledgment

The authors would like to thank ARC sign'it program and Siric Socrates INCA. This work was partially supported by the Fondation pour la Recherche Médicale (FRM; no. DIC20161236437) and by the ARC sign'it grant: Grant SIGNIT201801286.

References

[1] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proceedings of the 20th international conference on machine learning (ICML-03)*, 2003, pp. 856–863.

Third-order only	Balanced Accuracy			Weighted Precision			Weighted Sensitivity			Weighted Specificity		
p	Training	Validation	Test	Training	Validation	Test	Training	Validation	Test	Training	Validation	Test
0	1	1	1	1	1	1	1	1	1	1	1	1
0.1	0.97	0.91	0.96	0.97	0.92	0.96	0.97	0.92	0.96	0.97	0.91	0.96
0.2	0.9	0.92	0.91	0.9	0.93	0.92	0.89	0.92	0.9	0.9	0.93	0.92
0.3	0.59	0.58	0.64	0.78	0.77	0.79	0.61	0.6	0.61	0.63	0.62	0.67
0.4	0.62	0.62	0.68	0.78	0.78	0.8	0.59	0.58	0.65	0.66	0.65	0.71
0.5	0.54	0.52	0.48	0.57	0.54	0.48	0.51	0.5	0.46	0.57	0.55	0.5
Cluster metric only	**Balanced Accuracy**			**Weighted Precision**			**Weighted Sensitivity**			**Weighted Specificity**		
p	Training	Validation	Test	Training	Validation	Test	Training	Validation	Test	Training	Validation	Test
0	1	1	1	1	1	1	1	1	1	1	1	1
0.1	0.97	0.91	0.96	0.97	0.92	0.96	0.97	0.92	0.96	0.97	0.91	0.96
0.2	0.9	0.92	0.91	0.9	0.93	0.92	0.89	0.92	0.9	0.9	0.93	0.92
0.3	0.7	0.66	0.64	0.78	0.77	0.79	0.7	0.63	0.61	0.63	0.62	0.67
0.4	0.61	0.6	0.63	0.76	0.75	0.8	0.6	0.59	0.6	0.67	0.66	0.7
0.5	0.53	0.48	0.5	0.55	0.47	0.5	0.5	0.47	0.46	0.5	0.49	0.5
Ghost	**Balanced Accuracy**			**Weighted Precision**			**Weighted Sensitivity**			**Weighted Specificity**		
p	Training	Validation	Test	Training	Validation	Test	Training	Validation	Test	Training	Validation	Test
0	1	1	1	1	1	1	1	1	1	1	1	1
0.1	0.98	0.91	0.97	0.97	0.92	0.96	0.98	0.92	0.95	0.98	0.91	0.97
0.2	0.91	0.93	0.91	0.9	0.93	0.92	0.89	0.92	0.9	0.9	0.93	0.92
0.3	0.71	0.65	0.65	0.79	0.76	0.8	0.71	0.63	0.62	0.62	0.61	0.68
0.4	0.62	0.62	0.68	0.78	0.78	0.8	0.6	0.59	0.6	0.67	0.66	0.7
0.5	0.54	0.52	0.48	0.57	0.54	0.48	0.51	0.5	0.46	0.57	0.55	0.5

Table 2: Results on the synthetic dataset of the different higher-order experiments with the distance to center inference strategy.

Framework	Balanced Accuracy		Weighted Precision		Weighted Sensitivity		Weighted Specificity	
	Training	Test	Training	Test	Training	Test	Training	Test
Ghost	**0.67**	**0.71**	**0.78**	**0.8**	**0.69**	**0.73**	**0.65**	**0.69**
Ensemble	0.73	0.7	0.82	0.81	0.67	0.64	0.8	0.77
Deep Features	0.71	0.68	0.8	0.79	0.72	0.72	0.71	0.64
TPOT	0.84	0.64	0.87	0.76	0.82	0.71	0.86	0.56

Table 3: Performance of the different learning frameworks over the Covid-19 dataset. The first row is the Ghost framework, the following 3 rows are the performance of the ensemble of standard classifiers on classical imaging features as defined in [31], of the ensemble of standard classifiers on deep learning representations-based imaging features and of a feature selection and classification pipeline learnt by the genetic algorithm defined in [32].

[2] R. Sun, E. J. Limkin, M. Vakalopoulou, L. Dercle, S. Champiat, S. R. Han, L. Verlingue, D. Brandao, A. Lancia, and S. A. et al., "A radiomics approach to assess tumour-infiltrating CD 8 cells and response to anti-PD-1 or anti-PD-l1 immunotherapy: an imaging biomarker, retrospective multicohort study," *The Lancet Oncology*, vol. 19, no. 9, pp. 1180–1191, sep 2018.

[3] E. Battistella, M. Vakalopoulou, R. Sun, T. Estienne, M. Lerousseau, S. Nikolaev, E. A. Andres, A. Carré, S. Niyoteka, C. Robert *et al.*, "Cancer gene profiling through unsupervised discovery," *arXiv preprint arXiv:2102.07713*, 2021.

[4] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, "Distance metric learning with application to clustering with side-information," in *NIPS*, vol. 15, no. 505–512. Citeseer, 2002, p. 12.

[5] S. Xiang, F. Nie, and C. Zhang, "Learning a mahalanobis distance metric for data clustering and classification," *Pattern Recognition*, vol. 41, no. 12, pp. 3600–3612, Dec. 2008. [Online]. Available: https://doi.org/10.1016/j.patcog.2008.05.018

[6] N. Komodakis, "Learning to cluster using high order graphical models with latent variables," in *2011 International Conference on Computer Vision*. IEEE, Nov. 2011. [Online]. Available: https://doi.org/10.1109/iccv.2011.6126227

[7] U. Demšar, O. Špatenková, and K. Virrantaus, "Identifying critical locations in a spatial network with graph theory," *Transactions in GIS*, vol. 12, no. 1, pp. 61–82, 2008.

[8] D. Krioukov, K. Fall, and X. Yang, "Compact routing on internet-like graphs," in *IEEE INFOCOM 2004*, vol. 1. IEEE, 2004.

[9] E. Battistella and L. Cholvy, "Modelling and simulating extreme opinion diffusion," in *International Conference on Agents and Artificial Intelligence*. Springer, 2018, pp. 79–104.

[10] R. Vermeulen, E. L. Schymanski, A.-L. Barabási, and G. W. Miller, "The exposome and health: Where chemistry meets biology," *Science*, vol. 367, no. 6476, pp. 392–396, 2020.

[11] D. Szklarczyk, J. H. Morris, H. Cook, M. Kuhn, S. Wyder, M. Simonovic, A. Santos, N. T. Doncheva, A. Roth, P. Bork *et al.*, "The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible," *Nucleic acids research*, p. gkw937, 2016.

[12] D. K. Lê-Huu and N. Paragios, "Alternating direction graph matching," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 4914–4922.

[13] L. Torresani, V. Kolmogorov, and C. Rother, "Feature correspondence via graph matching: Models and global optimization," in *European conference on computer vision*. Springer, 2008, pp. 596–609.

[14] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" *arXiv preprint arXiv:1810.00826*, 2018.

[15] S. E. Schaeffer, "Graph clustering," *Computer Science Review*, vol. 1, no. 1, pp. 27–64, Aug. 2007. [Online]. Available: https://doi.org/10.1016/j.cosrev.2007.05.001

[16] H. Yin, A. R. Benson, J. Leskovec, and D. F. Gleich, "Local higher-order graph clustering," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 555–564. [Online]. Available: https://doi.org/10.1145/3097983.3098069

[17] A. R. Benson, D. F. Gleich, and J. Leskovec, "Higher-order organization of complex networks," *Science*, vol. 353, no. 6295, pp. 163–166, Jul. 2016. [Online]. Available: https://doi.org/10.1126/science.aad9029

[18] A. Grover and J. Leskovec, "node2vec: Scalable feature learning

for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 855–864.
[19] R. Lambiotte, M. Rosvall, and I. Scholtes, "From networks to optimal higher-order models of complex systems," *Nature Physics*, vol. 15, no. 4, pp. 313–320, Mar. 2019. [Online]. Available: https://doi.org/10.1038/s41567-019-0459-y
[20] E. Battistella, M. Vakalopoulou, T. Estienne, M. Lerousseau, R. Sun, C. Robert, N. Paragios, and E. Deutsch, "Gene expression high-dimensional clustering towards a novel, robust, clinically relevant and highly compact cancer signature," in *IWBBIO 2019*, Granada, Spain, May 2019. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02076104
[21] Z. Yu, Z. Kuang, J. Liu, H. Chen, J. Zhang, J. You, H.-S. Wong, and G. Han, "Adaptive ensembling of semi-supervised clustering solutions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 8, pp. 1577–1590, 2017.
[22] K. Wagstaff, "Refining inductive bias in unsupervised learning via constraints," in *AAAI/IAAI*, 2000, p. 1112.
[23] I. Davidson and S. Ravi, "Clustering with constraints: Feasibility issues and the k-means algorithm," in *Proceedings of the 2005 SIAM international conference on data mining*. SIAM, 2005, pp. 138–149.
[24] M. T. Law, R. Urtasun, and R. S. Zemel, "Deep spectral clustering learning," in *International conference on machine learning*. PMLR, 2017, pp. 1985–1994.
[25] T. Finley and T. Joachims, "Supervised clustering with support vector machines," in *Proceedings of the 22nd international conference on Machine learning*, 2005, pp. 217–224.
[26] A. Fix, A. Gruber, E. Boros, and R. Zabih, "A graph cut algorithm for higher-order markov random fields," in *2011 International Conference on Computer Vision*. IEEE, 2011, pp. 1020–1027.
[27] H. Ishikawa, "Transformation of general binary mrf minimization to the first-order case," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 6, pp. 1234–1249, 2010.
[28] N. Komodakis, B. Xiang, and N. Paragios, "A framework for efficient structured max-margin learning of high-order mrf models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 7, pp. 1425–1441, 2014.
[29] N. Komodakis, N. Paragios, and G. Tziritas, "Mrf energy minimization and beyond via dual decomposition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 3, pp. 531–552, 2010.
[30] N. Jarman, E. Steur, C. Trengove, I. Y. Tyukin, and C. Van Leeuwen, "Self-organisation of small-world networks by adaptive rewiring in response to graph diffusion," *Scientific Reports*, vol. 7, no. 1, pp. 1–9, 2017.
[31] E. Battistella, G. Chassagnon, M. Vakalopoulou, S. Christodoulidis, T.-N. Hoang-Thi, S. Dangeard, E. Deutsch, F. Andre, E. Guillo, N. Halm *et al.*, "Ai-driven quantification, staging and outcome prediction of covid-19 pneumonia," *Medical Image Analysis*, vol. 67, p. 101860, 2021.
[32] T. T. Le, W. Fu, and J. H. Moore, "Scaling tree-based automated machine learning to biomedical big data with a feature set selector," *Bioinformatics*, vol. 36, no. 1, pp. 250–256, 2020.

Enzo Battistella holds a MSc and a computer science engineer degree from Telecom Paris. He is currently pursuing a PhD in CentraleSupelec and Gustave Roussy about gene clustering and gene signature designing. He is under the joint supervision of Eric Deutsch and Nikos Paragios.

Maria Vakalopoulou is an Assistant Professor at MICS laboratory of CentraleSupelec, University Paris-Saclay, Paris, France. Prior to that and during 2017 – 2018, she was a postdoctoral researcher at the same university. She received her PhD, in 2017, from the National Technical University of Athens, Athens, Greece from where she also received her Engineering Diploma degree, graduating with excellence. Her research interests include medical imagery, remote sensing, computer vision, and machine learning. The researcher has published her research in international journals (Lancet Oncology, Radiology, European Radiology) and conferences (MICCAI, ISBI, IGARSS).

Nikos Paragios received his B.Sc. and M.Sc.degrees from the University of Crete, Gallos, Greece,in 1994 and 1996, respectively, his Ph.D. degree from I.N.R.I.A./University of Nice/Sophia Antipolis, Nice,France, in 2000 and his HDR (Habilitation à Diriger des Recherches) degree from the same university in 2005. He is a Professor of applied mathematics at CentraleSupelec/ Universit Paris-Saclay Orsay, France. He has published more than 200 papers in the areas of computer vision, biomedical imaging, and machine learning. Prof. Paragios is the Editor-in-Chief of the Computer Vision and Image Understanding journal and serves on the editorial board of the Medical Image Analysis and SIAM Journal on Imaging Sciences. He is a Senior Fellow of the Institut Universitaire de France and a Member of the Scientific Council of Safran conglomerate.

Éric Deutsch is an oncologist-radiotherapist, head of the Radiotherapy Department of Gustave Roussy and director of the joint research unit "Molecular Radiotherapy" (Inserm - University Paris-XI).

His research has been dedicated to combination of novel anti-cancer drugs either used alone or in combination to radiotherapy. Prof. Deutsch's interest focuses on cell death mechanisms, inflammation and immunotherapy, on the clinical side, he is versed into the field of translational research and early clinical trials. He has investigated several first in human novel drugs-radiotherapy combinations such as mTOR inhibitors, antiviral agents, immune modifiers and nanoparticles. Research axes: Early clinical trials and phase I, nanoparticles, oligometastasis, radiation biology, HPVs related tumors, combination with new drugs and Radiotherapy.

In addition, his clinical research focuses on developing new irradiation processes to better target and eliminate tumors. By using technological innovations such as radiomics, computational imaging, guidance by medical imaging or novel anti-cancer agents Prof. Deutsch works with his medical team on increasingly personalized protocols that can adapt to very specific cancers. Together with members of the research unit he leads, Prof. Deutsch studies the biological effects of irradiation. The objectives of this research are first to identify the biological characteristics of tumors, "tumor markers", to predict their sensitivity or resistance to radiotherapy.