Efficient Kernel Learning from Constraints and Unlabeled Data

Mahdieh Soleymani Baghshah
Computer Engineering Department
Sharif University of Technology
Tehran, Iran
soleyman@ce.sharif.edu

Saeed Bagheri Shouraki
Electrical Engineering Department
Sharif University of Technology
Tehran, Iran
Bagheri-s@sharif.edu

Abstract—Recently, distance metric learning has been received an increasing attention and found as a powerful approach for semi-supervised learning tasks. In the last few years, several methods have been proposed for metric learning when must-link and/or cannot-link constraints as supervisory information are available. Although many of these methods learn global Mahalanobis metrics, some recently introduced methods have tried to learn more flexible distance metrics using a kernel-based approach. In this paper, we consider the problem of kernel learning from both pairwise constraints and unlabeled data. We propose a method that adapts a flexible distance metric via learning a nonparametric kernel matrix. We formulate our method as an optimization problem that can be solved efficiently. Experimental evaluations show the effectiveness of our method compared to some recently introduced methods on a variety of data sets.

Keywords: kernel learning; pairwise constraints; efficient method; metric; semi-supervised; optimization problem

I. INTRODUCTION

Practical data mining is usually related to the situations where limited supervisory information together with unlabeled data is available [1]. In some clustering tasks, it is possible to obtain supervision in the form of pairwise must-link and cannot-link constraints that are more general than labels. This problem is known as semi-supervised clustering [2]. Many methods have been introduced for semi-supervised clustering tasks in the last decade.

Distance metric learning has been received a lot of research interest in the last decade [1] and considered as one of the most popular approaches for semi-supervised learning settings [3]. In the last few years, several algorithms [4-13] have been proposed for metric learning when side information in the form of must-link (positive) and cannot-link (negative) constraints is available. Many of the existing methods learn global Mahalanobis metrics or equivalently linear transformations [4-7]. Recently, some methods have been introduced for learning more flexible distance metrics [8-13]. Most of these methods [9-13] have used a kernel-based approach. Hoi et al. [11] and Zhuang et al. [12] have introduced the most flexible kernel learning methods. These methods compared to the presented methods in [9, 10] that learn a linear combination of some base kernels are much more powerful and capable of considering the complexity in the clusters structure. Although these methods have assessed the ability of learning non-parametric kernel matrices, their optimization problems are not such appropriate for metric learning from pairwise constraints [13]. In our previous work [13], we proposed a more appropriate optimization problem that the introduced one in [11]. However, this problem was a Semi-Definite Programming (SDP) problem and standard SDP solvers can require as high as $O(n^6)$ computational complexity.

In this paper, we propose a semi-supervised metric learning method that seeks an appropriate distance metric by learning a proper non-parametric kernel matrix. We present an optimization problem that uses pairwise constraints and unlabeled data and it can be solved more efficiently than the existing problems in [11, 13]. Although our method is relatively related to the previous work [11, 12] introduced for kernel learning, we propose a novel optimization problem that is more appropriate for distance metric learning. We will show empirically that the proposed method is more effective. Compared to our previous work in [13], the proposed method in this paper is more (computationally) effective and appropriate for larger scale problems.

The rest of this paper is organized as follows: In Section 2, we propose a kernel-based metric learning method by introducing an optimization problem and solving the proposed problem efficiently. Section 3 presents experimental results. Concluding remarks are given in the last section.

II. PROPOSED KERNEL LEARNING METHOD

We are given a set of data points $X = \{x_i\}_{i=1}^n \subset \mathbb{R}^d$ and two sets $S$ and $D$ including positive and negative constraints. We intend to learn a PSD kernel matrix $K \in \mathbb{R}^{n \times n}$ according to the pairwise constraints and the topological structure of the data. By learning a kernel matrix, indeed we find a similarity measure that can capture the complicity in the structure of the data. Let $X = \{x_1, \ldots, x_n\}$ be the matrix containing all the data points, $\Phi = [\phi(x_1), \ldots, \phi(x_n)]$ be the transformed data to the kernel space, and $K = [k(x_i, x_j)]_{i,j=1}^n = \Phi^T \Phi$ be the corresponding kernel matrix. Following our previous method [13], we propose a new optimization problem. We use a kernel capacity constraint that allows the problem to be solved much more efficiently than the introduced problems in [11, 13] that require standard SDP solvers.
We use the idea of the graph Laplacian [14] to preserve the topological structure of the data. The $k$-nearest neighbors graph is used to model the relation between close data points (i.e., its weight matrix is obtained by setting $w_{ij} = 1$ if $x_i$ is amongst the $k$-nearest neighbors of $x_j$, or vice versa). According to the spectral graph theory, the smoothness of $\phi(x)$ applied on graph vertices can be defined as:

$$
Y(\phi, X) = \frac{1}{2} \sum_{i,j=1}^{n} \left\| \phi(x_i) - \sqrt{d_i} \phi(x_j) \right\|_2^2 w_{ij}
$$

with the ideal kernel whose elements are 1 for similar pairs and -1 for dissimilar pairs.

Although, our optimization problem in (3) is different from the introduced problem in [12], we can use similar approach to the presented one in [12] to solve the proposed problem. We consider the dual problem of the SDP problem in (3) and construct the following Lagrangian function:

$$
\mathcal{L}(\lambda, K) = tr\left( (E_x + \alpha L)K \right) - \sum_{(i,j) \in \mathcal{D}} \lambda_{ij} \left( k_{ii} + k_{jj} - 2k_{ij} - c \right)
$$

where $\lambda_{ij} \geq 0$ are the Lagrangian multipliers. Thus, we can form this min-max problem:

$$
\max_{K \succeq 0} \quad tr(\mathcal{K}K) \leq b, \quad \forall (i, j) \in \mathcal{D}, \quad \lambda_{ij} \geq 0.
$$

Here, we present a proposition that enable us to solve the above min-max problem.

**Proposition 1** [12]. Given $A$ is any symmetric matrix such that $\Lambda = \mathbf{V} \Lambda \mathbf{V}^T$, where columns of $\mathbf{V}$ are orthonormal eigenvectors of $A$ and $\Lambda$ is a diagonal matrix containing corresponding eigenvalues. Let, $b$ be a positive constant, the optimal solution to the following SDP problem

$$
\max_{K \succeq 0} \quad tr(\mathcal{K}K) \leq b,
$$

can be expressed as the closed-form

$$
K^* = A \frac{b}{tr(A \Lambda A^T)},
$$

where $A = \mathbf{V} \Lambda \mathbf{V}^T$ and $\Lambda = \max(\Lambda, 0)$.

Based on Proposition 1, for fixed $\lambda = [\lambda_{ij}]$, we can set $A = \sum_{(i,j) \in \mathcal{D}} \lambda_{ij} (e_i - e_j)(e_i - e_j)^T - (E_x + \alpha L)$ and find the optimal kernel matrix using (7). Thus, similar strategy to the introduced one in [12] can be used to solve Problem (5); 1) Compute the closed-form solution $K^*$ for a fixed $\lambda^*$ using (7); 2) Update $\lambda^*$ by setting $\lambda^* = (\lambda^* - \eta_j (\partial L/\partial \lambda_j))_+$; 3) Repeat Step 1-2 until convergence. When $\eta_j$ is small enough or $\eta_j = O(1/t)$, the whole optimization problem is guaranteed to converge [12]. Table 1 shows the steps of the proposed kernel learning method.

The most costly step of our algorithm is the eigen-decomposition of an $n \times n$ matrix (required in Step 7) that must be computed until convergence or when maxiter = 50 is reached. Computing the eigen-decomposition for a dense
n × n matrix is $O(n^3)$ (sparse spectral decomposition can be performed more quickly). Therefore, the time complexity of our method is very low compared to that of the introduced methods in [11, 13] needing the SDP solvers whose time complexity can be as high as $O(n^{6.5})$ [12].

## III. EXPERIMENTS

In this section, we conduct experiments on synthetic and real-world data sets and evaluate the proposed method by comparing the following algorithms:

1) $k$-means without metric learning (Euclidean);
2) $k$-means with the Xiang et al.’s metric learning method [7] (Xiang’s);
3) $k$-means with the LLMA [8] method for metric learning (LLMA);
4) kernel $k$-means with the kernel obtained by the kernel-β method [9] (Kernel-β);
5) kernel $k$-means with the kernel learned by Hoi et al.’s method [11] (Hoi’s);
6) kernel $k$-means with the kernel learned by the proposed method (Proposed);

The parameters of our method have been set to $k = 5$, $\alpha = 0.4$, $c = 1$, and $b = 1$. The number of positive and negative constraints is set to be equal $nc = |S'| = |D|$. We generate 20 different $S'$ and $D$ sets for each data set and run the clustering algorithms 20 times with different random initializations for each set of constraints. All data sets are normalized before use in the clustering algorithms. To measure the performance in our experiments, we use the modified Rand index that has been introduced in [4] and widely used for evaluating the performance of metric learning methods from pairwise constraints [7-10, 13].

In Figs. 1(a)-(c), three synthetic data sets have been displayed (the data points belonging to the same class are shown with the same color and style). Figs. 1(d)-(f) show the results of applying different methods on these data sets as box-plots. According to this figure, our method generally yields the best results and outperforms Hoi’s on DS3. Moreover, the time complexity of our method is much lower than that of Hoi’s as mentioned in Section 2.

We conduct further experiments on real-world UCI data sets (available at: http://archive.ics.uci.edu/ml): Soybean (47/35/4), Protein (116/20/6), Iris (150/4/3), Wine (178/13/3), Sonar (208/60/2), Heart (270/13/2), Ionosphere (351/34/2), Boston housing (506/13/3). The numbers inside the parentheses ($n/d/c$) show the number of data points $n$, the number of features $d$, and the number of classes $c$. Fig. 2 displays the results of different methods on the UCI data sets. As we can see in this figure, our method generally provides better results than the other methods. By comparing our method with Hoi’s, we find that it is clearly better than Hoi’s on six out of the eight data sets (while it is also much more time efficient than Hoi’s as explained in Section 2).

## IV. CONCLUSION

In this paper, we proposed a novel kernel learning method using unlabeled data and pairwise constraints. This method is formulated as an optimization problem that appropriately incorporates positive and negative constraints to learn a proper distance metric. We used a kernel capacity constraint allowing us to solve the problem efficiently. Experimental results validated that the performance of the proposed kernel learning method for semi-supervised clustering is at least comparable to that of some existing methods while our method is computationally effective.

## REFERENCES


Figure 1. Three synthetic data sets and clustering results of different methods (numbered as in Section 3) on them.

Figure 2. Average Rand index curves of different methods on the UCI data sets.