Pattern Recognition 43 (2010) 2982–2992

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Non-linear metric learning using pairwise similarity and dissimilarity constraints and the geometrical structure of data

Mahdieh Soleymani Baghshah [a,*], Saeed Bagheri Shouraki [b]

[a] Computer Engineering Department, Sharif University of Technology (SUT), Azadi St., PO Box: 1458889694, Tehran, Iran
[b] Electrical Engineering Department, Sharif University of Technology, Tehran, Iran

ARTICLE INFO

Article history:
Received 18 December 2008
Received in revised form
18 January 2010
Accepted 26 February 2010

Keywords:
Metric learning
Positive and negative constraints
Semi-supervised clustering
Optimization problem
Non-linear
Topological structure
Kernel

ABSTRACT

The problem of clustering with side information has received much recent attention and metric learning has been considered as a powerful approach to this problem. Until now, various metric learning methods have been proposed for semi-supervised clustering. Although some of the existing methods can use both positive (must-link) and negative (cannot-link) constraints, they are usually limited to learning a linear transformation (i.e., finding a global Mahalanobis metric). In this paper, we propose a framework for learning linear and non-linear transformations efficiently. We use both positive and negative constraints and also the intrinsic topological structure of data. We formulate our metric learning method as an appropriate optimization problem and find the global optimum of this problem. The proposed non-linear method can be considered as an efficient kernel learning method that yields an explicit non-linear transformation and thus shows out-of-sample generalization ability. Experimental results on synthetic and real-world data sets show the effectiveness of our metric learning method for semi-supervised clustering tasks.

1. Introduction

Distance metrics are a key issue in many machine learning algorithms [1]. Over the past few years, there has been considerable research on distance metric learning [2]. Many of the earlier studies optimize the metric with class labels for classification tasks [3–8]. More recently, researchers have given much attention to distance learning for semi-supervised clustering tasks. As class label information is not generally available for clustering tasks, constraints are used as more natural supervisory information for these tasks. Pairwise similarity (positive) and dissimilarity (negative) constraints are the most popular kind of side information that has been used for semi-supervised clustering. However, other kinds of side information like relative comparisons have also been considered in some studies.

Over the last few years, the problem of clustering with side information (semi-supervised clustering) has received increasing attention [9,10] and distance learning has been considered as a powerful approach for this problem. The two most frequently used approaches that include side information in the clustering algorithms are *constraint-based* [11–21] and *distance function learning* [22–34] approaches [24]. In the former approach, the clustering algorithm itself is modified to use the available labels or constraints to bias the search for an appropriate data clustering. However, in the latter approach, the algorithm learns a distance function prior to clustering. The learned distance function tries to put similar points close together and dissimilar points far away from each other. This approach is more flexible in the choice of distance function [33]. Additionally, it has received considerable attention in recent studies [1,25,28–31,33,34] and we also use this approach.

Distance learning based on constraints has been studied by many researchers [22–34]. Klein et al. [22] introduced a metric adaptation method for semi-supervised clustering. This method finds a distance measure according to the shortest path in a version of the similarity graph that has been altered by positive constraints. However, negative constraints have been employed after the metric adaptation phase during the complete-link clustering. Some latter studies [1,23,25,28,34] have considered a more popular approach that learns a global Mahalanobis metric from pairwise constraints. Xing et al. [23] proposed a convex optimization problem to learn a global Mahalanobis metric according to pairwise constraints. Bar-Hillel et al. [25] devised a more efficient, non-iterative algorithm called *relevant component analysis* (RCA) for learning a Mahalanobis metric. This method can only incorporate positive constraints. An extension of the RCA method that can consider both positive and negative constraints has also been introduced by Yeung and Chang [28].

* Corresponding author. Tel.: +98 21 6616 4642; fax: +98 21 6601 9246.
E-mail addresses: soleyman@ce.sharif.edu, mahdiesoleymani@yahoo.com (M. Soleymani Baghshah), bagheri-s@sharif.edu (S. Bagheri Shouraki).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.02.022

More recently, some non-linear metric learning methods for semi-supervised clustering have been introduced. Chang and Yeung [29] proposed a locally linear metric learning method that considers only positive constraints. The objective function of this method has many local optima and the topology cannot be preserved well during this approach [30]. Chang and Yeung [31] proposed also a metric adaptation method. This method adjusts the location of data points iteratively, so that similar points tend to get closer and dissimilar points tend to move away from each other. As this method lacks an explicit transformation map, it cannot project new data points onto the transformed space straightforwardly [31]. Additionally, the movement of data points in this method may interfuse the structure of the data. In [30], two kernel-based metric learning methods have been presented that do have some limitations [30]. These kernel-based methods can use only positive constraints.

Among the existing metric learning methods, some of them [1,23,28,34,39,40] can incorporate both positive and negative constraints. However, most of these methods [1,23,28,34] learn only a linear transformation that corresponds to a Mahalanobis metric. Although some recent studies [39,40] have been introduced for kernel learning from positive and negative constraints, they are based on learning non-parametric kernel matrices. These methods can only find distances of the seen data. Additionally, the optimization problems in these methods are usually difficult to solve [40] and the degree of freedom of the corresponding models is very high (i.e., n^2 where n denotes the number of data points). In this paper, we propose an efficient non-linear metric learning method that considers both positive and negative constraints and also the topological structure of the data. We formulate the proposed method as a constrained trace ratio optimization problem that can be solved efficiently using algorithms introduced for this purpose (e.g., Xiang et al.'s method [1]). The proposed non-linear method can be considered as an efficient kernel learning method that does not need to learn all items of an $n \times n$ matrix. Our method yields an explicit transformation that can project new data points onto the transformed space.

The rest of this paper is organized as follows: Section 2 presents a brief review of related works. In Section 3, first the general form of the proposed optimization problems that incorporate both positive and negative constraints and also the topological structure of the data is introduced. Then, we present special problems that can be solved efficiently for learning linear and non-linear transformations. Finally, we present a kernel-based method and show the relation between the proposed non-linear method and a special form of this kernel-based method. Section 4 presents some experimental results on synthetic and real-world data sets. Concluding remarks are given in the last section.

2. Related works

In this section, we review those methods that can consider both positive and negative constraints to learn a transformation. A positive constraint denotes a pair of data points that must be in the same cluster while a negative constraint denotes two data points that must be in two different clusters [1]. Most of the existing methods that can use both positive and negative constraints learn a Mahalanobis metric $\mathbf{A}$ (where $\mathbf{A}$ is a positive semi-definite matrix) or, equivalently, find a transformation matrix $\mathbf{W}$ ($\mathbf{y}=\mathbf{W}^T\mathbf{x}$). Learning the transformation matrix $\mathbf{W}$ can yield the Mahalanobis metric $\mathbf{A}=\mathbf{W}\mathbf{W}^T$ according to:

$$\begin{aligned}\|\mathbf{y}_i-\mathbf{y}_j\|^2 &= (\mathbf{x}_i-\mathbf{x}_j)^T\mathbf{W}\mathbf{W}^T(\mathbf{x}_i-\mathbf{x}_j)\\ &= (\mathbf{x}_i-\mathbf{x}_j)^T\mathbf{A}(\mathbf{x}_i-\mathbf{x}_j)=\|\mathbf{x}_i-\mathbf{x}_j\|_{\mathbf{A}}^2.\end{aligned} \tag{1}$$

Xing et al. introduced the first metric learning method using both positive and negative constraints [23]. They presented the following objective function:

$$\begin{aligned}&\min_{\mathbf{A}} \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in P} \|\mathbf{x}_i-\mathbf{x}_j\|_{\mathbf{A}}^2,\\ &\text{s.t.} \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in D} \|\mathbf{x}_i-\mathbf{x}_j\|_{\mathbf{A}}^2 \geq 1,\\ &\mathbf{A}\geq\mathbf{0},\end{aligned} \tag{2}$$

where P is the set of positive constraints and D is the set of negative constraints. Xing et al. [23] used the gradient descent and the idea of iterative projection to solve this problem. Although the above optimization problem is convex, it is a hard problem to solve and the introduced solution in [23] is slow and somewhat unstable [25].

Chang et al. [28] introduced an extended version of the RCA [25] method. They proposed the transformation matrix $\mathbf{W}=(\mathbf{S}_b)^{1/2}(\mathbf{S}_w)^{1/2}$ where $\mathbf{S}_b$ denotes the inter-class (cluster) covariance matrix computed from negative constraints and $\mathbf{S}_w$ shows the intra-class (cluster) covariance matrix computed from positive constraints. Although this transformation can be found easily, the singularity problem may occur during the calculation of $\mathbf{S}_w^{-1/2}$. Additionally, it has not been obtained as a solution of an optimization problem.

Hoi et al. [34] proposed the *discriminative component analysis* (DCA) method using the ratio of determinants as the objective function:

$$\mathbf{W}^*=\arg\max_{\mathbf{W}}\frac{|\mathbf{W}^T\widehat{\mathbf{C}}_b\mathbf{W}|}{|\mathbf{W}^T\widehat{\mathbf{C}}_w\mathbf{W}|}, \tag{3}$$

where $\widehat{\mathbf{C}}_b$ shows the covariance between data of discriminative chunklets (cannot-links) and $\widehat{\mathbf{C}}_w$ shows the total covariance of data within the same chunklet (must-links) [34]. This problem can be solved analytically by the eigenvalue decomposition of $\widehat{\mathbf{C}}_w^{-1}\widehat{\mathbf{C}}_b$ [1]. However, the singularity problem may occur during the calculation of $\widehat{\mathbf{C}}_w^{-1}$ [34]. To avoid the singularity problem, DCA diagonalizes the covariance matrices $\widehat{\mathbf{C}}_b$ and $\widehat{\mathbf{C}}_w$ simultaneously and discards the eigenvectors corresponding to the zero eigenvalue [1].

Recently, Xiang et al. [1] introduced the trace ratio objective function (with the constraint $\mathbf{W}^T\mathbf{W}=\mathbf{I}$) as a more appropriate objective function:

$$\mathbf{W}^*=\arg\max_{\mathbf{W}^T\mathbf{W}=\mathbf{I}}\frac{tr(\mathbf{W}^T\mathbf{S}_b\mathbf{W})}{tr(\mathbf{W}^T\mathbf{S}_w\mathbf{W})}, \tag{4}$$

where $\mathbf{S}_w$ is the covariance matrix computed from positive constraints and $\mathbf{S}_b$ is the covariance matrix obtained from negative constraints. The constraint $\mathbf{W}^T\mathbf{W}=\mathbf{I}$ has been introduced to avoid degenerate solutions [1]. The optimization problem in (4) is similar to the problem introduced by Guo et al. [35] as the *generalized Foley–Sammon transform* (GFST). It seeks a transformation matrix in the global sense instead of learning individual transformation vectors for different dimensions like Fisher criterion. To solve the above optimization problem, Xiang et al. [1] have developed an iterative method exploring the optimum in way of binary search. Additionally, they have found a lower bound and an upper bound including the optimum to speed up the search. Their proposed method provides a heuristic search to solve the problem presented in (4) [1]. In this paper, we propose a generalized form of the objective function presented in (4) that can learn a non-linear transformation and also considers the topological structure of data.