Subclass-Oriented Dimension Reduction with Constraint Transformation and Manifold Regularization

Bin Tong and Einoshin Suzuki

Graduate School of Systems Life Sciences, Kyushu University, Japan
{bintong,suzuki}@i.kyushu-u.ac.jp

Abstract. We propose a new method, called Subclass-oriented Dimension Reduction with Pairwise Constraints (SODRPaC), for dimension reduction on high dimensional data. Current linear semi-supervised dimension reduction methods using pairwise constraints, e.g., must-link constraints and cannot-link constraints, can not handle appropriately the data of multiple subclasses where the points of a class are separately distributed in different groups. To illustrate this problem, we particularly classify the must-link constraint into two categories, which are the *inter-subclass must-link constraint* and the *intra-subclass must-link constraint*, respectively. We argue that handling the *inter-subclass must-link constraint* is challenging for current discriminant criteria. Inspired by the above observation and the cluster assumption that nearby points are possible in the same class, we carefully transform must-link constraints into cannot-link constraints, and then propose a new discriminant criterion by employing the cannot-link constraints and the compactness of shared nearest neighbors. For the reason that the local data structure is one of the most significant features for the data of multiple subclasses, manifold regularization is also incorporated in our dimension reduction framework. Extensive experiments on both synthetic and practical data sets illustrate the effectiveness of our method.

1 Introduction

In various applications, such as gene expression, image retrieval, etc., one is often confronted with high dimensional data [1]. Dimension reduction, which maps data points in a high-dimensional space into those in a low-dimensional space, is thus viewed as one of the most crucial preprocessing steps of data analysis. Dimension reduction methods can be divided into three categories, which are supervised ones [2], unsupervised ones[3], and semi-supervised ones[4]. The input data in these three categories are labeled data, unlabeled data, and both of them, respectively. In a typical real-world application, only a small number of labeled data points are available due to the high cost to obtain them [4]. Hence the semi-supervised dimension reduction may be considered to fit into the practical setting. Instead of labeled data points, some semi-supervised methods assume pairwise constraints, for it is easier for experts to specify them than to assign the class labels of data points. More specifically speaking, pairwise constraints consist of must-link constraints and cannot-link constraints. The pair of data points in

M.J. Zaki et al. (Eds.): PAKDD 2010, Part II, LNAI 6119, pp. 1–13, 2010.

a must-link constraint shares the same class label, while the pair of data points in a cannot-link constraint is given different class labels.

From another viewpoint, dimension reduction methods can be divided into nonlinear and linear ones. The former allows a nonlinear transformation in the mapping while the latter restricts itself to linear transformation. We consider a complex distribution of points that are distributed in multiple subclasses. In other words, the data points of one class form several separated clusters. A nonlinear method has a higher degree of freedom and hence can handle data with complex distribution effectively while a linear method tends to be incompetent in such a case.

In this paper, we restrict our attention to the linear semi-supervised dimension reduction for the data of multiple subclasses with pairwise constraints. Previously relevant methods [5] [6] [7] [8] implicitly assume that a class consists of a single cluster. If the points are of multiple subclasses, handling the pairwise constraints to project the points into multiple subclasses in the transformed space is challenging for linear dimension reduction. For a deep analysis, we particularly classify the must-link constraint into two categories. If two points in a must-link constraint reside in a single subclass, we define such a must-link constraint as an *intra-subclass must-link constraint*. On the contrary, if two points in a must-link constraint come from different subclasses, we define such kind of must-link constraint as an *inter-subclass must-link constraint*. We attribute the reason of the improper behavior of current linear methods to the fact that the *inter-subclass must-link constraint* most probably confuses the discriminant criteria of existing methods. The problem resulted from the *inter-subclass must-link constraint* is also encountered by the constraint transformation. For instance, a method in [9] transforms multiple must-link constraints, which are connected via points in two different classes, into a cannot-link constraint between the centroids of the points of two classes. This method fails to give a comprehensible meaning if the points belong to different subclasses because the centroids may fall into the region of another class.

To overcome above problems, we propose SODRPaC, which consists of two steps. In the first step, must-link constraints which satisfy several conditions are transformed into cannot-link constraints and the remaining must-link constraints are deleted. The idea behind this step is to reduce the harmfulness of the *inter-subclass must-link constraints* while exploiting the must-link constraint information as much as possible by respecting the cluster assumption [10]: nearby points on the same manifold structure in the original space are likely to belong to the same class. In the second step, we obtain a projection mapping by inventing a new discriminant criterion for dimension reduction, which is suitable for the data of multiple subclasses, and employing the manifold regularization [11], which is helpful for discovering the local structure of data.

2 Problem Setting and Motivation

The problem setting is defined as follows. We are given a set of N points $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, where $\mathbf{x}_i$ represents a point in a d-dimensional space, a set of must-link constraints $\mathbf{M} = \{m_1, m_2, \ldots, m_{N_{\mathrm{ML}}}\}$, and a set of cannot-link constraints $\mathbf{C} = \{c_1, c_2, \ldots, c_{N_{\mathrm{CL}}}\}$. Here m_i consists of a pair of points belonging to the same class while c_i consists of a pair of points belonging to different classes. The output is a

$d \times l$ transformation matrix W which consists of l projective vectors $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_l\}$ ($l \ll d$). W maps $\mathbf{x}_1$,$\mathbf{x}_2$,...,$\mathbf{x}_N$ to a set of lower dimensional points $Y = \{\mathbf{y}_1, ..., \mathbf{y}_N\}$. Hence $\mathbf{y}_i = W^T\mathbf{x}_i$ where $\mathbf{y}_i$ represents a point in a l-dimensional space. After making data projection, we only consider the classification task in the transformed space. For avoiding the bias caused by the choice of the classification method, the accuracy of nearest neighborhood (1-NN) classifier is considered as a measurement for the goodness of dimension reduction with the 20×5-fold cross validation.

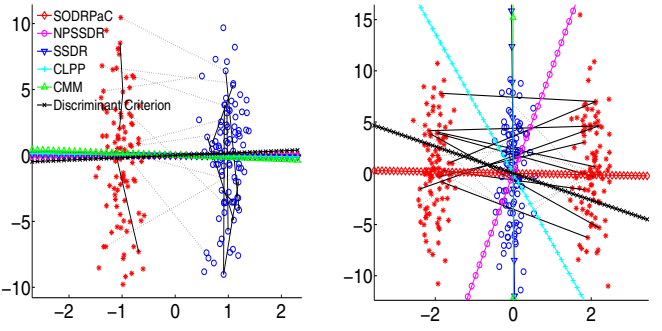


Fig. 1. Motivating examples. The data points are of Gaussian distribution. In (a), the blue and red points are distributed in different clusters. In (b), the red points reside in different subclasses. Must-link constraints and cannot-link constraints are denoted by black solid and dashed lines, respectively.

Fig. 1 presents the motivating examples, where $d = 2$ and $l = 1$. The task for dimension reduction here is thus to project the two dimensional data onto a line, where the points from different classes can be differentiated. A horizontal line is supposed to be the best projection while a vertical one is the worst projection. To better illustrate the motivation of our method, previously relevant methods are firstly retrospected. In the aspect of pairwise constraints, SSDR [5] and CMM [6] are to maximize the average distance between the points in cannot-link constraints, and to minimize the average distance between the points in must-link constraints simultaneously. We can see that minimizing the average distance between the points in must-link constraints is reasonable in the case shown in Fig. 1a, where all the must-link constraints are *intra-subclass must-link constraints*. However, it disturbs to maximize the average distance between the points in cannot-link constraints in the case shown in Fig. 1b, where all the must-link constraints are *inter-subclass must-link constraints*. CLPP [7] builds an affinity matrix, each entry of which indicates the similarity between two points. To utilize the constraint information, the affinity matrix is revised by setting the similarity degree between non-neighboring points involved in pairwise constraints. For example, given a must-link constraint, the similarity degree between two points is revised to be 1, indicating two points are close (similar) to each other, no matter the two points are distant (dissimilar) or not. Suppose that the must-link constraint is *inter-subclass must-link constraint*, it implies that the two points are not geometrically nearby each other. This

arbitrary updating may damage the geometrical structure of data points. This problem is also confronted by NPSSDR [8]. The above analysis explains the reason why CMM, SSDR, CLPP and NPSSDR are capable of obtaining excellent performance as shown in Fig. 1a, while they fail to reach the same fine performance in the multiple subclass case shown in Fig. 1b.

In the light of observations, we argue that the *inter-subclass must-link constraint* is probably harmful for the discriminant criteria of existing methods. For this reason, we attempt to design a new discriminant criterion that is able to behave appropriately in the case of multiple subclasses. The new discrimination criterion marked as 'Discriminant Criterion' is able to obtain almost the same performance as others, as shown in Fig. 1a, and can even outperform previous methods, as shown in Fig. 1b. Moreover, the manifold regularization is helpful for discovering the local structure of data which is considered as one of the most principle characteristics of the data of multiple subclasses [12]. We therefore consider to make the new discriminant criterion and the manifold regularization work together in a collaborative way. Fig. 1b also demonstrates that our method SODRPaC, which is the combination of the new discrimination criterion and the manifold regularization, can obtain the best performance.

3 Subclass-Oriented Dimension Reduction with Pairwise Constraints

The overview of our SODRPaC involves two steps described as follows:

(1) ***Transformation***. This step transforms must-link constraints into cannot-link constraints under the cluster assumption.
(2) ***Dimension reduction***. This step includes two components. The first component is the new discriminant criterion suitable for the case of multiple subclasses. The other one is the manifold regularization, which helps discovering the local structure of data.

3.1 Transformation from Must-Link Constraints

Although a method that transforms must-link constraints into cannot-link constraints is provided in [9], we would point out that its purpose that the plentiful amount of constraints are reduced is substantially different from ours. Moreover, it becomes ineffective due to the *inter-subclass must-link constraint*. In a high dimensional space, the boundaries of subclasses and the number of subclasses within one class can not be explicitly detected by using the unlabeled data and link constraints. Thus, it is difficult to identify whether a must-link constraint is of *inter-subclass must-link constraint* or not. To reduce the harmfulness of *inter-subclass must-link constraints*, removing all the must-link constraints is, therefore, the most straightforward way. However, it can be regarded as a waste of must-link constraint information. Preserving the useful must-link constraints as much as possible in the form of cannot-link constraints is then the fundamental idea behind the transformation.

In our method, the transformation from must-link constraints into cannot-link constraints basically occurs when a must-link constraint and a cannot-link constraint are

connected. Under the cluster assumption, it is natural to consider two nearby points as another form of must-link constraint, so that we have more opportunities to transform must-link constraints into cannot-link constraints. In this paper, we employ shared nearest neighbor (SNN) [13] to formulate the sense of 'nearby' points. A set of shared nearest neighbors is denoted by $\mathbf{N}_S = \{\mathbf{N}_S^{\mathbf{x}_1}, \mathbf{N}_S^{\mathbf{x}_2}, \ldots, \mathbf{N}_S^{\mathbf{x}_N}\}$ where $\mathbf{N}_S^{\mathbf{x}_i} = \{\{\mathbf{x}_i, \mathbf{x}_j\} | \mathbf{x}_i \in \mathbf{N}(\mathbf{x}_j), \mathbf{x}_j \in \mathbf{N}(\mathbf{x}_i)\}$. $\mathbf{N}(\mathbf{x}_i)$ denotes the k nearest neighbors set of $\mathbf{x}_i$. Let $|\mathbf{N}_S|$ be the number of the pairs of shared nearest neighbors, where $|\cdot|$ denotes the cardinality of a set. The value of SNN between $\mathbf{x}_i$ and $\mathbf{x}_j$ is defined as the number of points shared by their neighbors $SNN(i,j) = |\mathbf{N}(\mathbf{x}_i) \cap \mathbf{N}(\mathbf{x}_j)|$. The larger the value of SNN between two points is, the closer the two points are. It should be noted that we design a $N \times N$ matrix $\mathbf{L}$ to specify a kind of reliability for cannot-link constraints, which could be also deemed as the trustiness to them. Suppose that all the previously specified constraints are correct, for the previously given cannot-link constraints and the generated cannot-link constraints by using must-link constraints, their reliabilities are set to be 1. For the generated cannot-link constraints by using shared nearest neighbors, their reliabilities are equal to the similarities between the shared nearest neighbors. It is because that transformation by employing shared nearest neighbors are considered to be less trustful than that by using must-link constraints. We believe it is natural to take the similarity between the shared nearest neighbors as a measurement for the trustiness. For example, given a pair of shared nearest neighbors $\{\mathbf{x}_i, \mathbf{x}_j\}$, we represent the reliability of a generated cannot-link constraint by using it as a Gaussian kernel, which is a simple kernel and has been widely applied in research fields. The reliability is formulated as $\theta(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\gamma)$, where $\|\cdot\|$ denotes the Euclidian norm and γ is the kernel parameter. Note that, for the convenient access to the matrix $\mathbf{L}$, given a cannot-link constraint $c = \{\mathbf{x}_i, \mathbf{x}_j\}$, we use $\mathbf{L}(c)$ to denote the entries of $\mathbf{L}_{ij}$ and $\mathbf{L}_{ji}$, thus $\mathbf{L}$ is a symmetric matrix.

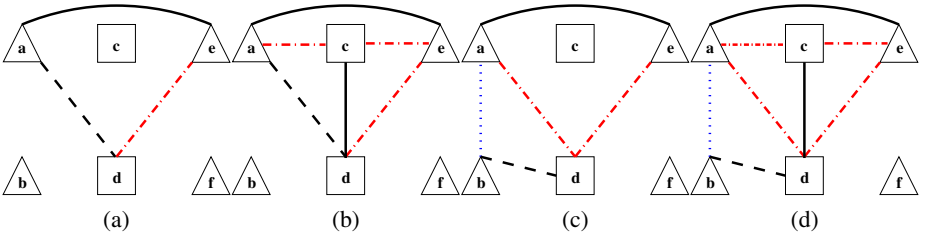


Fig. 2. Four simple cases for the transformation. The previously specified must-link constraints and cannot-link constraints are denoted by the black solid and dashed lines, respectively. The shared nearest neighbor is presented as the blue dotted line. The red dash-dotted line specifies the new cannot-link constraint.

Fig. 2 shows four fundamental scenarios of the transformation. The set $\{\mathrm{a}, \mathrm{b}, \mathrm{e}, \mathrm{f}\}$, and $\{\mathrm{c}, \mathrm{d}\}$ represent different classes of data points. We explain these four scenarios in a metaphorical way where the must-link constraint is taken as a friend relationship while the cannot-link constraint is considered as an enemy one. Standing from the viewpoint of point 'a', it is given a friend relationship, say $\{\mathrm{a}, \mathrm{e}\}$, as shown in Fig. 2a, which is

called as a basic form. If 'd' is my enemy, instead of keeping my friend 'e', consider that 'e' is the enemy of my enemy 'd'. Fig. 2b shows an extension of the basic form with an enemy's friend rule. If my enemy 'd' has a friend 'c', 'c' is the enemy of my friend 'e' and me. In these two cases, the reliabilities for the new enemy relationships are set to be 1. Fig. 2c presents an extension of the basic form, which is called as a proximity form. If I have no enemy but my neighbor 'b' has an enemy 'd', 'd' is the enemy of my friend 'e' and me. Fig. 2d shows an extension of the proximity form with the enemy's friend rule. Note that, in the latter two cases, the reliabilities for the new enemy relationships are set to be the similarity between my neighbor 'b' and me. The pseudo code for the summary of these four cases is illustrated in Algorithm 1.

Algorithm 1. Transformation from Must-link Constraints into Cannot-link Constraints

Input: $\mathbf{M}$, $\mathbf{C}$, k, γ.
Output: $\mathbf{C}$, $\mathbf{L}$.

create a $N \times N$ zero matrix $\mathbf{L}$.
for each $c \in \mathbf{C}$ **do**
  $\mathbf{L}(c) = 1$.
end for
if $\exists\, c \in \mathbf{C}, m \in \mathbf{M}$ s.t. $m \cap c \neq \emptyset$ **then**
  define $\mathrm{a} \in m \cap c$, $\mathrm{e} \in m - m \cap c$, $\mathrm{d} \in c - m \cap c$.
  create a new cannot-link constraint $c' = \{\mathrm{d}, \mathrm{e}\}$; if $c' \notin \mathbf{C}$ then $\mathbf{C} \leftarrow \mathbf{C} \cup \{c'\}$, $\mathbf{L}(c') = 1$.
  if $\exists\, m' \in \mathbf{M}$ s.t. $\mathrm{d} \in m'$, $m' \neq \{\mathrm{d}, \mathrm{e}\}$ **then**
    define $\mathrm{c} \in m' - m' \cap c$.
    create two new cannot-link constraints $c'_1 = \{\mathrm{a}, \mathrm{c}\}$, $c'_2 = \{\mathrm{e}, \mathrm{c}\}$; for each c'_i, $i = 1, 2$, if $c'_i \notin \mathbf{C}$, then $\mathbf{C} \leftarrow \mathbf{C} \cup \{c'_i\}$, $\mathbf{L}(c'_i) = 1$.
  end if
end if
if $\exists\, m \in \mathbf{M}, c \in \mathbf{C}, \forall \mathrm{a} \in m, \forall \mathbf{n}^{\mathrm{a}}_S \in \mathbf{N}^{\mathrm{a}}_S$ s.t. $c \notin \mathbf{N}^{\mathrm{a}}_S$, $c \cap \mathbf{n}^{\mathrm{a}}_S \neq \emptyset$, $\mathrm{a} \notin c \cap \mathbf{n}^{\mathrm{a}}_S$ **then**
  define $\mathrm{d} \in c - c \cap \mathbf{n}^{\mathrm{a}}_S$, $\mathrm{e} \in m - m \cap c$.
  create two new cannot-link constraints $c'_1 = \{\mathrm{a}, \mathrm{d}\}$, $c'_2 = \{\mathrm{e}, \mathrm{d}\}$ and $r = \theta(\mathrm{a}, \mathrm{b})$; for each c'_i, $i = 1, 2$, if $c'_i \notin \mathbf{C}$, then $\mathbf{C} \leftarrow \mathbf{C} \cup \{c'_i\}$, $\mathbf{L}(c'_i) = r$.
  if $\exists\, m' \in \mathbf{M}$ s.t. $\mathrm{d} \in m'$ and $m' \neq \{\mathrm{d}, \mathrm{e}\}$ **then**
    define $\mathrm{c} \in m' - m' \cup c$.
    create two new cannot-link constraints $c'_1 = \{\mathrm{a}, \mathrm{c}\}$, $c'_2 = \{\mathrm{e}, \mathrm{c}\}$ and $r = \theta(\mathrm{a}, \mathrm{b})$; for each c'_i, $i = 1, 2$, if $c'_i \notin \mathbf{C}$, then $\mathbf{C} \leftarrow \mathbf{C} \cup \{c'_i\}$, $\mathbf{L}(c'_i) = r$.
  end if
end if

3.2 Dimension Reduction

In this section, we explain the dimension reduction which is based on a novel discriminant criterion and the manifold regularization. As mentioned in section 2, minimizing the average distance between the points in must-link constraints is inappropriate when the must-link constraints are *inter-subclass must-link constraints*. Under the cluster assumption, the shared nearest neighbors could be naturally deemed as another kind of *intra-subclass must-link constraints*. Thus, minimizing the average distance between the points in *intra-subclass must-link constraints* could be relaxed as making the shared

nearest neighbors closer in the transformed space. Furthermore, the pair of points in the shared nearest neighbors probably resides in the same subclass, such that this relaxation would not suffer from the harmfulness of *inter-subclass must-link constraints*. Therefore, the discriminant criterion, which maximizes the average distance between the points in cannot-link constraints and minimizes the average distance between the shared nearest neighbors, is expected to be suitable for the data of multiple subclasses.

Suppose that $\mathbf{x}_i$ and $\mathbf{x}_j$ are projected to the image y_i^k and y_j^k along the direction $\mathbf{w}_k$, the new discriminant criterion is defined as follows:

$$\partial(\mathbf{w}_k) = \sum_{i,j:\{\mathbf{x}_i,\mathbf{x}_j\}\in\mathbf{C}} \mathbf{L}_{ij} \frac{\left\|y_i^k - y_j^k\right\|^2}{2|\mathbf{C}|} - \sum_{i,j:\{\mathbf{x}_i,\mathbf{x}_j\}\in\mathbf{N}_S} \mathbf{H}_{ij} \frac{\left\|y_i^k - y_j^k\right\|^2}{2|\mathbf{N}_S|} \tag{1}$$

where the elements of $\mathbf{H}$ are given below:

$$\mathbf{H}_{ij} = \begin{cases} SNN(i,j), & \{\mathbf{x}_i,\mathbf{x}_j\} \in \mathbf{N}_S^{\mathbf{x}_i} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

Inspired by the local scatter [14], the intuition behind the latter part of the right side of Eq. 1 could be regarded as the compactness of shared nearest neighbors, since two points are more likely to be close if the value of SNN between them is large. The difference from the local scatter lies in the fact that a weighted matrix $\mathbf{H}$ which handles the similarity degree between shared nearest neighbors is employed. Since SNN provides a robust property that the side effect caused by the noisy points could be reduced to some degree, the compactness of shared nearest neighbors is more reliable than that of local scatter. The compactness of shared nearest neighbors could be also re-written as follows:

$$\begin{aligned} \sum_{i,j:\{\mathbf{x}_i,\mathbf{x}_j\}\in\mathbf{N}_S} \mathbf{H}_{ij} \frac{\left\|y_i^k - y_j^k\right\|^2}{2|\mathbf{N}_S|} &= \frac{1}{2|\mathbf{N}_S|} \sum_i \sum_j \mathbf{H}_{ij} (\mathbf{w}_k^T \mathbf{x}_i - \mathbf{w}_k^T \mathbf{x}_j)^2 \\ &= \mathbf{w}_k^T \left[\frac{1}{2|\mathbf{N}_S|} \sum_i \sum_j \mathbf{H}_{ij} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T \right] \mathbf{w}_k \\ &= \mathbf{w}_k^T S_1 \mathbf{w}_k \end{aligned} \tag{3}$$

where $S_1 = \frac{1}{2|\mathbf{N}_S|} \sum_i \sum_j \mathbf{H}_{ij} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^T$. S_1 then could be computed as follows:

$$\begin{aligned} S_1 &= \frac{1}{2|\mathbf{N}_S|} \left(\sum_i \sum_j \mathbf{H}_{ij} \mathbf{x}_i \mathbf{x}_i^T + \sum_i \sum_j \mathbf{H}_{ij} \mathbf{x}_j \mathbf{x}_j^T - 2 \sum_i \sum_j \mathbf{H}_{ij} \mathbf{x}_i \mathbf{x}_j^T \right) \\ &= \frac{1}{|\mathbf{N}_S|} \left(\sum_i \mathbf{D}_{ii} \mathbf{x}_i \mathbf{x}_i^T - \sum_i \sum_j \mathbf{H}_{ij} \mathbf{x}_i \mathbf{x}_j^T \right) \\ &= \frac{1}{|\mathbf{N}_S|} \left(X \mathbf{D} X^T - X \mathbf{H} X^T \right) \end{aligned} \tag{4}$$

where $\mathbf{D}$ is a diagonal matrix whose entries are column sums of $\mathbf{H}$, $\mathbf{D}_{ii} = \sum_j \mathbf{H}_{ij}$. Similarly, the first part of right hand of Eq. 1 could be reformulated as:

$$\sum_{i,j:\{\mathbf{x}_i,\mathbf{x}_j\}\in\mathbf{C}} \mathbf{L}_{ij}\frac{\left\|y_i^k - y_j^k\right\|^2}{2|\mathbf{C}|} = \frac{1}{2|\mathbf{C}|}\sum_i\sum_j \mathbf{L}_{ij}(\mathbf{w}_k^T\mathbf{x}_i - \mathbf{w}_k^T\mathbf{x}_j)^2$$
$$= \mathbf{w}_k^T S_2 \mathbf{w}_k \tag{5}$$

where $S_2 = \frac{1}{|\mathbf{C}|}\left(X\mathbf{G}X^T - X\mathbf{L}X^T\right)$ where $\mathbf{G}$ is a diagonal matrix whose entries are column sums of $\mathbf{L}$, $\mathbf{G}_{ii} = \sum_j \mathbf{L}_{ij}$. Then, $\partial(\mathbf{w}_k)$ can be briefly written as:

$$\partial(\mathbf{w}_k) = \mathbf{w}_k^T X(\mathbf{P} - \mathbf{Q})X^T\mathbf{w}_k \tag{6}$$

where $\mathbf{P} = \mathbf{D} - \mathbf{H}$, and $\mathbf{Q} = \mathbf{G} - \mathbf{L}$. For all the $\mathbf{w}_k$, $k = 1, ..., l$, we can arrive at

$$\partial = tr\left[W^T X(\mathbf{P} - \mathbf{Q})X^T W\right] \tag{7}$$

where tr denotes the trace operator. As illustrated in Fig. 1b, the manifold regularization [11] is helpful for enhancing the performance obtained by the new discriminant criterion. We therefore incorporate it into our dimension reduction framework. The manifold regularization is defined as:

$$\xi = tr\left[W^T X\mathbf{M}X^T W\right] \tag{8}$$

where $\mathbf{M} = \mathbf{I} - \mathbf{K}^{-1/2}\mathbf{U}\mathbf{K}^{-1/2}$ is defined as a normalized graph Laplacian. $\mathbf{I}$ is a unit matrix, and $\mathbf{K}$ is a diagonal matrix whose entries are column sums of $\mathbf{U}$, $\mathbf{K}_{ii} = \sum_j \mathbf{U}_{ij}$, where $\mathbf{U}$ is defined as follows:

$$\mathbf{U}_{ij} = \begin{cases} \exp(\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^2/\gamma), & \mathbf{x}_i \in \mathbf{N}(\mathbf{x}_j) \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

ξ is expected to be minimized in order to preserve the sub-manifold of data. At last, the final objective function that combines Eq. 7 and Eq. 8 together is expected to be maximized, and is derived as

$$\underset{\substack{W\in\mathbb{R}^{d\times l} \\ \text{s.t.} W^T W = \mathbf{I}}}{\arg\max} \; tr\left[W^T X(\mathbf{P} - \mathbf{Q} - \lambda\mathbf{M})X^T W\right] \tag{10}$$

where λ is a parameter to control the impact of manifold regularization. By introducing the Lagrangian, the objective function is given by the maximum eigenvalue solution to the following generalized eigenvector problem:

$$X(\mathbf{P} - \mathbf{Q} - \lambda\mathbf{M})X^T\mathbf{w} = \phi\mathbf{w} \tag{11}$$

where ϕ is the eigenvalue of $\mathbf{P} - \mathbf{Q} - \lambda\mathbf{M}$, and $\mathbf{w}$ is the corresponding eigenvector. One may argue that, when the graph of SNN is equal to the k-NN graph of the manifold regularization, $\mathbf{Q}$ is almost equivalent to $\mathbf{M}$ on preserving the local structure. As shown in [13], this situation would rarely happen since the two types of graph are dramatically different from each other in the general case. Moreover, to minimize the average distance between the shared nearest neighbors, which are considered as another form of must-link constraints, is conceptually distinct from preserving the local structure.

4 Evaluation by Experiments

4.1 Experiments Setup

We use public data sets to evaluate the performance of SODRPaC. Table 1 summarizes the characteristics of the data sets. All the data come from the UCI repository [15] except for GCM [16] that is of very high dimensionality. For the 'monks-1', 'monks-2', and 'monks-3' data, we combined the train and test sets into a whole one. For the 'letter' data, we chose 'A', 'B', 'C', and 'D' letters from the train and test sets respectively by randomly picking up 100 samples for each letter, and then assembled them into a whole set.

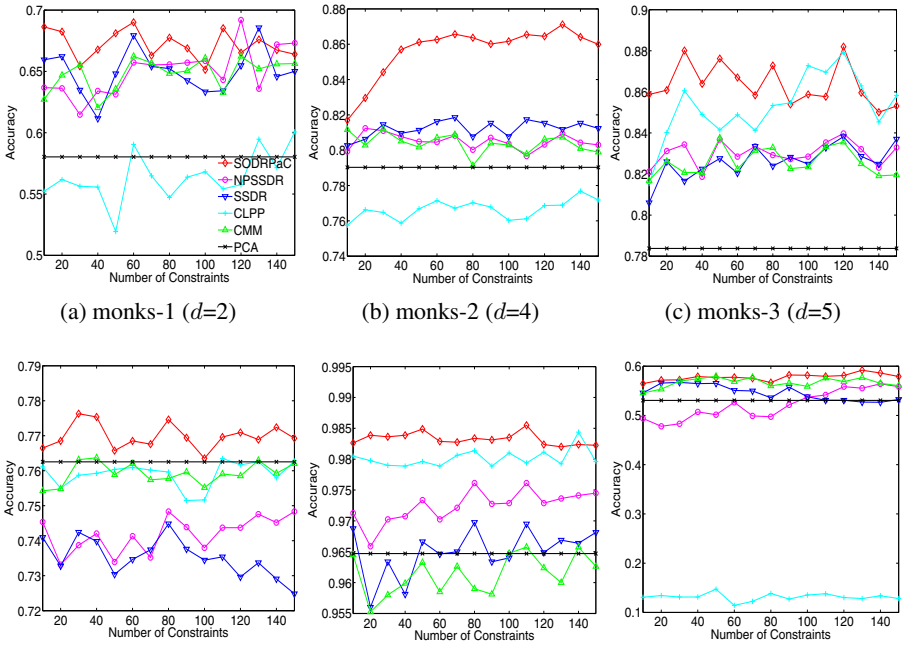
Table 1. Summary of the benchmark data sets

Data set	Dimension	Instance	Class	Data set	Dimension	Instance	Class
monks-1	6	556	2	monks-2	6	601	2
monks-3	6	554	2	letter(ABCD)	16	800	4
heart	13	270	4	GCM	16063	198	14

As shown in Eq. 10, λ is the parameter that controls the balance between $\mathbf{P} - \mathbf{Q}$ and $\mathbf{M}$. In this experiments setting, the parameter λ is searched from 2^{α}, where $\alpha \in \{\alpha | -5 \leq \alpha \leq 10, \alpha \in \mathbb{Z}\}$. A weighted 5-nearest-neighbor graph is employed to construct the manifold regularizer. In addition, the kernel parameter γ follows the suggestion in [17] that it is searched from the grid $\{\frac{\delta^2}{16}, \frac{\delta^2}{8}, \frac{\delta^2}{4}, \frac{\delta^2}{2}, \delta^2, 2\delta^2, 4\delta^2, 8\delta^2, 16\delta^2\}$, where δ is the mean norm of data. The parameter λ and the manifold regularizer are then optimized by means of the 5-fold cross-validation. As to the parameter settings of other competitive methods, we follow the parameters recommended by them, which are considered to be optimal. Without specific explanation, the number of must-link constraints is always set to be equal to that of cannot-link constraints, as the equal equilibrium between must-link constraints and cannot-link constraints is favorable for the existing methods. In addition, the value of k for searching shared nearest neighbors is set to be 3. The reason of this setting is to guarantee that the pairs of points in shared nearest neighbors reside in the same subclass, and to make the constraint transformation have more opportunities to be performed. In our experiments, must-link constraints and cannot-link constraints are selected according to the ground-truth of data labels.

4.2 Analysis of Experiments

First, the effectiveness of SODRPaC is exhibited by changing the number of constraints. Apart from the semi-supervised dimension reduction methods, we also take PCA as the baseline. As illustrated in Fig. 3, SODRPaC always keeps the best performance when the number of available constraints increases from 10 to 150. As seen in Fig. 3a, Fig. 3b, Fig. 3d, and Fig.3f, CLPP is inferior to PCA even if the number of constraints is small. The side effect of *inter-subclass must-link constraints*, in this case, can be neglected.



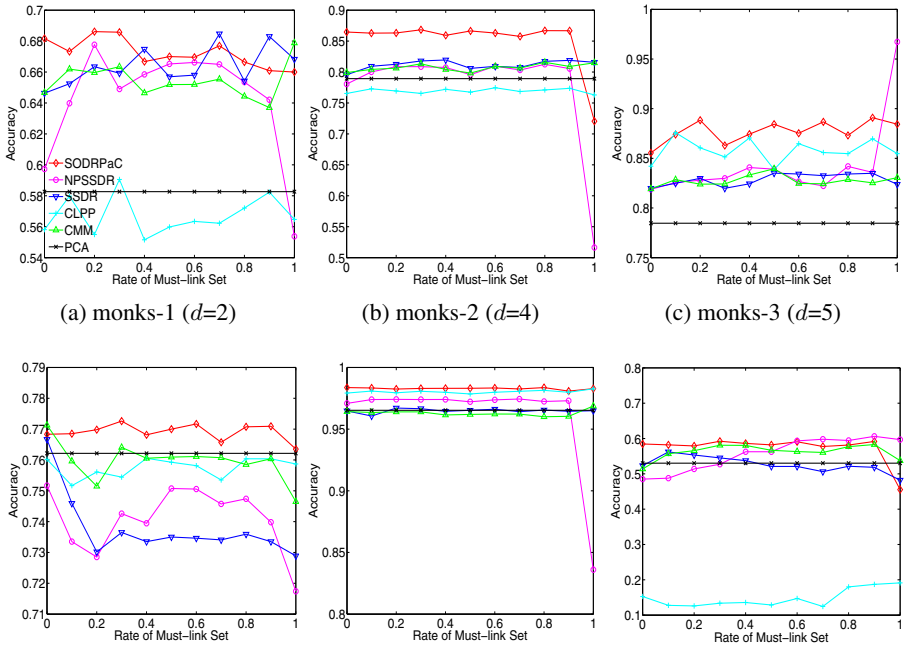
(a) monks-1 (d=2) (b) monks-2 (d=4) (c) monks-3 (d=5)

(d) heart (d=10) (e) letter(abcd) (d=10) (f) GCM (d=150)

Fig. 3. The performance with different numbers of constraints (d: reduced dimensionality)

The reason is probably that the feature of discovering the local structure of data points could not help CLPP to outperform PCA. However, our SODRPaC, which also utilizes the manifold regularization due to its property of discovering the local structure, obtains the best performance. We can judge that the new discriminant criterion boosts the performance. It is also presented in Fig. 3d that the performance of SSDR decreases to some extent with the increase of the number of constraints. The possible reason is that increasing the number of available constraints makes the opportunity higher that *inter-subclass must-link constraints* exist, which deteriorates the optimization on the fine dimension reduction. It should be also pointed out that SODRPaC does not significantly outperform other methods. A possible reason is that the Euclidean distance, which is employed to formulate the similarity between points in the original space, is likely to be meaningless in the high dimensional space.

We then examine the relative impact between must-link constraints and cannot-link constraints on the performance of SODRPaC. In this experiment, given 150 available constraints, the ratio of must-link constraints to cannot-link constraints is varied from 0 to 1. Fig. 4 presents that SODRPaC has a much smoother behavior than others with the change of ratio. It indicates that SODRPaC is more robust than other semi-supervised methods in terms of the imbalance between must-link constraints and cannot-link constraints. As shown in Fig. 4b and Fig. 4f, SODRPaC presents an obvious degradation of



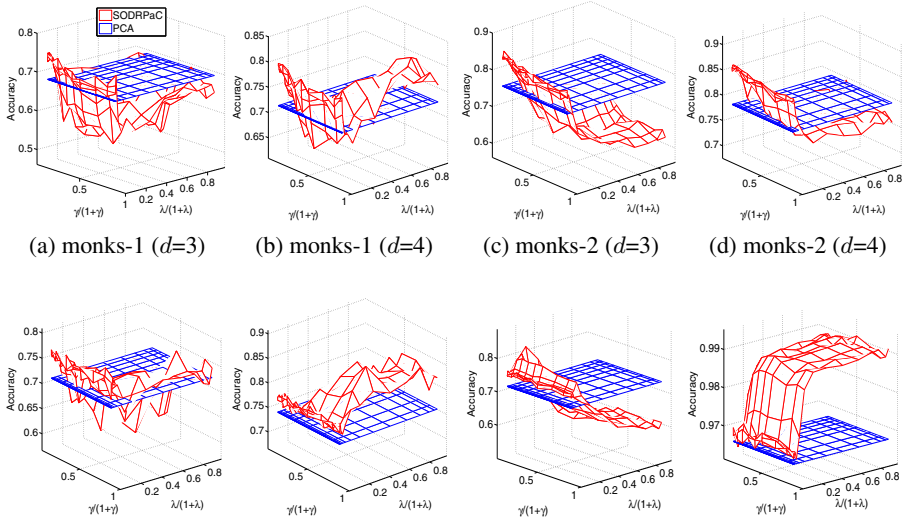
(a) monks-1 (d=2) (b) monks-2 (d=4) (c) monks-3 (d=5)

(d) heart (d=10) (e) letter (abcd) (d=10) (f) GCM (d=150)

Fig. 4. The performance with the change of rate for must-link set (d: reduced dimensionality)

performance when all constraints are must-link ones. The most probable reason would be that the transformation from must-link constraints into cannot-link constraints can not be performed when the necessary cannot-link constraints lack. This behavior is consistent with the conclusion demonstrated in [9] that cannot-link constraints are more important than must-link constraints in guiding the dimension reduction.

As implicated in the previous sections, the parameter λ that controls the balance between $\mathbf{P} - \mathbf{Q}$ and $\mathbf{M}$, and the factor γ that is related to computing the similarity between two points would influence the performance of SODRPaC. An analysis on the two parameters is necessary to provide the guideline about how to choose their values. PCA is employed as the baseline because existing methods can not hold such two parameters simultaneously. Because of the different scale between λ and γ, λ-axis and γ-axis are thus plotted as $\lambda/(1+\lambda)$ and $\gamma/(1+\gamma)$, respectively. The axis values are then in the interval $(0, 1)$. We empirically uncover two interesting patterns for most of data sets and reduced dimensions as well. There are two regions where SODRPaC are more likely to obtain its best performance. The first region is where $\lambda/(1+\lambda)$ is small, as shown Fig. 5a, Fig. 5b, Fig. 5c, Fig. 5d, Fig. 5e and Fig. 5g. In this situation, the variation of $\gamma/(1+\gamma)$ would not cause the dramatic change for the performance of SODRPaC. The second region is where both $\lambda/(1+\lambda)$ and $\gamma/(1+\gamma)$ are large, as shown in Fig. 5b, Fig. 5e, Fig. 5f, and Fig. 5h.



(a) monks-1 (d=3) (b) monks-1 (d=4) (c) monks-2 (d=3) (d) monks-2 (d=4)

(e) monks-3 (d=3) (f) monks-3 (d=4) (g) letter(abcd) (d=3) (h) letter(abcd) (d=15)

Fig. 5. The analysis for λ and γ (d: reduced dimensionality)

5 Conclusions and Future Works

In this paper, we have proposed a new linear dimension reduction method with must-link constraints and cannot-link constraints, called SODRPaC, that can deal with the multiple subclasses data. Inspired by the observation that handling the *inter-subclass must-link constraint* is challenging for the existing methods, a new discriminant criterion is invented by primarily transforming must-link constraints into cannot-link constraints. We also combine the manifold regularization into our dimension reduction framework. The results of extensive experiments show the effectiveness of our method.

There are some other aspects of this work that merit further research. Although the empirical choice of λ and γ is suggested, we do not as yet have a good understanding of how to choose these two parameters which are also correlated with choice of the number of the reduced dimensionality. Therefore, we are interested in automatically identifying these three parameters and uncovering relationships among them. Another possible would be to integrate the semi-supervised dimension reduction and clustering in a joint framework with automatic subspace selection.

Acknowledgments. A part of this research is supported by the Strategic International Cooperative Program funded by Japan Science and Technology Agency (JST) and by the grant-in-aid for scientific research on fundamental research (B) 21300053 from the Japanese Ministry of Education, Culture, Sports, Science and Technology. Bin Tong is sponsored by the China Scholarship Council (CSC).

References

1. Parsons, L., Haque, E., Liu, H.: Subspace Clustering for High Dimensional Data: A Review. In: SIGKDD Explorations, pp. 90–105 (2004)
2. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press, San Diego (1990)
3. Joliffe, I.: Principal Component Analysis. Springer, New York (1986)
4. Zhu, X.: Semi-supervised Learning Literature Survey. Technical Report Computer Sciences 1530, University of Wisconsin-Madison (2007)
5. Zhang, D., Zhou, Z.H., Chen, S.: Semi-supervised Dimensionality Reduction. In: Proceedings of the 7th SIAM International Conference on Data Mining (2007)
6. Wang, F., Chen, S., Li, T., Zhang, C.: Semi-supervised Metric Learning by Maximizing Constraint Margin. In: ACM 17th Conference on Information and Knowledge Management, pp. 1457–1458 (2008)
7. Cevikalp, H., Verbeek, J., Jurie, F., Klaser, A.: Semi-supervised Dimensionality Reduction Using Pairwise Equivalence Constraints. In: International Conference on Computer Vision Theory and Applications (VISAPP), pp. 489–496 (2008)
8. Wei, J., Peng, H.: Neighborhood Preserving Based Semi-supervised Dimensionality Reduction. Electronics Letters 44, 1190–1191 (2008)
9. Tang, W., Xiong, H., Zhong, S., Wu, J.: Enhancing Semi-supervised Clustering: A Feature Projection Perspective. In: Proc. of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 707–716 (2007)
10. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B.: Learning with Local and Global Consistency. In: Proc. Advances in Neural Information Processing Systems, pp. 321–328 (2004)
11. Belkin, M., Niyogi, P.: Manifold Regularization: A Geometric Framework for Learning From Labeled and Unlabeled Examples. Journal of Machine Learning Research 7, 2399–2434 (2006)
12. Sugiyama, M.: Dimensionality Reduction of Multimodal Labeled Data by Local Fisher Discriminant Analysis. Journal of Machine Learning Research 8, 1027–1061 (2007)
13. Ertöz, L., Steinbach, M., Kumar, V.: A New Shared Nearest Neighbor Clustering Algorithm and its Applications. In: Proc. of the Workshop on Clustering High Dimensional Data and its Applications, Second SIAM International Conference on Data Mining (2002)
14. Yang, J., Zhang, D., Yang, J.Y., Niu, B.: Globally Maximizing, Locally Minimizing: Unsupervised Discriminant Projection with Applications to Face and Palm Biometrics. IEEE Trans. Pattern Analysis and Machine Intelligence 29(4), 650–664 (2007)
15. Blake, C., Keogh, E., Merz, C.J.: UCI Repository of Machine Learning Databases, `http://www.ics.uci.edu/~mlearn/MLRRepository.html`
16. Ramaswamy, S., Tamayo, P., Rifkin, R., et al.: Multiclass Cancer Diagnosis Using Tumor Gene Expression Signatures. Proceedings of the National Academy of Sciences, 15149–15154 (1998)
17. He, X., Yan, S., Hu, Y., Niyogi, P.: Face Recognition Using Laplacianfaces. IEEE Transaction on Pattern Analysis and Machine Intelligence 27, 316–327 (2005)