



Semi-supervised discriminant Isomap with application to visualization, image retrieval and classification

Rui Huang¹ · Guopeng Zhang¹ · Junli Chen¹

Received: 4 April 2017 / Accepted: 17 March 2018 / Published online: 27 March 2018
© Springer-Verlag GmbH Germany, part of Springer Nature 2018

Abstract

As one of the most promising nonlinear unsupervised dimensionality reduction (DR) technique, the Isomap reveals the intrinsic geometric structure of manifold by preserving geodesic distance of all data pairs. Recently, some supervised versions of Isomap have been presented to guide the manifold learning and increase the discriminating capability. However, the performance may deteriorate when there is no sufficient prior information available. Hence, a novel semi-supervised discriminant Isomap (SSD-Isomap) is proposed in the paper. First, two pairwise constraints including must-link and likely-link (LL) are used to depict the neighborhoods of data points. Then, two graphs are constructed based on the two constraints, and distances between points belonging to the LL constraint are modified by a scale parameter. Finally, the geodesic distance metric is obtained based on the graphs, and the corresponding optimal nonlinear subspace is sought. The performance of SSD-Isomap is evaluated by extensive experiments of data visualization, image retrieval and classification. Compared with other state-of-the-art DR methods, SSD-Isomap presents more accurate and robust results.

Keywords Isomap · Semi-supervised dimensionality reduction · Visualization · Image retrieval · Classification

1 Introduction

Dimensionality reduction (DR), as a fundamental issue in pattern recognition, is to project the original high-dimensional data into a lower-dimensional space by getting rid of the redundant or even irrelevant information according to a certain criterion. In recent years, since many emerging applications are revolved with the high-dimensional data, such as gene expressions, text mining, image classification and retrieval, the technique has attracted considerable attention and extensive studies have been done.

So far, many DR methods have been proposed. Among them, principal component analysis (PCA) [1] and linear discriminant analysis (LDA) [2, 3] are two representative methods. Recently, some methods based on manifold learning have shown their advantages, e.g. multidimensional scaling (MDS) [4], laplacian eigenmaps (LE) [5, 6], locally linear embedding (LLE) [7], t-distributed stochastic neighbor embedding (t-SNE) [8] and Isomap [9, 10]. In particular,

Isomap, which can reveal the intrinsic geometric structure of manifold by preserving geodesic distance of all similarity pairs, has presented some encouraging results. However, the original Isomap is not good at extracting discriminative features for classification, as no class information of labeled data is considered. To handle the problem, several supervised versions of Isomap are proposed. WeightedIso uses a constant factor to change the Euclidean distance between two data points with the same class labels in the first step of Isomap [11]. In supervised Isomap (S-Isomap), two parameters are applied to update the distances among the pairwise points with the same and different class labels [12]. Zhang etc. [13] proposed a pairwise-constrained marginal Isomap (M-Isomap) which incorporates the pairwise cannot-link (CL) and must-link (ML) constraints induced from the neighborhood graph into Isomap to guide the discriminant manifold learning. Inspired by M-Isomap, multi-manifold discriminant Isomap (MMD-Isomap) [14] was presented by introducing two global pairwise constraints and defining a joint optimization objective.

Despite the supervised DR methods can generally perform better than the unsupervised ones, the performances are deeply influenced by the number of labeled samples, and performance deterioration becomes inevitable when

✉ Rui Huang
huangr@shu.edu.cn

¹ School of Communication and Information Engineering,
Shanghai University, 99 Shangda Road, Shanghai, China

there are not enough labeled data available [15, 16]. Active learning and semi-supervised learning are two promising learning paradigms to address the problem. Different from active learning which tries to select the most informative samples to be labeled [17], semi-supervised learning makes use of the prior knowledge of labeled samples and discriminative information hiding in unlabeled samples [18–21]. Semi-supervised discriminant analysis (SDA) [22] improves LDA by adding a regularization term to preserve the local structures of data. In [23], semi-supervised Isomap (SS-Isomap) is proposed by using prior information on exact mapping of certain data points to compute the low dimensional coordinates of unknown points. It is not in the typical sense of semi-supervised learning where both the labeled and unlabeled data are used for classification. It actually presents a method for out-of-sample mapping. Besides, the authors indicated that the improvement of SS-Isomap over the basic Isomap is not significant according to experimental results. In [24], multiple view semi-supervised dimensionality reduction (MVSSDR), an improved version of semi-supervised dimensionality reduction (SSDR) [25], uses the pairwise constraints to derive embedding in each view and makes these embeddings comparable through a linear transformation.

Although some semi-supervised DR methods have been proposed, most of them try to compute linear projections. However, non-linear DR may play an important role in human perception and learning. As a popular non-linear DR method, original Isomap is unsupervised and its performance can be improved by considering the class label information when a sufficient number of labeled samples available. In the case of limited training data, semi-supervised learning is helpful. In the paper, we study to apply useful information from the labeled and unlabeled samples to manifold learning for Isomap, and a semi-supervised discriminant Isomap (SSD-Isomap) is presented. In the method, two pairwise constraints including must-link (ML) and likely-link (LL) are first defined. Among the constraints, ML is constructed based on the labeled samples, while LL is built for the unlabeled samples. Then, two graphs based on the constraints are obtained, and the distances between points belonging to LL constraint are reset by a scale parameter. Finally, the corresponding optimal nonlinear subspace is sought to preserve the real distance of data.

The main contributions of this paper are summarized as follows:

1. Besides the common used constraint of must-link (ML), the constraint of likely-link (LL) is presented to depict the neighborhoods of data points without class labels. Compared with those supervised versions of Isomap, such as S-Isomap, M-Isomap and MMD-Isomap, the local structure information in the unlabeled sample
2. SSD-Isomap uses a similar procedure in Isomap to obtain the low dimensional embedding after the geodesic distance matrix is initialized. Unlike M-Isomap and MMD-Isomap, no extra optimization algorithms are needed. Extensive experiments on data visualization, image retrieval and classification show that our method has better performance compared with other state-of-the-art DR methods.

The rest of this paper is organized as follows. In Sect. 2, the related work is briefly introduced. Section 3 describes the details of the proposed SSD-Isomap. In Sect. 4, extensive experiments are carried out for performance evaluation of the proposed method. Finally, Sect. 5 gives the conclusion.

2 Related work

In the section, we first summarize some notations used throughout the paper. The closely related works about Isomap is then reviewed.

Given a data set $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N] \in R^{M \times N}$ with N points, DR is to find a mapping function that maps these points to a new data set $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N] \in R^{m \times N}$ in a lower dimensional space with dimension m ($m \ll M$). $d(\mathbf{x}_i, \mathbf{x}_j)$ denotes the Euclidean distance between \mathbf{x}_i and \mathbf{x}_j .

2.1 Isomap

Isomap is a classic global nonlinear DR algorithm which aims at seeking an optimal subspace that best preserves the geodesic distance in pair data. It can be summarized as follows:

1. Construct a weighted undirected neighborhood graph $G(V, E)$, where node $v_i \in V$ corresponds to point \mathbf{x}_i . For every pair of data points, if $d(\mathbf{x}_i, \mathbf{x}_j)$ is smaller than the fixed radius ϵ or $\mathbf{x}_j \in \text{KNN}(\mathbf{x}_i)$ (KNN means \mathbf{x}_j is the K -nearest neighbors of \mathbf{x}_i), the weight of edge $e(\mathbf{x}_i, \mathbf{x}_j) \in E$ is set to $d(\mathbf{x}_i, \mathbf{x}_j)$.
2. Compute geodesic distances. Initialize the distance $d_G(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_i, \mathbf{x}_j)$ if \mathbf{x}_j and \mathbf{x}_i are neighbors, otherwise, let $d_G(\mathbf{x}_i, \mathbf{x}_j) = \infty$. Estimate geodesic distances between all pairs of data points through computing all the shortest path distances $d_G(\mathbf{x}_i, \mathbf{x}_j)$ in G . Dijkstra's or Floyd's algorithm can be applied to find the shortest paths.
3. Construct m -dimensional embedding. Define the Isomap criterion as follows:

$$\min_Y \sum_{\mathbf{x}_i, \mathbf{x}_j} (d(\mathbf{y}_i, \mathbf{y}_j) - d_G(\mathbf{x}_i, \mathbf{x}_j))^2 \tag{1}$$

Let $\mathbf{H} = \mathbf{I} - (1/N)\mathbf{e}\mathbf{e}^T$, where \mathbf{I} is a $N \times N$ identity matrix, and \mathbf{e} is the vector of all ones. Let \mathbf{Q} be a $N \times N$ matrix with elements $\mathbf{Q}_{ij} = d_G^2(\mathbf{x}_i, \mathbf{x}_j)$. According to MDS, the lower-dimensional embedding \mathbf{Y} is obtained as $[\sqrt{\lambda_1}\mathbf{v}_1, \sqrt{\lambda_2}\mathbf{v}_2, \dots, \sqrt{\lambda_m}\mathbf{v}_m]^T$, where $\{\mathbf{v}_i\}_{i=1}^m$ denotes the eigenvector according to the first m leading eigenvalues of $\mathbf{R} = -\mathbf{H}\mathbf{Q}\mathbf{H}/2$.

2.2 Supervised versions of Isomap

Since Isomap is an unsupervised DR algorithm, it is not good at extracting discriminative features for classification task. Some supervised versions have been presented by considering the class label information [26]. In WeightedIso, the Euclidean distances between points with the same labels are reduced by a constant rescaling factor, and then Isomap is implemented based on these new distances. S-Isomap [12] develops this idea and defines a different distance metric as follows:

$$\hat{d}(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} \sqrt{1 - \exp(-d^2(\mathbf{x}_i, \mathbf{x}_j)/\beta)} & \text{if } l(\mathbf{x}_i) = l(\mathbf{x}_j) \\ \sqrt{\exp(d^2(\mathbf{x}_i, \mathbf{x}_j)/\beta) - \alpha} & \text{if } l(\mathbf{x}_i) \neq l(\mathbf{x}_j) \end{cases}, \tag{2}$$

where $l(\mathbf{x}_i)$ is the class label of \mathbf{x}_i . The parameters of α and β are used to control the range of $d(\mathbf{x}_i, \mathbf{x}_j)$. Usually, α is set to be a small positive value, and β is set to be the average Euclidean distance between all pairs of data point. The two parameters both modify the distances between data points with help of the class labels, and can improve the classification performance.

In M-Isomap [13], two local pairwise constraint sets, i.e., must-link (ML) and cannot-link (CL) are defined as

$$S_{ML} = \{(\mathbf{x}_i, \mathbf{x}_j) \mid l(\mathbf{x}_i) = l(\mathbf{x}_j), \mathbf{x}_j \in \text{KNN}(\mathbf{x}_i)\}, \tag{3}$$

$$S_{CL} = \{(\mathbf{x}_i, \mathbf{x}_j) \mid l(\mathbf{x}_i) \neq l(\mathbf{x}_j), \mathbf{x}_j \in \text{KNN}(\mathbf{x}_i)\}. \tag{4}$$

According to Eqs. (3) and (4), two pairwise-constrained neighborhood graphs G_{ML} and G_{CL} can be constructed. M-Isomap optimizes the following two criteria:

$$J_{ML} = \min_Y \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in S_{ML}} (d(\mathbf{y}_i, \mathbf{y}_j) - d_G^{ML}(\mathbf{x}_i, \mathbf{x}_j))^2, \tag{5}$$

$$J_{CL} = \max_Y \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in S_{CL}} (d(\mathbf{y}_i, \mathbf{y}_j) - d_G^{CL}(\mathbf{x}_i, \mathbf{x}_j))^2, \tag{6}$$

where $d_G^{ML}(\mathbf{x}_i, \mathbf{x}_j)$ and $d_G^{CL}(\mathbf{x}_i, \mathbf{x}_j)$ are the shortest path distances between \mathbf{x}_i and \mathbf{x}_j in G_{ML} and G_{CL} , respectively. By combing Eqs. (5) and (6), the optimization problem can be solved by using iterative trace ratio (ITR) algorithm. Inspired by M-Isomap, MMD-Isomap [14] is proposed by introducing two global pairwise constraints sets. The two optimization criteria are combined through a regularization parameter, and SMACOF is used to solve the objective.

3 Semi-supervised discriminant Isomap

The supervised versions of Isomap can usually perform better than the original one, but the performance improvement largely depends on whether there are enough number of labeled samples. According to the idea of semi-supervised learning, some discriminative information hiding in the unlabeled samples can be complementary to the prior knowledge. In view of this, a semi-supervised discriminant Isomap (SSD-Isomap) is presented. A ‘‘good’’ projection should be the one which two data points in the new subspace are close to each other if and only if they have the same labels or they are in neighborhood in the original feature space.

3.1 Pairwise-constrained graphs

Let $\mathbf{X}_L = \{\mathbf{x}_1, \dots, \mathbf{x}_L\} \in \mathbf{R}^{M \times L}$ and $\mathbf{X}_U = \{\mathbf{x}_{L+1}, \dots, \mathbf{x}_{L+U}\} \in \mathbf{R}^{M \times U}$ be the labeled and unlabeled data sets. A weighted undirected neighborhood graph $G(V, E)$ is first constructed. The edge weight $e(\mathbf{x}_i, \mathbf{x}_j) \in \{0, 0.5, 1\}$ is used to indicate three types (disconnection, strong connection and weak connection) of the neighboring points. Specifically, when the labeled samples \mathbf{x}_i and \mathbf{x}_j have the same class labels, the connection between the two points is strong with $e(\mathbf{x}_i, \mathbf{x}_j) = 1$; when $\mathbf{x}_i \in \mathbf{X}_U$ is among K -nearest neighbors of \mathbf{x}_j or $\mathbf{x}_j \in \mathbf{X}_U$ is among K -nearest neighbors of \mathbf{x}_i , the connection between the two points is weak with $e(\mathbf{x}_i, \mathbf{x}_j) = 0.5$; otherwise, there is no connection with $e(\mathbf{x}_i, \mathbf{x}_j) = 0$. Therefore, a global pairwise-constrained set (namely ML) for the labeled samples and a local pairwise-constrained set named by likely-link (LL) for the unlabeled samples are defined as follows:

$$S_{ML}^L = \{(\mathbf{x}_i, \mathbf{x}_j) \mid e(\mathbf{x}_i, \mathbf{x}_j) = 1, l(\mathbf{x}_i) = l(\mathbf{x}_j)\}, \tag{7}$$

$$S_{LL}^U = \{(\mathbf{x}_i, \mathbf{x}_j) \mid e(\mathbf{x}_i, \mathbf{x}_j) = 0.5, \mathbf{x}_j \in \text{KNN}(\mathbf{x}_i)\}, \tag{8}$$

where $l(\mathbf{x}_i)$ is the class label of \mathbf{x}_i ($i = 1, 2, \dots, L$). Based on these definitions, an ML-constrained graph $G^L(V, E)$ based on the labeled data and an LL-constrained graph $G^U(V, E)$ based on the unlabeled data are constructed by keeping the edges with $e(\mathbf{x}_i, \mathbf{x}_j) = 1$ and $e(\mathbf{x}_i, \mathbf{x}_j) = 0.5$, respectively.

3.2 Geodesic distance metrics

Based on the constrained graph $G^L(V, E)$, $d^{ML}(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2$ is computed for the linking pair $(\mathbf{x}_i, \mathbf{x}_j) \in S_{ML}^L$. For a pair $(\mathbf{x}_i, \mathbf{x}_j) \in S_{LL}^U$, they are likely to be connected, but there is a possibility that they come from different classes and thus no edge should be put between them. In view of this, a compromise factor γ is applied to reset their distance according to the graph $G^U(V, E)$:

$$d^{LL}(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2 / \gamma, \quad (9)$$

where $\gamma \in (0, 1)$. Through computing d^{ML} and d^{LL} for all pairs $(\mathbf{x}_i, \mathbf{x}_j)$ ($i, j = 1, 2, \dots, L + U$), the distance matrix D_{SS} is constructed. The geodesic distances between all pairs of points are estimated by computing their shortest path distances $d_{SSD}(\mathbf{x}_i, \mathbf{x}_j)$. Similar to Isomap, the shortest path between each pair of points is computed by Floyd's algorithm. In particular, initialize $d_{SSD}(\mathbf{x}_i, \mathbf{x}_j) = d_{SS}(\mathbf{x}_i, \mathbf{x}_j)$ for points \mathbf{x}_i and \mathbf{x}_j . Then for each value of $k = 1, 2, \dots, L + U$ in turn, replace all entries $d_{SSD}(\mathbf{x}_i, \mathbf{x}_j)$ by $\min\{d_{SSD}(\mathbf{x}_i, \mathbf{x}_j), d_{SSD}(\mathbf{x}_i, \mathbf{x}_k) + d_{SSD}(\mathbf{x}_k, \mathbf{x}_j)\}$. Finally, the matrix D_{SSD} will contain the shortest path distances between all pairs of points.

3.3 Objective function

Like Isomap and S-Isomap, SSD-Isomap seeks the projection which can preserve the original manifold structure between all pair of points in a lower dimensional space. The object function is defined as:

$$J_{SSD} = \min_{\mathbf{Y}} \sum_{\substack{(\mathbf{x}_i, \mathbf{x}_j) \in S_{ML}^L \\ \text{or } (\mathbf{x}_i, \mathbf{x}_j) \in S_{LL}^U}} (d(\mathbf{y}_i, \mathbf{y}_j) - d_{SSD}(\mathbf{x}_i, \mathbf{x}_j))^2, \quad (10)$$

where $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_N] \in R^{m \times (L+U)}$ is the lower dimensional representations of \mathbf{X} . Let $\mathbf{R}^{SSD} = -\mathbf{H}\mathbf{Q}^{SSD}\mathbf{H}/2$ and

$\mathbf{Q}_{ij}^{SSD} = (d_G^{SSD}(\mathbf{x}_i, \mathbf{x}_j))^2$; then Eq. (10) can be rewritten as follows:

$$J_{SSD}(\mathbf{Y}) = \min_{\mathbf{Y}} \|\mathbf{Y}^T \mathbf{Y} - \mathbf{R}^{SSD}\|_F, \quad (11)$$

where $\|\cdot\|_F$ is the Frobenius matrix norm. The final embedding \mathbf{Y} can be obtained by the classic MDS. The whole training procedure of SSD-Isomap is listed in Table 1.

As one of Isomap series methods, SSD-Isomap uses the constraints of ML and LL to guide the computation of geodesic distance metrics. The computational complexity is $O(N^3)$ when Floyd's algorithm is used, and it can be improved to $O(kN^2 \log N)$ when Dijkstra's algorithm is used, where N and k are the sample size and neighborhood size, respectively. Based on the shortest paths between all pairs of samples, eigen-decomposition is applied to obtain the lower-dimensional embedding. The time complexity is $O(N^3)$. In fact, the computational complexity is always a bottleneck for all the Isomap series methods when they are applied to large data sets. But landmark Isomap [10] presents an effective solution to the problem, and can reduce the computational cost to $O(knN \log N)$ for the shortest-paths calculation and $O(n^2N)$ for the MDS eigenvalue calculation, where n is the number of landmark points and $n \ll N$. So, SSD-Isomap can handle large-scale datasets by adopting the scheme.

4 Experiments

In this section, the extensive experiments are carried out to evaluate the performance of the proposed SSD-Isomap. The visual, classification and image retrieval performances are compared with those of some state-of-the-art DR algorithms including the unsupervised Isomap, MDS, Laplacian Eigenmaps (LE), LLE, the supervised LDA, S-Isomap, M-Isomap, and the semi-supervised SDA, SSDR. In this study, we test a synthetic data set, six benchmark image data sets including Corel [27–29], UC Merced LULC [30], Caltech101 [31], YALE [32, 33], ORL [34], MNIST [35], and six UCI

Table 1 SSD-Isomap

Input: labeled data set $\{\mathbf{x}_i, l(\mathbf{x}_i)\}_{i=1}^L$, unlabeled data set $\{\mathbf{x}_j\}_{j=1, \gamma}^U, K$

Output: \mathbf{Y}

1. Construct $G(V, E)$. Define a $(L + U) \times (L + U)$ graph G over all data points by setting the edge $e(\mathbf{x}_i, \mathbf{x}_j)$ between points \mathbf{x}_i and \mathbf{x}_j to 1 if they come from the same class, or to 0.5 if \mathbf{x}_i is one of the K nearest neighbors of \mathbf{x}_j and at least one of the two points belongs to $\{\mathbf{x}_j\}_{j=1}^U$.
2. Compute the distance between points \mathbf{x}_i and \mathbf{x}_j to $d_{SS}(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_i, \mathbf{x}_j)$ if $e(\mathbf{x}_i, \mathbf{x}_j) = 1$, or $d_{SS}(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_i, \mathbf{x}_j)/\gamma$ if $e(\mathbf{x}_i, \mathbf{x}_j) = 0.5$.
3. Compute shortest paths. Initialize $d_{SSD}(\mathbf{x}_i, \mathbf{x}_j) = d_{SS}(\mathbf{x}_i, \mathbf{x}_j)$, if $e(\mathbf{x}_i, \mathbf{x}_j) = 1$ or 0.5; $d_{SSD}(\mathbf{x}_i, \mathbf{x}_j) = \infty$ otherwise. Then, Floyd's or Dijkstra's algorithm is used to calculate the shortest paths between all pairs of points.
4. Construct m -dimensional embedding \mathbf{Y} . MDS is used to obtain the embedding by optimizing Eq. (11).

data sets [36]. All of the experiments were run in MATLAB with Intel(R) core(TM) i5-3470 CPU at 3.2 GHz and 12 GB RAM.

4.1 Visualization

An artificial data set of Swiss roll is used in the section. This 3-D data set with 1600 samples has 4 classes and is shown in Fig. 1. In the experiment, half of data is randomly selected as the training (labeled) samples, and the rest is treated as test (unlabeled) samples. In Figs. 1, 4-class samples are marked by different colors. Furthermore, labeled samples are denoted by the symbol of ‘circle’, and unlabeled samples from four classes are denoted by the symbol of ‘stars’, ‘plus’, ‘square’ and ‘diamond’, respectively. For unsupervised DR methods, all 1600 samples are used to obtain the lower dimensional representations. For supervised methods, the lower embedding is computed only based on the training samples, and the map approximated by a BP neural network is applied to all samples. For semi-supervised methods, the map is established based on both labeled and unlabeled samples.

To get the best performance of algorithms, their parameters are carefully adjusted. The number of neighbors K is set to 30 for S-Isomap, 7 for Isomap and LLE, 150 for the proposed SSD-Isomap. We follow the same settings as [12] for S-Isomap: the parameter α is set to 0.5, and the parameter β is set to be the average Euclidean distance between all pairs of data points. In our algorithm, the parameter of γ is determined by fivefold cross validation and is set to 0.1. The visualization results are shown in Fig. 2.

It can be observed that the unsupervised methods including Isomap, MDS, LE and LLE cannot separate the four-class data clearly. T-SNE can achieve a clear separation, but the clusters in green and blue are wrongly divided into two subparts. In supervised and semi-supervised methods, LDA, M-Isomap, SDA, SSDR and MVSSDR also fail to achieve separation. The clusters in the S-Isomap embedding space

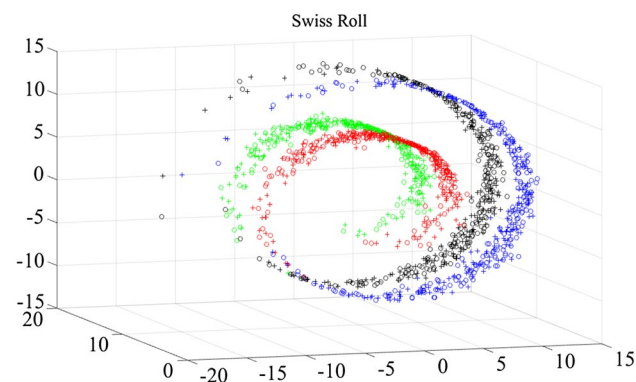


Fig. 1 Data set of Swiss roll

are generally separable, but there still exist overlaps between clusters in red and blue, and clusters in green and black. Compared with other methods, the proposed SSD-Isomap provides a better separation on the clusters.

4.2 Image retrieval

Three data sets including Caltech101, Corel1000 and UC Merced LULC are used in the section. Caltech101 data set, which has been widely adopted for object recognition and image retrieval tasks, has 101 categories with 40–800 images per category. The size of each image is roughly 300×200 pixels. Corel1000 data set is a part of the real-world photos from COREL Photo Gallery. It has 10 categories with 100 images per category. The UC Merced LULC data set, obtained from aerial imagery, consists of images from 20 classes, with a pixel resolution of 30 cm. Each class contains 100 images of size 256 by 256 pixels. We extract basic color features and wavelet texture features to describe images [37, 38]. The features include color histogram (32 dimensions), color moment (64 dimensions), color auto correlogram (6 dimensions), wavelet moment (40 dimensions) and Gabor transform where the number of scales was set 4 and orientation was set 6 (48 dimensions). All the features are concatenated into a long vector as an image feature and each image is represented by a 190-dimensional vector. The new feature is normalized to zero mean and unit variance.

In our experiment on Caltech101, we use 10 out of 101 categories, and images from each class are randomly split into a training set of 80 images and a test set of 20 images. For Corel1000 and UC Merced LULC data sets, image samples are evenly divided into training set and test set. In training set, samples are further split into labeled and unlabeled subsets equally, and only samples belonging to the labeled subset can be used for the supervised DR methods. For other methods, all samples in the training set are used to obtain the lower dimensional projection. For nonlinear DR methods, such as Isomap, S-Isomap, M-Isomap and SSD-Isomap, a BP neural network is constructed to simulate the mapping from high dimension to lower dimension based on the training set. In the reduced dimension space, L_2 distance metric is employed to measure the similarities between the query image from the test set and the labeled images. Precision is used as the quantitative index for performance evaluation.

The dimension of data is reduced to the number of classes of each data set for all DR methods except for LDA. The parameters of each method are carefully adjusted to get the best performance. Table 2 lists the comparison on precision of top 5 to 25 (with 5 step intervals) retrieved images, and average precisions are also presented. On each data set, the best performance is emphasized by bold. As we can see, SSD-Isomap obtains the highest precisions on the three data sets. According to the average performance, the 12 methods

Fig. 2 The 2D embedding obtained by different methods on the Swiss roll data set

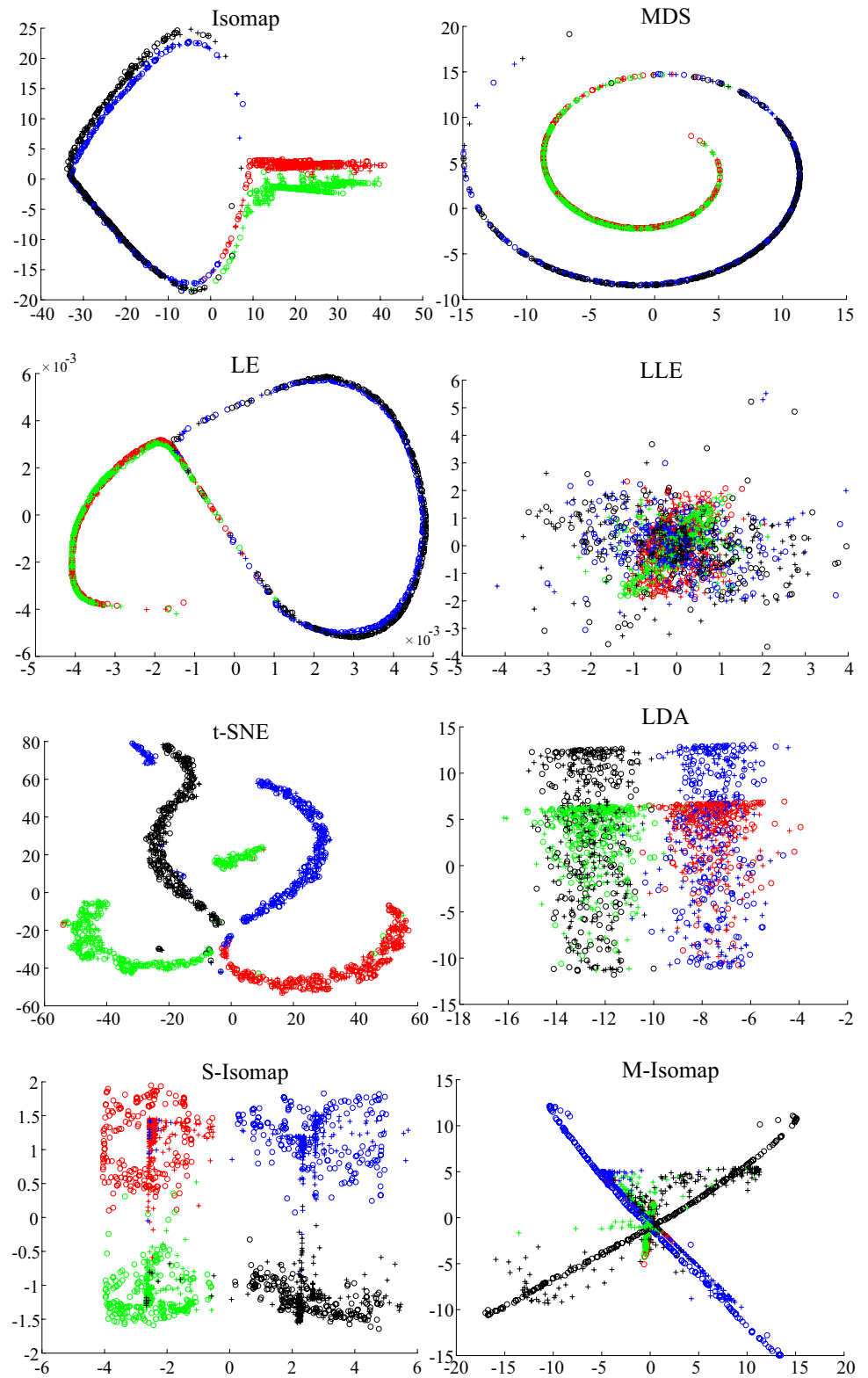
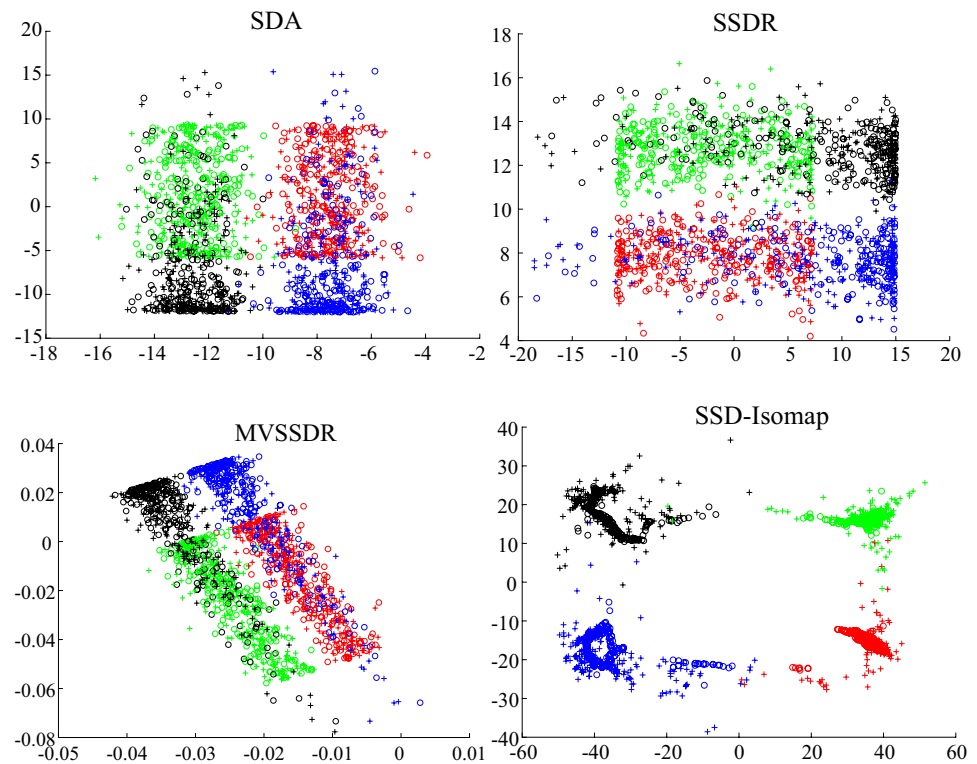


Fig. 2 (continued)



can be sorted as: (1) SSD-Isomap, SDA, S-Isomap, SSDR, LDA, MVSSDR, LE, MDS, Isomap, t-SNE, M-Isomap and LLE for Corel1000; (2) SSD-Isomap, LDA, SDA, SSDR, S-Isomap, MVSSDR, t-SNE, LE, MDS, Isomap, LLE and M-Isomap for LULC; (3) SSD-Isomap, S-Isomap, LDA, SDA, SSDR, M-Isomap, MVSSDR, MDS, LE, Isomap, t-SNE and LLE for Caltech101.

To compare the effect of number of dimensions on retrieval performance, Fig. 3 presents the average precisions of different methods when the number of dimension changes from 5 to 40 on the Caltech101 data set. It is obvious that the supervised and semi-supervised methods perform better than unsupervised ones. SSD-Isomap generally achieves the highest precisions except when the dimension is reduced to 5 and 30. SDA slightly outperforms S-Isomap, and SSDR ranks between S-Isomap and M-Isomap. In the figure, many methods achieve best performance when number of dimensions is 10. As known, the high-dimensional data can be efficiently represented in a space of a much lower dimension without losing much information. The number of reduced dimensions is a key parameter for DR. If the dimension is too small, important features are projected onto the same dimension, and if the dimension is too large, the projections become noisy. In Fig. 3, the performances of different methods exhibit the phenomenon. When the number of

dimensions grows, the retrieval precision of each method reaches a maximum in a ten-dimensional space and then decreases. Hence the intrinsic dimension of Caltech101 data set may be 10 by empirical analysis. Indeed, how to estimate intrinsic dimension is still an open issue, and it is beyond the scope of this paper. But Fig. 3 also shows that intrinsic dimension is significant for DR methods.

4.3 Classification

In classification experiments, three image data sets including MNIST, YALE and ORL, and six UCI data sets are used for performance comparison. The MNIST data set has 70,000 hand written digit images with sizes of 28×28 pixels. Each image is denoted by a 784-dimensional vector. YALE face data set contains 165 images of 15 individuals. These face images are resized to 32×32 pixels with 256 Gy levels, and each image is presented by a 1024-dimensional vector. In ORL data set, 400 images with sizes of 32×32 pixels are from 40 persons, so each image is also denoted by a 1024-dimensional vector. Table 3 gives the details of the six UCI data sets.

The same manner as in Sect. 4.2 is adopted to generate the training and test sets for MNIST, YALE and ORL data sets. In UCI data set, the two sets are obtained by fivefold

Table 2 Retrieval precisions of different methods

	5	10	15	20	25	Ave
Corel1000						
Isomap	0.4424	0.3950	0.3593	0.3369	0.3130	0.3793
MDS	0.4896	0.4308	0.3932	0.3603	0.3335	0.4015
LE	0.4952	0.4450	0.4085	0.3739	0.3410	0.4127
LLE	0.3478	0.3060	0.2697	0.2490	0.2322	0.2809
t-SNE	0.4148	0.3914	0.3724	0.3508	0.3254	0.3710
LDA	0.5204	0.5164	0.5144	0.5114	0.4964	0.5118
S-Isomap	0.6352	0.6356	0.6185	0.6184	0.6054	0.6226
M-Isomap	0.4220	0.3390	0.2739	0.2285	0.1996	0.2936
SDA	0.6720	0.6558	0.6440	0.6290	0.5998	0.6401
SSDR	0.6412	0.6196	0.5963	0.5669	0.5229	0.5893
MVSSDR	0.5444	0.5134	0.4833	0.4616	0.4373	0.4880
SSD-Isomap	0.6890	0.6844	0.6808	0.6754	0.6524	0.6764
LULC						
Isomap	0.1854	0.1628	0.1478	0.1397	0.1282	0.1524
MDS	0.1870	0.1660	0.1503	0.1401	0.1327	0.1552
LE	0.2026	0.1764	0.1536	0.1404	0.1279	0.1602
LLE	0.1630	0.1397	0.1218	0.1105	0.1044	0.1279
t-SNE	0.2028	0.1859	0.1680	0.1567	0.1446	0.1726
LDA	0.3990	0.3805	0.3643	0.3456	0.3267	0.3632
S-Isomap	0.2584	0.2582	0.2581	0.2581	0.2528	0.2571
M-Isomap	0.1380	0.1268	0.1197	0.1141	0.1082	0.1214
SDA	0.4028	0.3775	0.3523	0.3313	0.3092	0.3547
SSDR	0.3210	0.3047	0.2883	0.2777	0.2661	0.2916
MVSSDR	0.3154	0.2763	0.2448	0.2213	0.2038	0.2523
SSD-Isomap	0.4622	0.4561	0.4523	0.4465	0.4354	0.4505
Caltech101						
Isomap	0.3500	0.3300	0.3147	0.2960	0.2902	0.3167
MDS	0.3980	0.3765	0.3530	0.3322	0.3224	0.3564
LE	0.3430	0.3265	0.3110	0.2995	0.2854	0.3131
LLE	0.2690	0.2405	0.2290	0.2208	0.2178	0.2354
t-SNE	0.3270	0.3105	0.2983	0.2888	0.2800	0.3009
LDA	0.6140	0.6010	0.6023	0.5940	0.5830	0.5989
S-Isomap	0.6150	0.6125	0.6117	0.6112	0.6110	0.6123
M-Isomap	0.5520	0.5390	0.5260	0.5028	0.4846	0.5209
SDA	0.6210	0.6075	0.5967	0.5785	0.5654	0.5938
SSDR	0.5720	0.5520	0.5395	0.5215	0.5106	0.5391
MVSSDR	0.4490	0.4255	0.4080	0.3985	0.3880	0.4120
SSD-Isomap	0.6660	0.6640	0.6610	0.6587	0.6518	0.6613

cross validation. A BP neural network is used to approximate the maps of nonlinear DR methods. The final class labels of test data are determined by KNN classifier. The dimension of data is reduced to the number of classes of each data set for all DR methods except for LDA.

Table 4 presents the classification accuracies of 12 methods on the nine data sets. It is interesting to find that LDA

performs better than SSD-Isomap on ORL and wine data sets. It seems that manifold methods generally perform worse on the two data sets. For ORL data set, SSD-Isomap is inferior to LDA, but comparable with SDA and SSDR. For wine data set, LDA, SDA and SSDR outperform SSD-Isomap, but SSD-Isomap performs best in the manifold methods. In general, SSD-Isomap shows the best or second

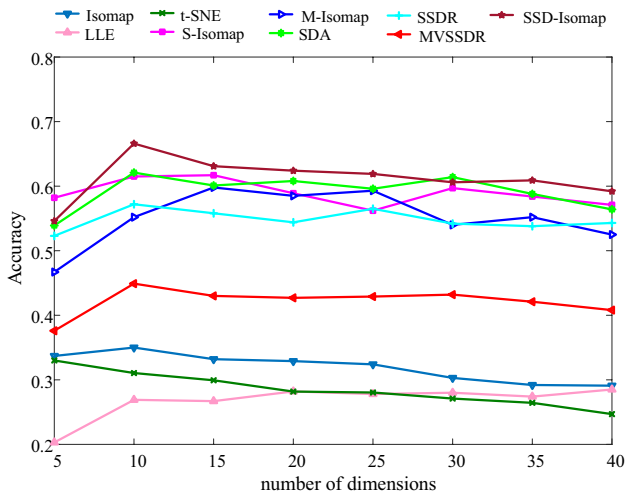


Fig. 3 Average precision comparison when number of dimensions changes on Caltech101 data set

Table 3 Six UCI data sets

Data set		Size	Classes	Attributes
Abbr.	Name			
Wine	Wine	178	3	13
Glass	Glass identification	214	6	10
Iris	Iris	150	3	4
Liv	Liver disorder	345	2	7
Seeds	Seeds	210	3	7
Sales	Wholesale customer	440	3	8

best performance in most data sets except wine, and achieves the highest averaged accuracy. Figure 4 shows the classification accuracy obtained by SSD-Isomap with different γ on wine data set. It can be seen that better performance can be achieved when γ ranges from 0.2 to 0.6.

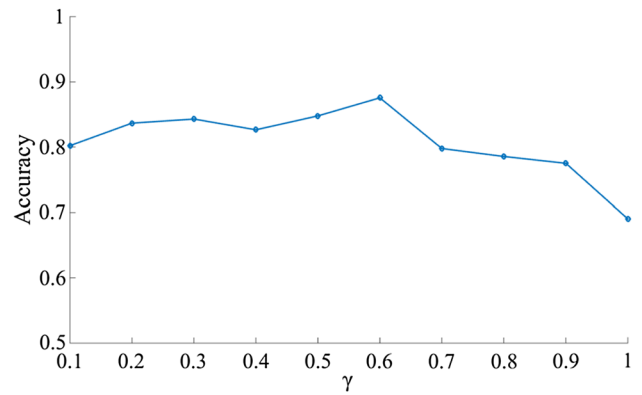


Fig. 4 Classification accuracy of SSD-Isomap with different values of γ on wine data set

5 Conclusions

A novel nonlinear DR method, SSD-Isomap, is presented in the paper. In SSD-Isomap, the original unsupervised Isomap is extended to the semi-supervised learning paradigm. We use a new constraint of LL from the unlabeled data to depict the local structure of data points, and the popular constraint of ML from the labeled data to present the class information. Based on the two constraints, graphs are constructed and the geodesic distance matrix is initialized. Then, the matrix is optimized by a similar procedure in Isomap and the low dimensional embedding is obtained. SSD-Isomap not only exploits useful information from both labeled and unlabeled data, but also applies the intrinsic nonlinear structure of the data. Therefore, it can achieve a discriminative lower-dimensional mapping. The extensive experimental comparisons between SSD-Isomap and other state-of-the-art methods have demonstrated that SSD-Isomap is more robust and effective in visualization, image retrieval and classification.

Table 4 Classification accuracy on the MNIST, YALE, ORL and UCI

	MNIST	YALE	ORL	Wine	Glass	Iris	Liv	Seeds	Sales	Ave.
Isomap	0.7625	0.3373	0.3650	0.7020	0.6356	0.9267	0.5913	0.8762	0.5593	0.6395
MDS	0.4775	0.3700	0.3210	0.6852	0.6166	0.9400	0.5623	0.8667	0.5141	0.5948
LE	0.8090	0.3171	0.4000	0.6743	0.6037	0.9133	0.5391	0.8810	0.5681	0.6340
LLE	0.4410	0.3457	0.4050	0.6967	0.4770	0.8267	0.5362	0.6667	0.5162	0.5457
t-SNE	0.8355	0.3253	0.4150	0.6681	0.6130	0.9667	0.5739	0.8333	0.5546	0.6428
LDA	0.6212	0.4321	0.4900	0.9717	0.5748	0.9600	0.6029	0.9238	0.5404	0.6797
S-Isomap	0.8175	0.1860	0.1600	0.9038	0.5977	0.9333	0.6870	0.8714	0.5932	0.6389
M-Isomap	0.7385	0.1220	0.1150	0.6575	0.6171	0.8467	0.7619	0.8095	0.7182	0.5985
SDA	0.7820	0.4375	0.4200	0.9606	0.6268	0.9533	0.5826	0.9381	0.5789	0.6978
SSDR	0.7960	0.4850	0.4280	0.9495	0.5741	0.9400	0.5971	0.9286	0.6043	0.7003
MVSSDR	0.7445	0.3452	0.3750	0.6685	0.5272	0.9160	0.5681	0.8095	0.5866	0.6156
SSD-Isomap	0.8410	0.4634	0.4300	0.8755	0.6512	0.9724	0.7449	0.9524	0.7259	0.7396

References

- Musa AB (2014) A comparison of ℓ_1 -regularization, PCA, KPCA and ICA for dimensionality reduction in logistic regression. *Int J Mach Learn Cybern* 5(6):861–873
- Sharma A, Paliwal KK (2015) Linear discriminant analysis for the small sample size problem: an overview. *Int J Mach Learn Cybern* 6(3):443–454
- Cai D, He X, Han J (2008) Training linear discriminant analysis in linear time. In: *IEEE 24th international conference on data engineering*, Cancun, pp 209–217
- Liu Y, Rong J (2006) Distance metric learning: a comprehensive survey. http://www.cs.cmu.edu/~liuy/frame_survey_v2.pdf. Accessed 6 May 2015
- Belkin M, Niyogi P (2002) Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. *Adv Neural Inf Process Syst* 14(6):585–591
- Raducanu B, Dornaika F (2012) A supervised non-linear dimensionality reduction approach for manifold learning. *Pattern Recognit* 45(6):2432–2444
- Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. *Science* 290(5500):2323–2326
- Laurens VDM (2014) Accelerating t-SNE using tree-based algorithms. *J Mach Learn Res* 15(1):3221–3245
- Tenenbaum JB, De SV, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. *Science* 290(5500):2319–2323
- Silva VD, Tenenbaum JB (2003) Global versus local approaches to nonlinear dimensionality reduction. In: *Advances in neural information processing systems*, pp 705–712
- Vlachos M, Domeniconi C, Gunopulos D (2002) Non-linear dimensionality reduction techniques for classification and visualization. In: *Proceeding of 8th ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, pp 645–651
- Geng X, Zhan DC, Zhou ZH (2005) Supervised nonlinear dimensionality reduction for visualization and classification. *IEEE Trans Syst Man Cybern Part B Cybern* 35(6):1098–1107
- Zhang Z, Chow TW, Zhao M (2012) M-Isomap: orthogonal constrained marginal isomap for nonlinear dimensionality reduction. *IEEE Trans Syst Man Cybern Part B Cybern* 43(1):180–191
- Yang B, Xiang M, Zhang Y (2016) Multi-manifold discriminant Isomap for visualization and classification. *Pattern Recognit* 55:215–230
- Meng M, Wei J, Wang J et al (2015) Adaptive semi-supervised dimensionality reduction based on pairwise constraints weighting and graph optimizing. *Int J Mach Learn Cybern* 8(3):793–805
- Chen WJ, Shao YH, Hong N (2013) Laplacian smooth twin support vector machine for semi-supervised classification. *Int J Mach Learn Cybern* 5(3):459–468
- Wang R, Wang XZ, Kwong S et al (2017) Incorporating diversity and informativeness in multiple-instance active learning. *IEEE Trans Fuzzy Syst* 25(6):1460–1475
- Luo Y, Tao D, Xu C (2013) Vector-valued multi-view semi-supervised learning for multi-label image classification. In: *Proceeding of 27th AAAI conference on artificial intelligence*, pp 647–653
- Zhu S, Sun X, Jin D (2016) Multi-view semi-supervised learning for image classification. *Neurocomputing* 208:136–142
- Zhu H, Wang X (2017) A cost-sensitive semi-supervised learning model based on uncertainty. *Neurocomputing* 251:106–114
- Ashfaq RAR, Wang XZ, Huang JZ et al (2017) Fuzziness based semi-supervised learning approach for intrusion detection system. *Inf Sci* 378:484–497
- Cai D, He X, Han J (2007) Semi-supervised discriminant analysis. In: *IEEE 11th international conference on computer vision*, pp 1–7
- Yang X, Fu H, Zha H, Barlow J (2006) Semi-supervised nonlinear dimensionality reduction. In: *Proceeding of 23th international conference on machine learning*, pp 1065–1072
- Hou V, Zhang C, Wu Y, Nie F (2010) Multiple view semi-supervised dimensionality reduction. *Pattern Recognit* 43(3):720–730
- Zhang D, Zhou ZH, Chen S (2007) Semi-supervised dimensional reduction. In: *Proceeding of the 7th SIAM international conference on data mining (SDM'07)*, pp 629–634
- Xing EP, Ng AY, Jordan MI (2003) Distance metric learning with application to clustering with side-information. In: *Advances in neural information processing systems*, pp 505–512
- Hoi SCH, Liu W, Lyu MR, Ma WY (2006) Learning distance metrics with contextual constraints for image retrieval. In: *IEEE Computer Society conference on computer vision and pattern recognition (CVPR'06)*, pp 2072–2078
- Xia H, Hoi SCH, Jin R, Zhao P (2014) Online multiple kernel similarity learning for visual search. *IEEE Trans Pattern Anal Mach Intell* 36(3):536–549
- Oliveira GL, Vieira AW, Vieira AW (2014) Sparse spatial coding: a novel approach to visual recognition. *IEEE Trans Image Process* 23(6):2719–2731
- Yang Y, Newsam S (2010) Bag-of-visual-words and spatial extensions for land-use classification. In: *Sigspatial international conference on advances in geographic information systems*. ACM, pp 270–279
- Li F-F, Fergus R, Perona P (2004) Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In: *Conference on computer vision and pattern recognition workshop*, pp 178–178
- Georghiadis AS, Belhumeur PN, Kriegman DJ (2001) From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans Pattern Anal Mach Intell* 23(6):643–660
- Peng X, Yu Z, Yi Z (2017) Constructing the L2-graph for robust subspace learning and subspace clustering. *IEEE Trans Cybern* 47(4):1053–1066
- Samaria FS, Harter AC (1994) Parameterization of a stochastic model for human face identification. In: *Proceedings of IEEE workshop on applications of computer vision*, pp 138–142
- Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. *Proc IEEE* 86(11):2278–2324
- Bache K, Lichman M (2013) UCI machine learning repository. <http://archive.ics.uci.edu/ml>. Accessed 26 June 2016
- Yu J, Tao D, Li J (2014) Semantic preserving distance metric learning and applications. *Inf Sci* 281:674–686
- Wu P, Hoi SCH, Zhao P, Miao C, Liu Z. Y (2016) Online multimodal distance metric learning with application to image retrieval. *IEEE Trans Knowl Data Eng* 28(2):454–467