Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 29 (2012) 3552 – 3556

Procedia Engineering

www.elsevier.com/locate/procedia

2012 International Workshop on Information and Electronics Engineering (IWIEE)

A General Model for Semi-Supervised Dimensionality Reduction

Xuesong Yin [a]*, Ting Shu[b], Qi Huang[c]

[a]*Department of Computer Science & Technology, Zhejiang Radio & TV University, Hangzhou, 310030, China*
[b]*College of Informatics and Electronics, Zhejiang Sci-Tech University, Hangzhou 310018, China*
[c]*School of Biological and Chemical Engineering, Zhejiang University of Science & Technology, 310023, China*

Abstract

This paper focuses on semi-supervised dimensionality reduction. In this scenario, we present a general model for semi-supervised dimensionality reduction with pairwise constraints (SSPC). Through defining a discriminant adjacent matrix, SSPC learns a projection embedding the data from the original space to the low-dimensional space such that intra-cluster instances become even more nearby while extra-cluster instances become as far away from each other as possible. Experimental results on a collection of benchmark data sets show that SSPC is superior to many established dimensionality reduction methods.

Keywords: Data mining; Dimensionality reduction; Pairwise constraints; Adjacent matrix

1. Introduction

Semi-supervised dimensionality reduction, aiming to obtain a low-dimensional faithful representation of high-dimensional data with side information, has been applied to computer vision, statistical learning and pattern recognition [1-4]. In fact, Utilizing side information has been an important issue in many data mining tasks [5-7]. Generally, in semi-supervised scenarios, side information may show diverse forms, such as class labels, pairwise constraints, prior membership degree or other prior information. According to given side information, semi-supervised dimensionality reduction methods can generally fall into two approaches. The first kind of approaches adopts pairwise constraints to guide the dimensionality reduction process. Other semi-supervised DR methods based on pairwise constraints are related to semi-supervised clustering [8-10]. The second kind of approaches applies available class labels to steer the DR process. In

* Corresponding author. Tel.: +0-86-571-88086839; fax: +0-86-571-88823529.
E-mail address: yinxs@nuaa.edu.cn.

1877-7058 © 2011 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.proeng.2012.01.529

[11], a semi-supervised DR framework is proposed, which adds a regularized term to the objective function of SDA and MMC [12]. Another general semi-supervised DR framework applies pairwise distances of embedded points to find projective direction [13].

In this paper, we focus on side information in the form of pairwise constraints, and present a general model for semi-supervised dimensionality reduction with pairwise constraints (SSPC). Concretely, we define a discriminant adjacent matrix in support of clustering and then learn a projection mapping the input data into a embedding space such that instances involved by must-link constraints become even more close while instances involved by cannot-link constraints are as far away from each other as possible. Moreover, SSPC can perform linear and non-linear mappings.

2. The algorithm

2.1. Model formulation

Given a set of data set $X = [x_1,\ldots, x_i,\ldots, x_n]$ ($x_i \in R^D$) together with some must-link constraints (M) and cannot-link constraints (C), we aim at finding a group of projective vectors $A = [a_1, a_2,\ldots,a_d]$ such that instances in the same cluster should be close while ones in different clusters should be far in the transformed low-dimensional space. To this end, we learn a transformation $f(.)$ from the input space to convert all the instances to the transformed space. Thus, we minimize the following loss function:

$$J = \frac{1}{2}\sum_{i,j}\left\| f(x_i) - f(x_j)\right\|^2 S_{ij} \tag{1}$$

where S is the discriminant adjacent matrix and is given its value later.

In order to clearly express Eq. (1), we split it into two parts:

$$J = \Delta + \delta \tag{2}$$

where

$$\Delta = \frac{1}{2n^2}\sum_{i,j}\left\| f(x_i) - f(x_j)\right\|^2 P_{ij} \tag{3}$$

and

$$\delta = \frac{\alpha}{2n_M}\sum_{(x_i,x_j)\in M}\left\| f(x_i) - f(x_j)\right\|^2 - \frac{\beta}{2n_C}\sum_{(x_i,x_j)\in C}\left\| f(x_i) - f(x_j)\right\|^2 \tag{4}$$

An edge is put between nodes i and j if x_i and x_j are close, i.e. x_i and x_j are among k nearest neighbors. S is the corresponding weight matrix, which is defined as follows:

$$F_{ij} = \begin{cases} 1 & if\ x_i \in N_k(x_j)\ or\ x_j \in N_k(x_i) \\ 0 & otherwise \end{cases} \tag{5}$$

where $N_k(x_i)$ denotes the set of k nearest neighbors of x_i. α and β are the trade-off parameters.

According to Eqs. (3) and (4), S can be expressed as follows:

$$S_{ij} = \begin{cases} \frac{F_{ij}}{n^2} + \frac{\alpha}{n_M} & if\ (x_i,x_j)\in M \\ \frac{F_{ij}}{n^2} - \frac{\beta}{n_C} & if\ (x_i,x_j)\in C \\ \frac{F_{ij}}{n^2} & otherwise \end{cases} \tag{6}$$

2.2. Optimization of the objective function

We use $f(x_i) = [f_1(x_i), \ldots, f_d(x_i)]^{\mathrm{T}}$ to project each instance x_i from R^D to the new space R^d. We assume that each component $f_j(.)$ is a linear combination of r basis functions and get the following equation:

$$f_j(x_i) = \sum_{l=1}^{r} P_{lj}\varphi_l(x_i) \quad (i = 1,\ldots,n, j = 1,\ldots,d) \tag{7}$$

where $\varphi_l(.)$s are the linear or non-linear basis functions and $P = [P_{lj}]_{\mathrm{D}\times\mathrm{d}}$ contains the weights. For ease of calculation, we constrain that $\sum_{l=1}^{r} P_{lj}^2 = 1$. If denoting $\varphi(x_i) = [\varphi_1(x_i), \ldots, \varphi_r(x_i)]^{\mathrm{T}}$, we have:

$$f(x_i) = [f_1(x_i),\ldots, f_d(x_i)]^T = P^T\varphi(x_i) \tag{8}$$

We use Eq. (8) to display Eq. (1) and obtain as follows:

$$\begin{aligned} J &= \frac{1}{2}\sum_{i,j}\left\|f(x_i) - f(x_j)\right\|^2 S_{ij} \\ &= \frac{1}{2}\sum_{i,j}\left\|P^T\varphi(x_i) - P^T\varphi(x_j)\right\|^2 S_{ij} \\ &= Trace(P^T\varphi(X)(D-S)\varphi^T(X)P) \\ &= \sum_{l=1}^{d} P_l^T[\varphi(X)(D-S)\varphi^T(X)]P_l \end{aligned}$$

where $P = [P_1,\ldots,P_d]$, $X = [X_1,\ldots,X_n]$ and $\varphi(X) = [\varphi_1(x_1), \ldots, \varphi_r(x_n)]$. $D \in R^{n\times n}$ is a diagonal matrix; its entries are column (or row) sum of S, $D_{ii} = \sum_j S_{ij}$.

We can obtain the optimal matrix P as follows:

$$\begin{aligned} &\min \sum_{l=1}^{d} P_l^T[\varphi(X)(D-S)\varphi^T(X)]P_l \\ &\text{s.t. } P_l^T P_l = 1 \end{aligned} \tag{9}$$

By introducing the Lagrangian, we have

$$L(P_l, \lambda_l) = \sum_{l=1}^{d} P_l^T[\varphi(X)(D-S)\varphi^T(X)]P_l - \lambda_l(P_l^T P_l - 1) \tag{10}$$

with the multiplier λ_l. This Lagrangian is minimized with respect to λ_l and P_l. By taking the derivatives of L and setting it to zero, we find that the solution is

$$\varphi(X)(D-S)\varphi^T(X)P_l = \lambda_l P_l \tag{11}$$

According to Eq. (11), we get

$$P_l = \varphi(X)w_l \tag{12}$$

where

$$w_l = \varphi(X)(D-S)\varphi^T(X)P_l / \lambda_l \in \Re^n \tag{13}$$

We substitute the expression of P_l into Eq. (11) and get

$$\varphi(X)(D-S)\varphi^T(X)\varphi(X)w_l = \lambda_l\varphi(X)w_l \tag{14}$$

Further, if both sides of Eq. (14) are left multiplied with $\varphi^{\mathrm{T}}(X)$, we have

$$K(D-S)Kw_l = \lambda_l Kw_l \tag{15}$$

where $K=\varphi^{\mathrm{T}}(X)\,\varphi(X)$ is a kernel matrix.

Zhang et al. [14] proved that Eq (15) has the same eigenvalues as the following equation:

$$(D-S)Kw_l=\lambda_l w_l \tag{16}$$

Thus, w_l can be solved by Eq (16).

2.3. Basis function φ

We assume that the transformed low-dimensional representations $Y=[f(x_1),\ldots,f(x_n)]$ can preserve the structure of the original data set as well as instances in the same cluster should be close while ones in different clusters should be far. Concretely,

$$\begin{aligned} Y &= [\,P^T\varphi(x_1),\ldots,P^T\varphi(x_n)\,] \\ &= [w_1,\ldots,w_d]^T \begin{bmatrix} \varphi^T(x_1)\varphi(x_1),\ldots,\varphi^T(x_1)\varphi(x_n) \\ \vdots \\ \varphi^T(x_n)\varphi(x_1),\ldots,\varphi^T(x_n)\varphi(x_n) \end{bmatrix} \\ &= [w_1,\ldots,w_d]^T K \end{aligned}$$

where $K=\varphi^{\mathrm{T}}(X)\,\varphi(X)$.

Since we only need operate the dot product $\varphi^{\mathrm{T}}(X)\,\varphi(X)$, Gaussian kernel is used to get kernel functions.

$$K(x_1,x_2)=\exp(-\|x_1-x_2\|^2/2\sigma^2) \tag{17}$$

Thus, Based on the analysis above, we can obtain the transformed low-dimensional representations Y.

3. Experiments

Table 1. Clustering performance on four data sets (%)

Data set	SSDR	LMDM	SCREEN	SSPC
Segment	78.37	75.24	74.13	**81.55**
Letter	66.34	61.25	62.28	**70.44**
USPS	75.99	70.83	71.22	**79.58**
YaleFaceB	89.10	86.25	86.16	**93.73**

In this section, we present an empirical study to evaluate the SSPC algorithm in comparison with several other representative semi-supervised learning algorithms, such as SSDR [2], LMDM [10], SCREEN [9]. We compared all these algorithms on four benchmark data sets, including Segment, Letter (a-d), USPS and YaleFaceB. Segment contains 7classes with 2309 instances and 19 dimensions. Letter has 4 classes with 3096 instances and 16 dimensions. USPS contains 10 classes with 6000 instances and 256 dimensions. YaleFaceB contains 10 classes with 5850 instances and 1200 dimensions. We use K-means to all the data in the low-dimensional space. The clustering result is evaluated by *NMI* [1,8,9]. The experimental results are shown in Table 1. We observe that our algorithm can relatively outperform the other methods on almost all the experiments.

4. Conclusion

In this paper, we propose a general model for semi-supervised dimensionality reduction called SSPC, which exploits both cannot-link and must-link constraints together with unlabeled data. SSPC can preserve the intrinsic local structure of the data set as well as the pairwise constraints specified by users in

the transformed low-dimensional space. In addition, our method is non-iterative and immune to small sample size (SSS) problem. Experimental results on four benchmark data sets demonstrated the effectiveness of the proposed algorithm.

Acknowledgements

The research reported in this paper has been partially supported by Natural Science Foundation of Zhejiang Province under Grant No. Y1100349, National Science Foundation of China under Grant No. 61101111, Foundation of Open University of China under Grant No. GFQ1601, Doctoral Foundation of Zhejiang Radio & TV University and Science Foundation of Zhejiang Sci-Tech University (ZSTU)under Grant No.1004839-Y.

References

[1] Yin XS, Chen S, Hu E, Zhang D. Semi-Supervised Clustering with Metric Learning: An Adaptive Kernel Method. *Pattern Recognition*; 2010, 43(4), 1320-1333.

[2] Zhang D, Zhou Z, Chen S. Semi-supervised dimensionality reduction. *In Proceedings of SIAM Conference on Data Mining*; 2007, 629-634.

[3] Cai D, He X, Han J, Semi-supervised discriminant analysis, *in Proc. ICCV*; 2007, 1-7.

[4] Yin XS, Huang Q. Integrating Global and Local Structures in Semi-supervised Discriminant Analysis. *The Third International Symposium on Intelligent Information Technology Application*; 2009, 757-761.

[5] Bar-Hillel A, Hertz T, Shental N, Weinshall D. Learning a mahalanobis metric from equivalence constraints. *Journal of Machine Learning Research*; 2005, 6, 937–965.

[6] Wagsta K, Cardie C, Schroedl S. Constrained k-means clustering with background knowledge. *in ICML'01, Williamstown, MA*; 2001, 577–584.

[7] XING EP, NG AY, JORDAN MI, RUSSELL S. Distance metric learning, with application to clustering with side-information. *in NIPS 15*, MIT Press, Cambridge, MA; 2003, 505–512.

[8] Yin X, Hu E. Distance Metric Learning Guided Adaptive Subspace Semi-supervised Clustering. *Frontiers of computer science in china*; 2011, 5(1), 100-108.

[9] Tang W, Xiong H, Zhong S, Wu J. Enhancing semi-supervised clustering: a feature projection perspective, *in: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2007, pp. 707–716.

[10] Xiang S, Nie F and Zhang C. Learning a Mahalanobis distance metric for data clustering and classification. *Pattern Recognition*; 2008, 41(12), 3600 - 3612.

[11] Song Y, Nie F, Zhang C, Xiang S. A Unified framework for semi-supervised dimensionality reduction. *Pattern Recognition*; 2008, 41(3), 2789–2799.

[12] Li H, Jiang T, Zhang K. Efficient and robust feature extraction by maximum margin criterion, *IEEE Trans. Neural Networks*; 2006, 17 (1), 157–165.

[13] Chatpatanasiri R, Kijsirikul B. A unified semi-supervised dimensionality reduction framework for manifold learning. *Neurocomputing*; 2010, 73(3): 1631-1640.

[14] Zhang W, Xue X. Sun Z, Lu H, Guo Y. Metric learning by discriminant neighborhood embedding. *Pattern Recognition*; 2008, 41(6), 2086~2096.