Semi-Supervised Discriminant Analysis via Spectral Transduction

Deming Zhai[1]
dmzhai@jdl.ac.cn
Hong Chang[2]
hchang@jdl.ac.cn
Bo Li[1]
bli@jdl.ac.cn
Shiguang Shan[2]
sgshan@jdl.ac.cn
Xilin Chen[2]
xlchen@jdl.ac.cn
Wen Gao[13]
wgao@jdl.ac.cn

[1] School of Computer Science and Technology,
Harbin Institute of Technology,
Harbin, China

[2] Key Laboratory of Intelligent Information Processing,
Chinese Academy of Sciences,
Beijing,China

[3] Institute of Digital Media,
Peking University,
Beijing, China

Linear Discriminant Analysis (LDA) is a popular method for dimensionality reduction and classification. In real-world applications when there is no sufficient labeled data, LDA suffers from serious performance drop or even fails to work.In order to address this problem, several numerical solutions have been proposed.Another possible solution for SSS problem is to learn with both labeled and unlabeled data. It is more natural and reasonable since in reality we usually have a large supply of unlabeled data and comparatively insufficient labeled data.

In this paper, we continue to pursue in the direction of exploring label information from unlabeled training samples for LDA. Our proposed method(STSDA), comprises three stages. First, we formulate label transduction with labeled and unlabeled data as a convex optimization problem with pairwise constraints and solve it efficiently with a closed-form solution. Then, some unlabeled data with reliable class estimations are selected through a balanced strategy to augment the original labeled data set. At last, LDA with manifold regularization is performed. Compared with previous related methods, our work has advantages in the following aspects: 1) We take into account both label augmenting and local structure preserving. 2) The optimization problem is convex which could be solved effectively in an analytical manner with a global optimal solution. 3) The balanced data selection strategy is more effective than the preliminary method.

Let us consider the supervisory information in the form of pairwise similarity and dissimilarity constraints, which are included in $\mathscr{S}$ and $\mathscr{D}$, respectively.

$$\begin{aligned}\mathscr{S} &= \{(\mathbf{x}_i,\mathbf{x}_j)|\mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to the same class}\},\\ \mathscr{D} &= \{(\mathbf{x}_m,\mathbf{x}_n)|\mathbf{x}_m \text{ and } \mathbf{x}_n \text{ belong to different classes}\}.\end{aligned} \tag{1}$$

We denote each pairwise similarity constraint $(\mathbf{x}_i,\mathbf{x}_j)\in\mathscr{S}$ by an n-dimensi -onal indicator vector $\mathbf{u}_k$ (k-th, $k=1,\ldots,|\mathscr{S}|$), which has only two non-zero elements: $\mathbf{u}_k(i)=1$ and $\mathbf{u}_k(j)=-1$. Since the class indicators $\mathbf{z}_i$ and $\mathbf{z}_j$ are equal (both 1 or -1 for two class problem), we have $\mathbf{u}_k^T\mathbf{z}=0$. Let $\mathbf{U}=[\mathbf{u}_1,\ldots,\mathbf{u}_{|\mathscr{S}|}]$ be the positive constraints matrix, where $|\mathscr{S}|$ denotes the cardinality of $\mathscr{S}$. Then, the pairwise similarity constraints can be expressed as $\mathbf{U}^T\mathbf{z}=\mathbf{0}$. Similarly, each dissimilarity constraint $(\mathbf{x}_m,\mathbf{x}_n)\in\mathscr{D}$ can be represented by an indicator vector $\mathbf{v}_k$ with only two non-zero elements: $\mathbf{v}_k(m)=1$ and $\mathbf{v}_k(n)=1$. Define $\mathbf{V}=[\mathbf{v}_1,\ldots,\mathbf{v}_{|\mathscr{D}|}]$ as the negative constraints matrix. The pairwise dissimilarity constraints can be expressed as $\mathbf{V}^T\mathbf{z}=\mathbf{0}$. Consequently, we have the following constrained Normalized Cuts formulation which is an extension of [3]:

$$\min_{\mathbf{z}} \frac{\mathbf{z}^T\mathbf{L}\mathbf{z}}{\mathbf{z}^T\mathbf{Q}\mathbf{z}} \text{ s.t. } \mathbf{U}^T\mathbf{z}=\mathbf{0}\,;\,\mathbf{V}^T\mathbf{z}=\mathbf{0}, \tag{2}$$

where $\mathbf{L}$ is called Laplacian matrix and $\mathbf{Q}$ is computed according to the similarities between each two samples.

Solving this constrained optimization problem can be facilitated by using orthogonal projection matrices. Let $\mathbb{U}\in\mathbb{R}^n$ be a subspace spanned by columns of $\mathbf{U}$ with $\mathbb{U}^\perp$ as its null orthogonal space. $\mathbf{P}_\mathbf{U}$ and $\mathbf{P}_{\mathbf{U}^\perp}$ are the orthogonal projection matrices onto $\mathbb{U}$ and $\mathbb{U}^\perp$, respectively. From the definition above, $\mathbf{P}_{\mathbf{U}^\perp}\mathbf{z}$ is the projection of $\mathbf{z}$ onto $\mathbb{U}^\perp$ and it satisfies the property as follows:

$$\forall\,\mathbf{z}\in\mathbb{U}^\perp\,,\,\mathbf{P}_\mathbf{U}\mathbf{z}=\mathbf{0}\,;\,\mathbf{P}_{\mathbf{U}^\perp}\mathbf{z}=\mathbf{z}. \tag{3}$$

According to [2], $\mathbf{P}_\mathbf{U}$ can be calculated directly from $\mathbf{U}$ as $\mathbf{P}_{\mathbf{U}^\perp}=\mathbf{I}-\mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T$, where $\mathbf{I}$ is the identity matrix. Therefore, if $\mathbf{z}$ is a feasible solution, it must satisfy $\mathbf{P}_{\mathbf{U}^\perp}\mathbf{z}=\mathbf{z}$. In the same sense, $\mathbf{P}_{\mathbf{V}^\perp}$ is defined as the orthogonal projection matrix on the null orthogonal space spanned by $\mathbf{V}$ and it can be computed in a similar way. With these transformations, Eq. (2) can be expressed as:

$$\min_{\mathbf{z}} \frac{\mathbf{z}^T\mathbf{L}\mathbf{z}}{\mathbf{z}^T\mathbf{Q}\mathbf{z}} \text{ s.t. } \mathbf{P}_{\mathbf{U}^\perp}\mathbf{z}=\mathbf{z}\,;\,\mathbf{P}_{\mathbf{V}^\perp}\mathbf{z}=\mathbf{z}. \tag{4}$$

The solution to this optimization problem can be finally obtained by solving a generalized eigenvalue problem which be formulated as Eq. (5).

$$\mathbf{L}\mathbf{P}_{\mathbf{U}^\perp}\mathbf{P}_{\mathbf{V}^\perp}\mathbf{z}=\lambda\mathbf{Q}\mathbf{z} \tag{5}$$

Subsequently,we aim to select some unlabel data points whose label estimation are with sufficiently high confidence. We first perform LDA using all the training data with their estimated labels. Then, in the embedding space, the label confidence is defined according to local data distribution. Specifically, for each unlabeled data $\mathbf{x}_i$ $(i=l+1,\ldots,n)$, its label confidence is defined as the proportion of data points with the same estimated label as $\mathbf{x}_i$ among its k nearest neighbors. For all the unlabeled data with the same estimated class labels, we sort their confidence values in descending order and select the first m samples to augment the original labeled data set where m is a selection scale factor.

With the augmented labeled data set, we seek to find a global projection that can not only improve class discriminative ability but also preserve local data structure. We take into account local structure preserving through Laplacian regularization [1]. The optimization problem of the regularized LDA can be written as:

$$\mathbf{W}^*=\max_{\mathbf{W}}\operatorname{trace}\left(\frac{\mathbf{W}^T\mathbf{S}_b\mathbf{W}}{\mathbf{W}^T\mathbf{S}_\mathrm{w}\mathbf{W}+\alpha K(\mathbf{W})}\right). \tag{6}$$

The scatter matrixes $\mathbf{S}_b$ and $\mathbf{S}_w$ are computed with the augmented labeled data set. $K(\mathbf{W})$ is the Laplacian regularization term and the coefficient α controls the relative importance of the discrimination and regularization. $K(\mathbf{W})$ is defined as:

$$K(\mathbf{W})=\sum_{i,j=1}^{n}(\mathbf{W}^T\mathbf{x}_i-\mathbf{W}^T\mathbf{x}_j)^2\mathbf{S}_{ij}=\mathbf{W}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{W}, \tag{7}$$

where $\mathbf{X}=[\mathbf{x}_1,\ldots,\mathbf{x}_n]$ is the matrix form of the whole data set. With the Laplacian regularizer, all data points are involved in the optimization. Thus the local geometric structure of both labeled and unlabeled data tends to be preserved with the transformation $\mathbf{W}$. The discriminant projector $\mathbf{W}$ can be computed efficiently by solving the following generalized eigendecomposition problem: $\mathbf{S}_b\mathbf{W}=\lambda(\mathbf{S}_w+\alpha\mathbf{X}\mathbf{L}\mathbf{X}^T)\mathbf{W}$.

[1] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from examples. *Journal of Machine Learning Research*, pages 2399–2434, 2004.

[2] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in Mathematical Sciences, 1996.

[3] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8): 888–905, 2000.