SCIENCE CHINA
Information Sciences

• LETTER •

January 2020, Vol. 63 119204:1–119204:3
https://doi.org/10.1007/s11432-018-9535-9

Kernel semi-supervised graph embedding model for multimodal and mixmodal data

Qi ZHANG[1,4], Rui LI[2] & Tianguang CHU[3*]

[1]*School of Information Technology & Management, University of International Business & Economics, Beijing* 100029*, China;*
[2]*School of Mathematical Sciences, Dalian University of Technology, Dalian* 116024*, China;*
[3]*State Key Laboratory for Turbulence and Complex Systems, College of Engineering, Peking University, Beijing* 100871*, China;*
[4]*Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing* 100871*, China*

Received 23 May 2018/Accepted 29 June 2018/Published online 8 October 2019

Citation Zhang Q, Li R, Chu T G. Kernel semi-supervised graph embedding model for multimodal and mixmodal data. Sci China Inf Sci, 2020, 63(1): 119204, https://doi.org/10.1007/s11432-018-9535-9

Dear editor,

Semi-supervised learning has obtained increasing interests in machine learning, because making use of both labeled and unlabeled training samples helps extracting discriminative features and meanwhile reduces the time-consuming and labor-intensive labeling burden. For extracting features upon multimodal (i.e., data of the same class exhibits separate clustering) and mixmodal (i.e., data from different classes has mixed modality) data [1], we have presented a semi-supervised graph embedding (SGE) model in [2] to incorporate the soft label information with hierarchical locality of data. Through the maximizing process upon the weighted between-class separability as well as the minimizing processes upon the locality-preserved within-class and scaled overall-class data distances respectively, the intrinsic characters of data with multimodal or mixmodal distributing properties can be well captured.

However, as the SGE model is a linear technique, it might not always give satisfying results in capturing the nonlinear structural characteristics of multimodal and mixmodal data. According to the kernel theory, when data is mapped nonlinearly with a kernel operator into a high-dimensional dot product space, the nonlinear dimensionality reduction problems can be efficiently solved linearly [3, 4]. This motivates us to extend the SGE model to nonlinear case with the aid of kernel technique. To be specific, to better solve the nonlinear dimensionality reduction problem upon multimodal and mixmodal data, we present herein a kernel semi-supervised graph embedding (KSGE) approach by incorporating SGE with kernel theory.

In literature, a number of kernel learning methods have been proposed, such as the MSSKSC model [5], and KSPP model [6]. But most of the existing models merely take into account the local correlations of data samples and thus might not always achieve satisfying results when dealing with multimodal and mixmodal data. Compared with these models, our KSGE model exploits the "hierarchical locality" preservation of data to make better manipulation of multimodal and mixmodal data in semi-supervised settings.

Experiments for handwriting recognition in both multimodal or mixmodal situations are carried out to evaluate the feasibility and effectiveness of the proposed KSGE model.

Methodology and analysis. To capture the hierarchical local geometric information of data, we first introduce a number of affinity matrices to incorporate the soft label information with the local affinity information of within-class, between-class, and overall-class data.

Let $\boldsymbol{X} = [\boldsymbol{X}_L, \boldsymbol{X}_U] \in \mathbb{R}^{n\times(l+u)}$ denote the

* Corresponding author (email: chutg@pku.edu.cn)

training data matrix, where $\boldsymbol{X}_L = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_l] \in \mathbb{R}^{n\times l}$ represents the labeled set, $\boldsymbol{X}_U = [\boldsymbol{x}_{l+1}, \boldsymbol{x}_{l+2}, \ldots, \boldsymbol{x}_{l+u}] \in \mathbb{R}^{n\times u}$ is the unlabeled set, and $l+u=p$. Utilizing the soft labels for $\boldsymbol{X}_U$ obtained from the label propagation process [1], the data matrix can be rewritten as $\boldsymbol{X} = [\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \ldots, \boldsymbol{X}^{(c)}]$. We first introduce a within-class affinity matrix for the m-th class by

$$W_{ij}^{(m)} = \begin{cases} \exp\left[-\dfrac{\|\boldsymbol{x}_i^{(m)} - \boldsymbol{x}_j^{(m)}\|^2}{{\sigma^{(m)}}^2}\right], \\ \qquad \text{if } \boldsymbol{x}_i^{(m)} \in N_m\big(\boldsymbol{x}_j^{(m)}\big) \\ \qquad \text{or } \boldsymbol{x}_j^{(m)} \in N_m\big(\boldsymbol{x}_i^{(m)}\big), \\ 0, \text{ otherwise}, \end{cases}$$

where $\boldsymbol{x}_i^{(m)}$ and $\boldsymbol{x}_j^{(m)}$ denote the i-th and j-th samples of the m-th class respectively, $i, j = 1, 2, \ldots, p_m$, $\sigma^{(m)}$ is the Gaussian kernel parameter, $N_m(\cdot)$ represents the set of $k^{(m)}$ nearest neighbors, and $m = 1, \ldots, c$. We also define the between-class weight matrix and the overall-class affinity matrix by

$$W_{ij} = \begin{cases} \exp\left[\dfrac{\|\boldsymbol{u}_i - \boldsymbol{u}_j\|^2}{\sigma^2}\right], \\ \qquad \text{if } \boldsymbol{u}_i \in N_B(\boldsymbol{u}_j) \text{ or } \boldsymbol{u}_j \in N_B(\boldsymbol{u}_i), \\ 0, \text{ otherwise}, \end{cases}$$

and

$$A_{ij} = \begin{cases} \exp\left[-\dfrac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{\sigma^2}\right], \\ \qquad \text{if } \boldsymbol{x}_i \in N_A(\boldsymbol{x}_j) \text{ or } \boldsymbol{x}_j \in N_A(\boldsymbol{x}_i), \\ 0, \text{ otherwise}, \end{cases}$$

where $\boldsymbol{u}_i = \frac{1}{p_i}\sum_{t=1}^{p_i} \boldsymbol{x}_t^{(i)}$ and $\boldsymbol{u}_j = \frac{1}{p_j}\sum_{t=1}^{p_j} \boldsymbol{x}_t^{(j)}$ are the mean vectors for the i-th and j-th classes, and $N_B(\cdot)$ and $N_A(\cdot)$ indicate the neighborhood set of k_B and k_A nearest neighbors, respectively.

Note that to deal with the within-class multimodality issue, $\boldsymbol{W}^{(m)}$ assigns smaller weights for sample pairs from different modalities and relatively larger values for those from the same modality, by which the local neighborhood correlation of the multimodal data in each class can be captured. For mixmodal data, unlike many existing methods that directly computing the between-class distances, the weight matrix $\boldsymbol{W}$ defined above calculates the weighted between-class distances to capture the local correlations of different classes. Besides, the overall-class locality of data is preserved by $\boldsymbol{A}$ to balance the soft label supervision and data distribution, avoiding the over-reliance of the errors in the estimated soft labels. By using the above affinity matrices, the "hierarchical locality" of data can be well captured.

With the affinity matrices, we can calculate the within-class, between-class, and overall-class scatter matrices as

$$\boldsymbol{S}_W = \frac{1}{2}\sum_{m=1}^{c}\sum_{i,j}\big[\boldsymbol{x}_i^{(m)} - \boldsymbol{x}_j^{(m)}\big]W_{ij}^{(m)}\big[\boldsymbol{x}_i^{(m)} - \boldsymbol{x}_j^{(m)}\big]^{\mathrm{T}},$$

$$\boldsymbol{S}_B = \frac{1}{2}\sum_{i,j}(\boldsymbol{u}_i - \boldsymbol{u}_j)W_{ij}(\boldsymbol{u}_i - \boldsymbol{u}_j)^{\mathrm{T}},$$

$$\boldsymbol{S}_A = \frac{1}{2}\sum_{i,j}(\boldsymbol{x}_i - \boldsymbol{x}_j)A_{ij}(\boldsymbol{x}_i - \boldsymbol{x}_j)^{\mathrm{T}} = \boldsymbol{X}\boldsymbol{L}_A\boldsymbol{X}^{\mathrm{T}},$$

which give the locality-preserved distances of the samples within each class, the weighted locality-preserved distances of different classes, and local covariances of all the samples, respectively. For later use, we transform $\boldsymbol{S}_W$ and $\boldsymbol{S}_B$ as $\boldsymbol{S}_W = \boldsymbol{X}\boldsymbol{S}_1\boldsymbol{X}^{\mathrm{T}}$ and $\boldsymbol{S}_B = \boldsymbol{X}\boldsymbol{S}_2\boldsymbol{L}\boldsymbol{S}_2^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}$, where

$$\boldsymbol{S}_1 = \mathrm{diag}\big[\boldsymbol{L}^{(1)}, \boldsymbol{L}^{(2)}, \ldots, \boldsymbol{L}^{(c)}\big]_{p\times p}, \tag{1}$$

$$\boldsymbol{S}_2 = \mathrm{diag}\left[\frac{\boldsymbol{E}(p_1)}{p_1}, \frac{\boldsymbol{E}(p_2)}{p_2}, \ldots, \frac{\boldsymbol{E}(p_c)}{p_c}\right]_{p\times c}, \tag{2}$$

$\boldsymbol{E}(p_m) = [1, 1, \ldots, 1]^{\mathrm{T}} \in \mathbb{R}^{p_m}$, p_m is the number of the samples in the m-th class, and $\boldsymbol{L}^{(m)}$ is the Laplacian matrix corresponding to $\boldsymbol{W}^{(m)}$.

Further, to invoke the kernel theory, we shall make use of a nonlinear mapping $\phi(\boldsymbol{x})$ from $\mathbb{R}^n$ to a reproducing kernel Hilbert space $\mathcal{H}$ with the reproducing kernel $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle\phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j)\rangle$, where $\langle\cdot,\cdot\rangle$ denotes the inner product in $\mathcal{H}$ [7],

$$K_{ij} = \langle\phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j)\rangle = \exp\left[-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{\sigma^2}\right]. \tag{3}$$

Now we explore the basis vector $\boldsymbol{\beta}$ of the desired subspace by the following KSGE model:

$$\max_{\boldsymbol{\beta}} \frac{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{S}_2\boldsymbol{L}\boldsymbol{S}_2^{\mathrm{T}}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{\beta}}{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{K}(\boldsymbol{S}_1 + \zeta\boldsymbol{L}_A)\boldsymbol{K}^{\mathrm{T}}\boldsymbol{\beta}},$$

which can be tackled by solving the following generalized eigenvalue problem:

$$\boldsymbol{K}\boldsymbol{S}_2\boldsymbol{L}\boldsymbol{S}_2^{\mathrm{T}}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{\beta} = \lambda\boldsymbol{K}(\boldsymbol{S}_1 + \zeta\boldsymbol{L}_A)\boldsymbol{K}^{\mathrm{T}}\boldsymbol{\beta}, \tag{4}$$

where ζ is a confidence parameter that is used to avoid the over-reliance of the estimated soft-label information. The generalized eigenvectors $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_k$ of (4) are corresponding to the first k largest generalized eigenvalues. Algorithm 1 gives the optimization details of KSGE.

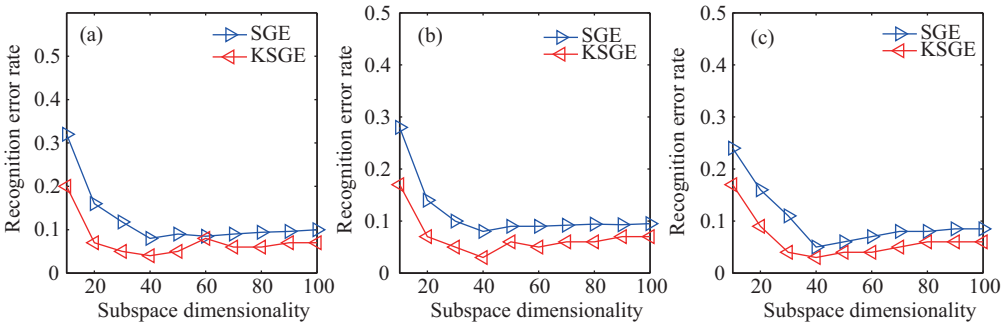

Figure 1 (Color online) Handwriting recognition. Average recognition error rates obtained by k-nn classifier ($k = 5$) for (a) USPS-eo, (b) USPS-sl, and (c) USPS-MNIST tasks.

Algorithm 1 KSGE

1: Input: Data matrix $\boldsymbol{X} = [\boldsymbol{X}_L, \boldsymbol{X}_U] \in \mathbb{R}^{n\times p}$.
2: Compute the within-class and between-class transform matrices $\boldsymbol{S}_1$, $\boldsymbol{S}_2$ in (1) and (2).
3: Compute the kernel matrix $\boldsymbol{K}$ in (3).
4: Solve the generalized eigenvalue problem (4).
5: Output: Subspace basis vectors $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_k$.

Discussion and analysis. The proposed KSGE model is tested by multimodal and mixmodal handwriting data. Three experiments are carried out including USPS-eo (to separate even numbers from odd numbers), USPS-sl (to separate small numbers ('0' to '4') from large numbers ('5' to '9')), and USPS-MNIST (to separate numbers from USPS and MNIST datasets). In USPS-eo and USPS-sl tasks, 1500 images are randomly chosen for training with 1/4 samples labelled. Another 1500 images are randomly selected for testing. In USPS-MNIST experiments, we select 700 images including equal amounts of digits "1" to "7" for training and another 700 images following the same choosing rule for testing. Similarly, 1/4 of the training samples are labelled. Binary class labels are generated for each task to indicate different classes of digits. Note that the USPS-eo and USPS-sl data follow multimodal distributions and the USPS-MNIST data exhibits mixmodal property. The testing images are mapped into the learned embedding subspaces with lower-dimensional representations, and we finish the recognition tasks by using the k-nearest neighbour (k-nn) classifier, where $k = 5$. For comparison, the SGE model is also applied in each experiment with 10 trials. The average handwriting recognition error rates for both methods are illustrated in Figure 1, with subspace dimensionalities ranging from 10 to 100. It is clear by Figure 1 that the proposed KSGE model exhibits better recognition performance.

Conclusion. A semi-supervised kernel method KSGE was proposed for multimodal and mixmodal data. The method can preserve the hierarchical locality, capturing the local geometric information of data in within-class, between-class, and overall-class scales. In addition, the present KSGE model can well incorporate the soft-label supervision with the data distributing information by using the introduced confidence parameter. Experiments for both multimodal and mixmodal data verify the efficiency of the proposed KSGE model.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61673027, 61503375) and Fundamental Research Funds for the Central Universities (Grant Nos. CXTD10-05, 18QD18 in UIBE, DUT19LK18).

References

1 Zhang Q, Chu T G. Semi-supervised discriminant analysis based on sparse-coding theory. In: Proceedings of the 35th Chinese Control Conference, 2016. 7082–7087
2 Zhang Q, Chu T G. Learning in multimodal and mixmodal data: locality preserving discriminant analysis with kernel and sparse representation techniques. Multimed Tools Appl, 2017, 76: 15465–15489
3 Liu Y, Liao S Z. Kernel selection with spectral perturbation stability of kernel matrix. Sci China Inf Sci, 2014, 57: 112103
4 Tao J W, Chung F L, Wang S T. A kernel learning framework for domain adaptation learning. Sci China Inf Sci, 2012, 55: 1983–2007
5 Mehrkanoon S, Suykens J A K. Scalable semi-supervised kernel spectral learning using random fourier features. In: Proceedings of IEEE Symposium Series on Computational Intelligence, 2017
6 Zhang A, Gao X W. Data-dependent kernel sparsity preserving projection and its application for semi-supervised classification. Multimed Tools Appl, 2018, 77: 24459–24475
7 Schölkopf B, Smola A J. Learning with Kernels. Cambridge: MIT Press, 2002