Semi-supervised Low-Rank Mapping Learning for Multi-label Classification

Liping Jing[1], Liu Yang[1], Jian Yu[1], Michael K. Ng[2]
[1]Beijing Key Lab of Traffic Data Analysis and Mining, Beijing Jiaotong University. [2]Department of Mathematics, Hong Kong Baptist University.

With the rapid growth of online content such as images, videos, web pages, it is crucial to design a scalable and effective classification system to automatically organize, store, and search the content. In conventional classification, each instance is assumed to belong to exactly one class among a finite number of candidate classes. However, in modern applications, an instance can have multiple labels. For example, an image can be annotated by many conceptual tags in semantic scene classification. Multi-label data have ubiquitously occurred in many application domains: multimedia information retrieval, tag recommendation, query categorization, gene function prediction, medical diagnosis, drug discovery and marketing. An important and challenging research problem [1, 4] in multi-label learning is how to exploit and make use of label correlations.

In this paper, we develop a novel method for multi-label learning when there is only a small number of labeled data. Our main idea is to design a Semi-supervised Low-Rank Mapping (SLRM) from a feature space to a label space based on given multi-label data. More specifically, the SLRM model can be formularized as

$$\min_{\mathbf{U}} ||\mathbf{U}\hat{\mathbf{X}}-\mathbf{Y}||_F^2+\lambda||\mathbf{U}||_*+\gamma tr((\mathbf{U}\mathbf{X})\mathbf{L}(\mathbf{U}\mathbf{X})^T), \tag{1}$$

where $\hat{\mathbf{X}}=[\hat{\boldsymbol{x}}_1\ \hat{\boldsymbol{x}}_2\ \cdots\ \hat{\boldsymbol{x}}_{n_l}]$ indicates a set of labeled data with n_l instances ($\hat{\boldsymbol{x}}_i\in\mathbb{R}^d$ is a d-dimensional feature vector) and $\mathbf{Y}=[\boldsymbol{y}_1\ \boldsymbol{y}_2\ \cdots\ \boldsymbol{y}_{n_l}]$ is the corresponding label information ($\boldsymbol{y}_i\in\mathbb{R}^k$ is a k-dimensional label vector of the ith instance). $\check{\mathbf{X}}=[\check{\boldsymbol{x}}_1\ \check{\boldsymbol{x}}_2\ \cdots\ \check{\boldsymbol{x}}_{n_u}]$ ($\check{\boldsymbol{x}}_i\in\mathbb{R}^d$) refers to a set of unlabeled data with n_u instances. The whole data set is denoted as $\mathbf{X}=[\hat{\mathbf{X}},\check{\mathbf{X}}]$ with n instances ($n=n_l+n_u$). The main goal of SLRM is to effectively and efficiently find a good mapping from $\hat{\mathbf{X}}$ to $\mathbf{Y}$ using the whole data set $\mathbf{X}$.

In SLRM model, the least square loss function is used for data fitting between $\hat{\mathbf{X}}$ and $\mathbf{Y}$ for easy optimization. The second term is designed for exploiting the label correlations via low-rank regularization on the linear transformation $\mathbf{U}$. The linear transformation $\mathbf{U}$ can be characterized by its singular value decomposition:

$$\sum_{j=1}^{r}\mathbf{p}_j(\mathbf{U})\sigma_j(\mathbf{U})\big(\mathbf{q}_j(\mathbf{U})\big)^T \tag{2}$$

where $r=\min\{k,d\}$, $p_j(\mathbf{U})\in\mathbb{R}^k$ and $q_j(\mathbf{U})\in\mathbb{R}^d$ are singular vectors of $\mathbf{U}$, and $\sigma_j(\mathbf{U})$ is the jth singular value of $\mathbf{U}$. With loss of generality, we assume that $\sigma_1(\mathbf{U})\geq\sigma_2(\mathbf{U})\geq\cdots\geq\sigma_r(\mathbf{U})$. Then, the nuclear norm regularization can be employed to measure the complexity of $\mathbf{U}$: $\|\mathbf{U}\|_*=\sum_{j=1}^{r}\sigma_j(\mathbf{U})$. In this case, the linear transformation of each data point $\hat{\mathbf{x}}_i$ can be given by

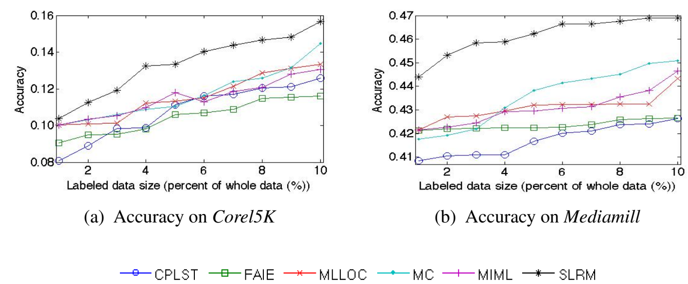
$$\mathbf{U}\hat{\mathbf{x}}_i=\sum_{j=1}^{r'}\sigma_j(\mathbf{U})[\big(\mathbf{q}_j(\mathbf{U})\big)^T\hat{\mathbf{x}}_i]\mathbf{p}_j(\mathbf{U}). \tag{3}$$

Obviously, the resulting vector is in the label space, and it is a linear combination of label-component vectors: $\mathbf{p}_1(\mathbf{U}),\mathbf{p}_2(\mathbf{U}),\cdots,\mathbf{p}_{r'}(\mathbf{U})$ which correspond to the largest r' singular values in the singular value decomposition of $\mathbf{U}$. Therefore, the label correlations can be recognized and represented by these label-component vectors.

The third term aims to make the mapping U capture the intrinsic geometric structure among data. Here, the heat kernel weight with self-tuning technique is used to construct a nearest neighbor graph for both labeled and unlabeled data $\mathbf{X}=[\hat{\mathbf{X}},\check{\mathbf{X}}]$. If two points are connected, $a_{i,j}=\exp\left(\frac{-\|\mathbf{x}_i-\mathbf{x}_j\|^2}{\sigma}\right)$, otherwise $a_{i,j}=0$. Then, we can generate an edge matrix $\mathbf{A}=[a_{i,j}]$ corresponding to the nearest neighbor graph. $\mathbf{L}$ is graph Laplacian of matrix $\mathbf{A}$ defined as $\mathbf{L}=\mathbf{D}-\mathbf{A}$, and $\mathbf{D}$ is a diagonal matrix whose main diagonal entries are column sums of $\mathbf{A}$, i.e., $d_{i,i}=\sum_{i=1}^{n}a_{i,j}$. This manifold regularization can model the local invariance assumption that when two instances are close in the feature space, their new representations based on mapping should be close.

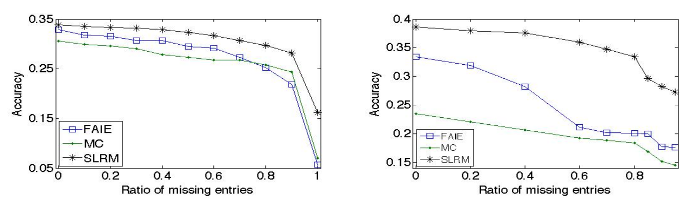
As a virtuous by-product, SLRM can handle missing labels because it has ability to fill such missing entries with label correlations and intrinsic structure among data, which is crucial as we may not have access to all the true labels of each training instance in most real applications [3].

The performance of SLRM is evaluated on four data sets including *MSRC*, *SUN*attribute database [2] and two Mulan multimedia datasets (*Core5K* and *Mediamill*). Five state-of-the-art multi-label classification methods (CPLST, FAIE, MLLOC, MC and MIML) are taken in our comparison.



(a) Accuracy on *Corel5K* (b) Accuracy on *Mediamill*

Figure 1: Comparison of six methods under varying the labeled data sizes on *Corel5K* and *Mediamill*.



(a) Accuracy on *SUN* (b) Accuracy on *MSRC*

Figure 2: Comparison results under varying the ratio of missing entries in label matrix (Y) on *SUN* and *MSRC*.

Figure 1 gives the label prediction performance under varying the labeled data size, where 1%-10% of data in each category are employed as training set. Obviously, the proposed method SLRM outperforms the other methods. It is interesting that SLRM performs well even when there are very few labeled data. Figure 2 demonstrates the performance by varying the ratio of missing labels (including positive and negative). SLRM clearly shows superior performance over other methods, especially on *MSRC*. We remark that handling *MSRC* is more difficult than handling *SUN* when partial labels are missing, because *MSRC* has less average labels in each instance than *SUN*.

[1] K. Dembczynski, W. Waegeman, W. Cheng, and E. Hullermeier. On label dependence and loss minimization in multi-label classification. *Machine Learning*, 88(1):5–45, 2012.

[2] G. Patterson, C. Xu, H. Su, and J. Hays. The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*, 108:59–81, 2014.

[3] H. Yu, P. Jain, P. Kar, and I. Dhillon. Large-scale multi-label learning with missing labels. In *Proc. of ICML*, pages 17–26, 2014.

[4] M. Zhang and Z. Zhou. A review on multi-label learning algorithms. *IEEE Trans. on Knowledge and Data Engineering*, 26(8):1819–1837, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.