Multi-label Learning with Specific Graph Laplacians and Specific Co-occurrence Matrices under Orthogonal constrains

FAN Min
Department of Mathematics
HKUST

ZHU Weizhi
Department of Mathematics
HKUST

Abstract

Differenet with common supervised learning, each example is represented by a single instance while associated with a set of labels simultaneously in multi-label learning. During the past decade, significant amount of progresses have been made towards this emerging machine learning paradigm[ZZ14]. Based on the way to predict test samples, multi-label learning algorithms can be categorized into two categories: inductive method and transductive method, where learning and prediction are separated in former one while learning and prediction are combined together in latter one. WU Baoyuan's work, which is transductive method, considers a ubiquitous problems in multi-label learning, that is, class imbalance. He put up with two types of class imbalance bounds. However, with unsuitable trade-off parameter, his model could get trivial result with all prediction positive or negative which stems from the property of graph laplacian. This report proposes orthogonal constrains to solve this trivial result and make algorithms more effective and robust with trade-off parameters. Besides, intuitively, different class should have its own graph laplacian and samples in different region should have different co-occurrence matrix. Specific graph laplacians and specific co-occurrence matrices are proposed to capture these properties. In the end, one assumption that some classes is generated from several sources linearly. Independent Component Analysis could be used to find a tighter bound, called independent component bound (ICA bound), compared with CIB bound.

1 Introduction

Multi-label learning (ML) assumes that one instance can be assigned to multiple classes simultaneously. For example, one image can be annotated with several tags, and one document can be associated with multiple topics. Significant amount of progresses have been made towards this emerging machine learning paradigm in recent years [ZZ14]. There are several ways to categorize ML algorithms like probabilistic [KVJ12] or deterministic [YJKD14] model, label space reduction [LDHW14] or not [KVJ12], handling missing labels [WSW$^+$14] or not [CL12], Large-scale [WSZ14] or not [TL12] and inductive [CL12] or transductive [WLW$^+$14]. In inductive method, learning could be separate as training process and prediction process individually. In prediction process, only the predictor/model after training process is needed and all training data is not. While transductive method combines training and prediction process together. There is no predictor and the prediction of testing label could be done without training data.

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Wu Baoyuan [WLG16] proposed a transductive model which consisted of label consistency, class-level label smoothness and instance-level label smoothness.

$$\min_{\mathbf{Z}\in\{-1,+1\}^{n\times m}} \mathrm{tr}(\mathbf{Y}_\mathbf{P}^\mathrm{T}(\mathbf{Y}-\mathbf{Z})) + \beta\mathrm{tr}(\mathbf{Z}\mathbf{L}_X\mathbf{Z}^\mathrm{T}) + \gamma\mathrm{tr}(\mathbf{Z}^\mathrm{T}\mathbf{L}_C\mathbf{Z}), \tag{1.1}$$

where $\mathbf{Z}$ is prediction with labels of samples as columns, $\mathbf{Y}$ is true label, $\mathbf{Y}_P$ is a constanct matrix to compute label consistency, $\mathbb{L}_X$ is grah laplacian and $\mathbf{L}_C$ is co-occurrence matrix.

He also considered another important challenge, class imbalance (CIB), which has two different phenomena. Firstly, each instance is assigned with only a few positive labels, while most other labels are negative. He referred to this type of class imbalance as CIB-1. Secondly, the proportions of positive instances of different classes may be significantly different. He referred to this type of CIB as CIB-2. CIB-1 often occurs in binary classification problems, while CIB-2 is more widely encountered in multi-label classification problems. Then he put up with two types of CIB bounds to alleviate CIB problems, CIB-1 bound

$$\bar{v}_1^l\mathbf{1}_n \le \mathbf{1}_m^\mathrm{T}\mathbf{Z} \le \bar{v}_1^u\mathbf{1}_n, \tag{1.2}$$

and CIB-2 bound

$$\bar{\boldsymbol{v}}_2^l \le \mathbf{Z}\mathbf{1}_n \le \bar{\boldsymbol{v}}_2^u. \tag{1.3}$$

As for class-level label smoothness, $\mathbf{L}_X$ is used to let samples near from each other have the same label as much as possible regardless different classes. As for instance-level label smoothness, $\mathbf{L}_C$ is used to push classes with high co-occurrence the same label regardless different instances.

2 Shortcomings in preivous model

Even though the previous model has good performence in many datasets, it still has some shortcomings in two aspects, roughly speaking.

Because of $\mathrm{tr}(\mathbf{Z}\mathbf{L}_X\mathbf{Z}^\mathrm{T})$, all rows of $\mathbf{Z}$ are pushed to $\mathbf{1}_n^\mathrm{T}\mathbf{D}_X^{\frac{1}{2}}$, where $\mathbf{D}_X$ is the degree matrix of $\mathbf{L}_X$. If β is very big compared with label consistency and γ, $\mathrm{tr}(\mathbf{Z}\mathbf{L}_X\mathbf{Z}^\mathrm{T})$ will dominate, which leads to one phenomenon that in one class, all samples are predicted to positive or negative at the same time. And similarly, because of $\mathrm{tr}(\mathbf{Z}^\mathrm{T}\mathbf{L}_C\mathbf{Z})$, all columns of $\mathbf{Z}$ are pushed to $\mathbf{D}_C^{\frac{1}{2}}\mathbf{1}$, where $\mathbf{D}_C$ is the degree matrix of $\mathbf{L}_C$. If γ is very big compared with label consistency and β, $\mathrm{tr}(\mathbf{Z}^\mathrm{T}\mathbf{L}_C\mathbf{Z})$ will dominate, which leads to one phenomenon that in one instance, all classed are assigned to positive or negative at the same time. At the worst case, when β and γ are both big enough, label consistency will not work at all. Under the influence of both $\mathrm{tr}(\mathbf{Z}\mathbf{L}_X\mathbf{Z}^\mathrm{T})$ and $\mathrm{tr}(\mathbf{Z}^\mathrm{T}\mathbf{L}_C\mathbf{Z})$, it is possible that $\mathbf{Z}$ could be predicted to all positive or negative.

CIB-1 and CIB-2 bounds are comparaly loose bounds. Since $\mathbf{Z}$ are loosed from $\{-1,+1\}^{n\times m}$ to $[-1,+1]^{n\times m}$ in optimization, CIB-1 and CIB-2 bounds only put constrains on sum of labels in columns or rows rather than more detailed constrains. This is the reason why no matter what CIB-2 bounds is put on, the prediction $\mathbf{Z}$ could be averaged in rows to be in proportion to $\mathbf{1}_n^\mathrm{T}\mathbf{D}_X^{\frac{1}{2}}$ and no matter what CIB-1 bounds is put on, the prediction $\mathbf{Z}$ could be averaged in columns to be in proportion to $\mathbf{D}_C^{\frac{1}{2}}\mathbf{1}$.

3 Proposed model

Intuitively, the rows and columns of prediction $\mathbf{Z}$ should be pushed away from $\mathbf{1}_n^\mathrm{T}\mathbf{D}_X^{\frac{1}{2}}$ and $\mathbf{D}_C^{\frac{1}{2}}\mathbf{1}$, which are eigenvectors of $\mathbf{L}_X$ and $\mathbf{L}_C$ both with eigenvalue 0 separately in order to get rid of trivial result. So orthogonal constrains will be implemented to eliminate trivial result.

Different classes have their own feature types. It is not suitable to adopt the sampe graph laplacian to depict the similarity between instances in different classes. Furthermore, co-occurrence matrix should be different in different regions or subspaces. One same co-occurrence matrix will lose some information. Then, specific graph laplacian and specific co-occurrence matrices will be used in the new model to caputure these information as much as possible whereby to improve the performence.

The proposed model is showed below

$$\begin{aligned}\min_{\mathbf{Z}\in\mathbf{R}^{n\times m}} \operatorname{tr}(\mathbf{Y}_{\mathbf{P}}^{\mathrm{T}}(\mathbf{Y}-\tanh(\mathbf{Z})))+\beta\sum_{i=1}^{m}\boldsymbol{e}_i\tanh(\mathbf{Z})\mathbf{L}_X^i\tanh(\mathbf{Z})^{\mathrm{T}}\boldsymbol{e}_i^{\mathrm{T}} \\ +\gamma\sum_{i=1}^{k}\operatorname{tr}(\boldsymbol{e'}_i\tanh(\mathbf{Z})^{\mathrm{T}}\mathbf{L}_C^i\tanh(\mathbf{Z})\boldsymbol{e'}_i^{\mathrm{T}})\end{aligned} \tag{3.1}$$

with modified CIB-1, CIB-2 bounds and orthogonal constrains

$$\begin{aligned}&\bar{\boldsymbol{v}}_1^l \le \mathbf{1}_m^{\mathrm{T}}\tanh(\mathbf{Z}) \le \bar{\boldsymbol{v}}_1^u\mathbf{1}_n, \\ &\bar{\boldsymbol{v}}_2^l \le \tanh(\mathbf{Z})\mathbf{1}_n \le \bar{\boldsymbol{v}}_2^u, \\ &\boldsymbol{e}_i\mathbf{Z}\mathbf{D}_X^{\frac{1}{2}}\mathbf{1}_n = 0, \quad \text{with}\, i=1,...,m, \\ &\boldsymbol{e'}_i\mathbf{Z}^{\mathrm{T}}\mathbf{D}_C^{\frac{1}{2}}\mathbf{1}_m = \mathbf{0}^{n^k}, \quad \text{with}\, i=1,...,k.\end{aligned} \tag{3.2}$$

m is the number of classes and $\boldsymbol{e}_i$ refers to $[0,0,...,1,...,0]$ with only i-th entry 1 and the other $n-1$ entries 0. k is the number of regions or subspaces. $\boldsymbol{e'}_i$ refers to $\{0,1\}^{n^k\times m}$ where n^k is the number of instances in i-th region. Each row represents one instance with its entry 1 and the others 0. $\mathbf{L}_X^i$ and $\mathbf{L}_C^i$ are specific graph laplacians and specific co-occurrence matrices.

4 Theoretical Illustration and Experiments

4.1 Why Specific Graph Laplacian?

Different class have their own feature type. If we capture the different similarity matrix with respect to each class, we couldprobably improve performence of the algorithm.

In order to compute specific graph laplacian accurately, we need to get distance matrix as percise as possible with respect to each class. To achieve this goal, it is necessary to alleviate the influence of the manifold or subspace, which positive instance of each class lie on or which could discriminate positive and negative instances as much as possible, on the computation of distance matrix. We can use dimension reduction techniques to erase these dimensions of the manifold or subspace.

Experiments show that using principal component analysis (PCA) on positive instances, new coordinates of positive and negative instances in new subspace, which is orthogonal complement to the subspace spanned by principal components, did not separate very well. This might be because that the feature of each class lies on a nonlinear manifold instead of a linear subspaceAnd this leads to considerable error in computation of distance matrix.

Linear discriminative analysis (LDA) finds one vector/direction to discriminate positive and negative samples most significantly. Experiments show that it did work well but it has one defect that LDA reduce original space from dimension p to just 1 and could lose much information during this process.

4.2 Why Orthogonal constrains?

We mainly talk about orthogonal constrains of $\mathbf{L}_X^i$ and the circumstances of orthogonal constrains of $\mathbf{L}_C^i$ are similar with $\mathbf{L}_X^i$.

The spectrum of $\mathbf{L}_X^i$ is spec$\left\{\lambda_0^i,\lambda_1^i,...,\lambda_n^i\right\}$, where $\lambda_0^i = 0 \le \lambda_1^i \le ... \le \lambda_n^i$. And the corrosponding eigenvector are $\left\{\boldsymbol{v}_0^i,\boldsymbol{v}_1^i,...,\boldsymbol{v}_n^i\right\}$, where $\boldsymbol{v}_0^i = \mathbf{D}_X^{i\,\frac{1}{2}}\mathbf{1}_n$. The orthogonal constrain are $\boldsymbol{e}_i\mathbf{Z}\mathbf{D}_X^{i\,\frac{1}{2}}\mathbf{1}_n = 0$ with $i=1,...,m$. This means each rows of Z will be orthogonal to $\boldsymbol{v}_0 = \mathbf{D}_X^{i\,\frac{1}{2}}\mathbf{1}_n$. And these constrains guarantee the entries of each row of Z can not be all positive or negative as entries of $\mathbf{D}_X^{i\,\frac{1}{2}}\mathbf{1}_n$ are all positive whereby we can avoid trivial result.

Laplacian eigenmaps and spectral clustering show that using eigenvectors of $\mathbf{L}_X^i$, $\left[\boldsymbol{v}_1^i, ..., \boldsymbol{v}_q^i\right]$ is good representation of new coordinates reserving distance of instances well, where trivial one $\boldsymbol{v}_o^i$ is dropped out and q is the dimension of new space after dimension reduction. If two categories of data has its own feature type, spectral clustering could cluster or separate two categroies of data with new coordinates well.

Orthogonal constrains $\boldsymbol{e}_i\mathbf{Z}\mathbf{D}_X^{i\,\frac{1}{2}}\mathbf{1}_n = 0$ where $i = 1, ..., m$ will push i-th row of Z to ${\boldsymbol{v}_1^i}^{\mathrm{T}}$ under process of optimization of $\boldsymbol{e}_i\tanh(\mathbf{Z})\mathbf{L}_X^i\tanh(\mathbf{Z})^{\mathrm{T}}\boldsymbol{e}_i^{\mathrm{T}}$. After dimension reduction with respect to the feature type of each class, positive and negative sample are assumed to enable to be seperated or clustered with each center. So the eigenvector $\boldsymbol{v}_1^i$ of specific graph laplacian $\mathbf{L}_X^i$ is supposed to cluster positive and negative instances comparaly well despite it uses only one eigenvector. In the case of big β, i-th row of result Z is ${\boldsymbol{v}_1^i}^{\mathrm{T}}$. It is supposed to have good performence by using rounding or top-k on Z to have the prediction, while it will lead to trivial prediction witout orthogonal constrains with big β.

5 References

References

[CL12] Yao-Nan Chen and Hsuan-Tien Lin. Feature-aware label space dimension reduction for multi-label classification. In *Advances in Neural Information Processing Systems*, pages 1529–1537, 2012.

[KVJ12] Ashish Kapoor, Raajay Viswanathan, and Prateek Jain. Multilabel classification using bayesian compressed sensing. In *Advances in Neural Information Processing Systems*, pages 2645–2653, 2012.

[LDHW14] Zijia Lin, Guiguang Ding, Mingqing Hu, and Jianmin Wang. Multi-label classification via feature-aware implicit label space encoding. In *International Conference on Machine Learning*, pages 325–333, 2014.

[TL12] Farbound Tai and Hsuan-Tien Lin. Multilabel classification with principal label space transformation. *Neural Computation*, 24(9):2508–2542, 2012.

[WLG16] Baoyuan Wu, Siwei Lyu, and Bernard Ghanem. Constrained submodular minimization for missing labels and class imbalance in multi-label learning. In *AAAI*, pages 2229–2236, 2016.

[WLW$^+$14] Baoyuan Wu, Zhilei Liu, Shangfei Wang, Bao-Gang Hu, and Qiang Ji. Multi-label learning with missing labels. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 1964–1968. IEEE, 2014.

[WSW$^+$14] Qifan Wang, Bin Shen, Shumiao Wang, Liang Li, and Luo Si. Binary codes embedding for fast image tagging with incomplete labels. In *European Conference on Computer Vision*, pages 425–439. Springer, 2014.

[WSZ14] Qifan Wang, Luo Si, and Dan Zhang. Learning to hash with partial tags: Exploring correlation between tags and hashing bits for large scale image retrieval. In *European Conference on Computer Vision*, pages 378–392. Springer, 2014.

[YJKD14] Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit Dhillon. Large-scale multi-label learning with missing labels. In *International Conference on Machine Learning*, pages 593–601, 2014.

[ZZ14] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2014.