Maximum Margin Matrix Factorization for Ensemble Clustering

Abstract

We formulate the late score fusion algorithm as ensemble clustering problem, and propose an efficient optimization algorithm, namely Score Fusion via Ensemble Cluster(SFEC). SFEC has capability to remove outliers of label assignments from multiple classifiers using a $\ell_{2,1}$ loss, which improves the performance of late fusion (large scale ability). To the best of our knowledge, this is the first work to connect ensemble clustering with late fusion. Compared to the state-of-the-art fusion algorithm, the SFEC achieves promising improvements on large scale dataset such as UCF101

Introduction

A single clustering result could be inaccurate, so some researchers study ensemble clustering methods to boost the performance, such as the algorithms proposed in (Yi et al. 2012; Gao et al.), to name a few.

The key aim of matrix completion based methods is to detect the anomalous clusters (bad partitions/assigments) and reconstruct the real assignments, which is shown in Figure 1. A typical formulation of matrix completion based methods then can be written as below:

$$\begin{aligned} \min_{\mathbf{X}} \ & ||\mathbf{X}-\mathbf{L}||_2^2 \\ \text{s.t. } & \text{rank}(\mathbf{X}) \leq k, \end{aligned} \tag{1}$$

where k is a pre-defined rank constraint, and $\mathbf{L} \in \mathbb{R}^{N \times MK}$. Here we denote M and K as the number of single clustering assignments and the number of clusters, respectively. However, the least square loss used in matrix completion is sensitive to the abnormal errors, namely outliers, the not appropriate to capture the anomalous assignments. The authors in (Gao et al.) propose to achieve cluster-wise (column-wise) sparsity of $\mathbf{X}$ by the $\ell_{1,2}$ norm on $\mathbf{X}$, which is as follow

$$\begin{aligned} \min_{\mathbf{X}} \ & ||\mathbf{L}-\mathbf{X}||_F^2 + \beta||\mathbf{X}||_{1,2} \\ \text{s.t. } & \text{rank}(\mathbf{X}) \leq k, \end{aligned} \tag{2}$$

where $||\mathbf{X}||_{1,2} = \sum_{j=1}^{MK} \sqrt{\sum_{i=1}^{N} \mathbf{X}_{ij}} = \sum_{j=1}^{MK} ||\mathbf{X}_{.j}||_2$, which would make $\mathbf{X}$ column-wise sparse (beware the direction of the computation of the $\ell_{1,2}$ norm, more details in (Nie et al. 2010)). Nevertheless, it does not consider the outliers in their loss function. In fact, the term $||\mathbf{L}-\mathbf{X}||_F^2$ is sensitive to outliers, and the term $||\mathbf{X}||_{1,2}$ is equivalent to $||\mathbf{X}-\mathbf{O}||_{1,2}$, which makes $\mathbf{X}$ fit $\mathbf{O}$, where $\mathbf{O}$ is a matrix with all entries as zero. Therefore, we propose to apply $\ell_{1,2}$ to capture the outliers as well as anomalous columns, which can be written as below

$$\begin{aligned} \min_{\mathbf{X}} \ & ||\mathbf{L}-\mathbf{X}||_{1,2} \\ \text{s.t. } & \text{rank}(\mathbf{X}) \leq k. \end{aligned} \tag{3}$$

Problem (3) is NP-hard due to the presence of the low rank constraint. For the sake of efficiency on large scale matrices, we consider matrix factorization approaches to optimize it, which can be written as

$$\begin{aligned} \min_{\mathbf{X}} \ & ||\mathbf{L}-\mathbf{X}||_{1,2} \\ \text{s.t. } & \mathbf{X}=\mathbf{U}\mathbf{V},\ \mathbf{U} \in \mathbb{R}^{N \times k},\ \mathbf{V} \in \mathbb{R}^{k \times MK}. \end{aligned} \tag{4}$$

Note that we have totally K single clustering assignments. Intuitively, we could set $k = K$ in Problem (4). Most existing matrix factorization optimization methods only consider smooth loss functions, but the $\ell_{1,2}$ loss in Problem (4) is nonsmooth. In this paper, we apply augmented Lagrangian multiplier (ALM) to optimize the nonsmooth objective function efficiently.

Related Work

The Proposed Approach

We first elaborate the formulation of our SFEC algorithm. Then we show the details about how to use ensemble clustering to fuse multiple classifiers' scores. Finally, we introduce a novel optimization for the $\ell_{2,1}$ loss ensemble clustering problem.

ALM Optimization

By introducing a new variable $\mathbf{E} = \mathbf{L} - \mathbf{X}$, we can develop Problem (4) as below:

$$\begin{aligned} \min_{\mathbf{E},\mathbf{X}} \ & ||\mathbf{E}||_{1,2} \\ \text{s.t. } & \mathbf{E}=\mathbf{L}-\mathbf{X} \\ & \mathbf{X}=\mathbf{U}\mathbf{V},\ \mathbf{U} \in \mathbb{R}^{N \times k},\ \mathbf{V} \in \mathbb{R}^{k \times MK} \end{aligned} \tag{5}$$

The augmented Lagrangian function of Problem (4) is as below

$$\mathcal{L} = ||\mathbf{E}||_{1,2} + \langle \boldsymbol{\lambda}, \mathbf{L} - \mathbf{X} - \mathbf{E} \rangle + \frac{\mu}{2}||\mathbf{L} - \mathbf{X} - \mathbf{E}||_F^2 \quad (6)$$

where $\mathbf{X} = \mathbf{UV}$, and $\boldsymbol{\lambda} \in \mathbb{R}^{MK \times N}$ are the Lagrangian multipliers (or dual variables). Then we could update $\mathbf{X}$ and $\mathbf{E}$ alternatively.

Specifically, to update $\mathbf{X}$, we consider the following problem:

$$\begin{aligned} \min_{\mathbf{X}} \ & \langle \boldsymbol{\lambda}, \mathbf{L} - \mathbf{X} - \mathbf{E} \rangle + \frac{\mu}{2}||\mathbf{L} - \mathbf{X} - \mathbf{E}||_F^2 \\ \text{s.t.} \ & \mathbf{X} = \mathbf{UV}, \ \mathbf{U} \in \mathbb{R}^{N \times k}, \ \mathbf{V} \in \mathbb{R}^{k \times MK}. \end{aligned} \quad (7)$$

Problem (7) includes a smooth loss function w.r.t. $\mathbf{X}$, thus can be solved by a standard matrix factorization method, such as (Yan et al. 2015; Tan et al. 2014; Vandereycken 2013).

Problem (7) can also be rewritten as below

$$\begin{aligned} \min_{\mathbf{X}} \ & \frac{\mu}{2}\Big(||\mathbf{L} - \mathbf{X} - \mathbf{E}||_F^2 + \frac{2}{\mu}\langle \boldsymbol{\lambda}, \mathbf{L} - \mathbf{X} - \mathbf{E} \rangle + \frac{||\boldsymbol{\lambda}||_F^2}{\mu^2}\Big) \\ & - \frac{\mu}{2}\frac{||\boldsymbol{\lambda}||_F^2}{\mu^2} \\ \text{s.t.} \ & \mathbf{X} = \mathbf{UV}, \ \mathbf{U} \in \mathbb{R}^{N \times k}, \ \mathbf{V} \in \mathbb{R}^{k \times MK}. \end{aligned} \quad (8)$$

We can obtain the final problem w.r.t. $\mathbf{X}$:

$$\begin{aligned} \min_{\mathbf{X}} \ & ||\mathbf{L} - \mathbf{X} - \mathbf{E} + \frac{\boldsymbol{\lambda}}{\mu}||_F^2 \\ \text{s.t.} \ & \mathbf{X} = \mathbf{UV}, \ \mathbf{U} \in \mathbb{R}^{N \times k}, \ \mathbf{V} \in \mathbb{R}^{k \times MK}. \end{aligned} \quad (9)$$

There are a number of optimization algorithms for Problem (9), such as LRGeomCG[1].

To update $\mathbf{E}$, we consider the following problem:

$$\min_{\mathbf{E}} \ ||\mathbf{E}||_{1,2} + \langle \boldsymbol{\lambda}, \mathbf{L} - \mathbf{X} - \mathbf{E} \rangle + \frac{\mu}{2}||\mathbf{L} - \mathbf{X} - \mathbf{E}||_F^2. \quad (10)$$

Similarly, we could reformulate the above problem as below:

$$\min_{\mathbf{E}} \frac{2}{\mu}||\mathbf{E}||_{1,2} + ||\mathbf{L} - \mathbf{X} - \mathbf{E} + \frac{\boldsymbol{\lambda}}{\mu}||_F^2. \quad (11)$$

Let $\mathbf{Y} = \mathbf{L} - \mathbf{X} + \frac{\boldsymbol{\lambda}}{\mu}$. The above problem can be efficiently solved by the following column-wise soft-thresholding operator:

$$\mathbf{E}_i = \mathcal{S}_\alpha(\mathbf{Y}_i) = \begin{cases} \mathbf{0}, \ \text{if } ||\mathbf{Y}_i||_2 \leq \alpha \\ \mathbf{Y}_i - \dfrac{\alpha \mathbf{Y}_i}{||\mathbf{Y}_i||_2}, \ \text{otherwise}, \end{cases} \quad (12)$$

where $\mathbf{E}_i$ and $\mathbf{Y}_i$ denote the i-th column of $\mathbf{E}$ and $\mathbf{Y}$, and $\alpha = \frac{2}{\mu}$.

We summarize the proposed algorithm for Problem (5) in Algorithm

[1]Code available at http://www.unige.ch/math/vandereycken/matrix_completion.html

Algorithm 1 The ALM algorithm for Problem (5)

Initialize $\rho > 1$, $t = 0$, $\boldsymbol{\lambda}^{(t)} = 0$, and $\mu^{(t)} > 0$.
1: Obtain $\mathbf{X}^{(t)}$ by solving Problem (9) via LRGeomCG.
2: Obtain $\mathbf{E}^{(t)}$ by column-wise soft-thresholding (12).
2: $\boldsymbol{\lambda}^{(t+1)} = \boldsymbol{\lambda}^{(t)} + \mu^{(t)}(\mathbf{L} - \mathbf{X}^{(t+1)} - \mathbf{E}^{(t+1)})$.
3: $\mu^{(t+1)} = \rho\mu^{(t)}$.
4: t = t + 1.
5: Go to step 1 until convergence.

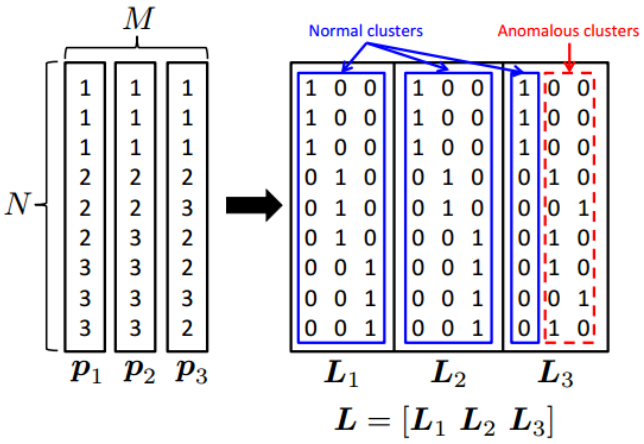

Figure 1: An example of anomalous cluster assignments which is retrieved from (Gao et al.)

An Option on Loss Function: Max Margin

In fact, the assignments $\mathbf{L}$ contains only $0/1$ values (discrete ordinal values). This fits the setting in (Yan et al. 2015) very much. It could be a contribution, since maximum margin loss achieves better performance on discrete ordinal values. Note: we may leave this extension as a journal?

Another Option on Loss Function: $\ell_{1,2}$ to ℓ_1 loss

$ell_{1,2}$ loss could be difficult to optimize. A simple way is to replace $\ell_{1,2}$ to ℓ_1, which would be much easier.

Experiments

We perform two experiments to show the effectiveness of the proposed outlier label pruning method. The first experiment is image classification with multiple features. The second experiment is ensemble classification. By these two experiments, we aim to demonstrate that our method is able to effectively remove the anomalous labels from the class indicator matrix, and thus improve the fusion performance of classification. We have the following datasets:

- Oxford flower 17
- Pascal Sentences
- Wikipedia
- mnist

- CIFAR10
- Imagenet

We compare our proposed method with the following five baselines:

- Score average
- MKL of multiple features (or base learners)
- LPBoost
- The method proposed in (Gao et al.)
- the method proposed in (Xu et al. 2013)

References

Gao, J.; Yamada, M.; Kaski, S.; Mamitsuka, H.; and Zhu, S. A robust convex formulation for ensemble clustering.

Nie, F.; Huang, H.; Cai, X.; and Ding, C. H. 2010. Efficient and robust feature selection via joint 2, 1-norms minimization. In *NIPS*, 1813–1821.

Tan, M.; Tsang, I. W.; Wang, L.; Vandereycken, B.; and Pan, S. J. 2014. Riemannian pursuit for big matrix recovery. In *ICML*, volume 32, 1539–1547.

Vandereycken, B. 2013. Low-rank matrix completion by riemannian optimization. *SIAM Journal on Optimization* 23(2):1214–1236.

Xu, Z.; Yang, Y.; Tsang, I.; Sebe, N.; and Hauptmann, A. G. 2013. Feature weighting via optimal thresholding for video analysis. In *ICCV*, 3440–3447.

Yan, Y.; Tan, M.; Tsang, I.; Yang, Y.; Zhang, C.; and Shi, Q. 2015. Scalable maximum margin matrix factorization by active riemannian subspace search. In *IJCAI*.

Yi, J.; Yang, T.; Jin, R.; Jain, A. K.; and Mahdavi, M. 2012. Robust ensemble clustering by matrix completion. In *ICDM*, 1176–1181.