Multi-view Multi-instance Multi-label Learning based on Collaborative Matrix Factorization

Supplementary file*

Yuying Xing¹, Guoxian Yu¹,²*, Carlotta Domeniconi³, Jun Wang¹, Zili Zhang¹,⁴ and Maozu Guo⁵

¹College of Computer and Information Science, Southwest University, Chongqing 400715, China
²Hubei Key Laboratory of Intelligent Geo-Information Processing, China University of Geosciences
³Department of Computer Science, George Mason University, Fairfax 22030, USA
⁴School of Information Technology, Deakin University, Geelong, VIC 3220, Australia
⁵School of Electrical and Information Engineering, Beijing University Of Civil Engineering and Architecture, China

This supplementary file elaborates on how to iteratively optimize G, α and β in the objective function (Eq. (3) of the main text) of M3L. Due to the fact that the objective function is not convex with respect to G1, G2 and G3, it is therefore unrealistic to find the global optimal solutions for them at the same time. Follow the idea of finding nonnegative matrix factorization (Lee and Seung 2001), we optimize G1, G2 and G3 using an alternative optimization technique by iteratively fixing two of them as constants while optimizing the other one. Due to the fact that both \( tr(R_1^T R_1) \) and \( tr(R_2^T R_2) \) are constants during optimization, we can rewrite the optimization problem as follows:

\[
\begin{align*}
\arg\min_{G_1, G_2, G_3} Z(G_1, G_2, G_3) &= tr(R_{12} G_1 R_{12}^T) + 2 tr(R_{22} G_2 R_{22}^T) \\
&= -2 tr(R_{12} G_1 G_1^T) - 2 tr(R_{12} G_2 G_3) - 2 tr(R_{22} G_2 G_2^T) \\
&= -2 tr(G_1 R_{12}^T G_1^T) + tr(G_1 G_2^T G_2) + tr(G_2 G_1^T G_1) + tr(G_2 G_3^T G_3) + tr(G_3 G_2^T G_2) \\
&= 2\sum_{v=1}^{V} \alpha_v (tr(D_{1v}^T - R_{1v}) G_1) \\
&= 2\sum_{v=1}^{V} \beta_v (tr(D_{2v}^T - R_{2v}) G_2) \\
&= 2\sum_{v=1}^{V} \gamma_v (tr(D_{3v}^T - R_{3v}) G_3)
\end{align*}
\]

The partial derivatives of \( Z(G_1, G_2, G_3) \) with respect to \( G_1, G_2 \) and \( G_3 \) are:

\[
\begin{align*}
\frac{\partial Z(G_1, G_2, G_3)}{\partial G_1} &= -2R_{12} G_2 - 2R_{13} G_3 + 2G_1 G_2^T G_2 \\
&= +2G_1 G_2^T G_2^+ + \sum_{v=1}^{V} \alpha_v (D_{1v}^T - R_{1v}) G_1 \\
\frac{\partial Z(G_1, G_2, G_3)}{\partial G_2} &= -2R_{12}^T G_1 - 2R_{23}^T G_3 + 2G_2 G_2^T G_2 \\
&= +2R_{23} G_3^T G_3^+ + \sum_{v=1}^{V} \beta_v (D_{2v}^T - R_{2v}) G_2 \\
\frac{\partial Z(G_1, G_2, G_3)}{\partial G_3} &= -2R_{13}^T G_1 - 2R_{12}^T G_2 + 2G_3 G_3^T G_3 \\
&= +2G_3 G_3^T G_3^+ + \sum_{v=1}^{V} \gamma_v (D_{3v}^T - R_{3v}) G_3 \\
\end{align*}
\]

We can then use the Karush-Kuhn-Tucker (KKT) conditions (Boyd and Vandenberghe 2004) for the nonnegativity of \( G_1, G_2, G_3 \):

\[
\begin{align*}
-\sum_{v=1}^{V} \alpha_v R_{1v} G_1 &= 0 \\
-\sum_{v=1}^{V} \beta_v R_{2v} G_2 &= 0 \\
-\sum_{v=1}^{V} \gamma_v R_{3v} G_3 &= 0
\end{align*}
\]

These nonnegative constraints give the fixed point relationship that the solution must satisfy. As such, we can update \( G_1, G_2, G_3 \) using the following iterative rules:

\[
\begin{align*}
[G_1]_{ij} &= (R_{12} G_2 + R_{13} G_3 + \sum_{v=1}^{V} \alpha_v R_{1v} G_1) / (G_2 G_2^T G_2 + G_3 G_3^T G_3 + \sum_{v=1}^{V} \alpha_v D_{1v} G_1) \\
[G_2]_{ij} &= (R_{23} G_3 + R_{23}^T G_3 + \sum_{v=1}^{V} \beta_v R_{2v} G_2) / (G_2 G_2^T G_2 + R_{23}^T G_3 + \sum_{v=1}^{V} \beta_v D_{2v} G_2) \\
[G_3]_{ij} &= (R_{13} G_1 + R_{13}^T G_1 + \sum_{v=1}^{V} \gamma_v R_{3v} G_3) / (G_1 G_1^T G_1 + R_{13}^T G_3 + \sum_{v=1}^{V} \gamma_v D_{3v} G_3)
\end{align*}
\]

We can iteratively update \( G_1, G_2, G_3 \) by using Eq. (8), Eq. (9) and Eq. (10) until convergence. Although the third term of Eq. (3) in the manuscript is not jointly convex for \( G_1, G_2, G_3 \), our empirical study shows that \( G_1, G_2, G_3 \) often converge in less than 20 iterations.

After updating \( G_1 \) and \( G_2 \), we view them as known and take the partial derivative of \( \hat{MR}(G) \) with respect to \( \alpha_v \) and \( \beta_v \). In this case, the second, the third and the fifth terms on the right of Eq. (3) are irrelevant to \( \alpha_v \), and can be ignored. Then we can obtain:

\[
\hat{MR}(G_1, \alpha) = \sum_{v=1}^{V} \alpha_v (D_{1v}^T - R_{1v}) G_1 + \lambda_1 \| \alpha \|_F^2
\]

subject to \( \alpha_v \geq 0, \sum_{v=1}^{V} \alpha_v = 1 \).
Let \( H_v = tr(G_1^T(D_{11}^v - R_{11}^v)G_1) \) be the smoothness loss on the \( v \)-th bag view for \( G_1 \), then Eq. (11) can be update as:

\[
\tilde{MR}(H, \alpha) = vec(\alpha)^T vec(H) + \lambda_1 vec(\alpha)^T vec(\alpha) \\
\text{s.t. } \alpha_v \geq 0, vec(\alpha) = 1.
\]

Eq. (12) is a quadratic optimization problem with respect to \( vec(\alpha) \). By introducing the Lagrangian multipliers \( (\gamma, \mu) \) for the constraints of \( \alpha \), Eq. (12) is formulated as:

\[
\tilde{MR}(H, \alpha, \gamma, \mu) = vec(\alpha)^T vec(H) + \lambda_1 vec(\alpha)^T vec(\alpha) \\
- \sum_{v=1}^{V} \gamma_v \alpha_v - \mu(\sum_{v=1}^{V} \alpha_v - 1) \\
\text{s.t. } \alpha_v \geq 0, vec(\alpha) = 1.
\]

Base on the KKT conditions, the optional \( \alpha \) should satisfy the following four conditions:

(i) Stationary condition: \( \frac{\partial \tilde{MR}}{\partial \alpha} = H + 2\lambda_1 \alpha - \gamma - \mu = 0 \)

(ii) Feasible condition: \( \alpha_v \geq 0, \sum_{v=1}^{V} \alpha_v - 1 = 0 \)

(iii) Dual feasibility: \( \gamma_v \geq 0 \)

(iv) Complementary slackness: \( \gamma_v \alpha_v = 0 \)

From the stationary condition, \( \alpha_v \) can be computed as follows:

\[
\alpha_v = \frac{\gamma_v + \mu - H_v}{2\lambda_1}.
\]

We can find that \( \alpha_v \) depends on the specification of \( \gamma_v \) and \( \mu \), and the specification of \( \gamma_v \) and \( \mu \) can be analyzed in the following cases:

(i) If \( \mu > H_v \), then \( \alpha_v > 0 \); because of the complementary slackness \( \gamma_v \alpha_v = 0, \gamma_v = 0 \) and \( \alpha_v = \frac{\mu - H_v}{2\lambda_1} \), then \( \gamma_v = 0 \) and \( \alpha_v = 0 \).

(ii) If \( \mu < H_v \), since \( \alpha_v \geq 0 \), it requires \( \gamma_v > 0 \); because \( \gamma_v \alpha_v = 0 \), then \( \alpha_v = 0 \).

From the above analysis, we can set \( \alpha_v \) as:

\[
\alpha_v = \begin{cases} 
\frac{\mu - H_v}{2\lambda_1} & \text{if } \mu > H_v \\
0 & \text{if } \mu \leq H_v
\end{cases},
\]

Let \( v_H \) store the entries of vector \( vec(H) \) in ascending order with entries corresponding to \( tr(G_1^T(D_{11}^v - R_{11}^v)G_1) \). Accordingly, \( v^a \) stores the corresponding entries of \( vec(\alpha) \) with entries corresponding to \( tr(G_1^T(D_{11}^v - R_{11}^v)G_1) \). For a not too big predefined \( \lambda_1 \), there exists \( h \in \{1, 2, \ldots, V\} \) with \( v_H(h) < \mu \) and \( v_H(h + 1) \geq \mu \), satisfying \( \sum_{v_H(h) < \mu} \frac{\mu - v_H(h)}{2\lambda_1} = 1 \). Then \( v^a(h') \) has the following explicit solution:

\[
v^a(h') = \begin{cases} 
\frac{\mu - v_H(h')}{2\lambda_1} & \text{if } h' \leq h \\
0 & \text{if } h' > h
\end{cases},
\]

From \( \sum_{h'=1}^{V} v^a(h') = \sum_{h'=1}^{h} \frac{\mu - v_H(h')}{2\lambda_1} = 1 \), we can get the value for \( \mu \) as:

\[
\mu = \frac{2\lambda_1 + \sum_{h'=1}^{h} v_H(h')}{h}.
\]

To find an appropriate \( h \) that satisfies \( \mu - v_H(h) > 0 \) and \( \mu - v_H(h + 1) \leq 0 \), we decrease \( h \) from \( V \) to 1 step by step, and list the procedure in Algorithm 1. In this way, we obtain the explicit solution of \( \alpha_h \).

**Algorithm 1** A method to seek \( h \) and compute \( \alpha_v \)

**Input:** \( v_H, \lambda_1 \).

**Output:** \( h, \alpha_v \).

1: Initialize \( h = V, \mu = 0 \)
2: While \( h > 0 \)
3: Update \( \mu \) using Eq. (17)
4: If \( \mu - v_H(h) > 0 \)
5: break;
6: \( h \leftarrow h - 1; \)
7: \( \alpha_v \leftarrow \frac{\mu - v_H(v)}{2\lambda_1} \) for \( v = 1, \ldots, h \)
8: \( \alpha_v \leftarrow 0 \) for \( v = h + 1, \ldots, V \)
9: Return \( h \) and \( \alpha_v \)

As to \( \beta \), we find the first, the third and the fourth terms on the right of Eq. (3) in the manuscript are irrelevant to \( \beta \), then we can seek \( \beta \) by minimizing the following objective:

\[
\tilde{MR}(G_2, \beta) = \sum_{v=1}^{V} \beta_v tr(G_1^T(D_{22} - R_{22})G_2) + \lambda_2 ||\beta||_F^2
\]

The explicit solution of \( \beta_v \) can be obtained following the similar procedure of \( \alpha_v \). In this way, we can alternatively optimize \( G_1, G_2, G_3, \alpha \) and \( \beta \).

**Time complexity of M3Lcmf**

Let \( n, m, c, V \) be the number of bags, instances, labels, and views, respectively. Suppose \( d \) is the decomposed dimension of \( G \), and \( t \) is the maximum number of iterations in the optimization process. The time complexity of optimizing \( G_1, G_2, G_3, \alpha \) and \( \beta \) is \( O(tmdn(d + 1)), O(tm(2n^2 + nm + md + 2dc)), O(2tc(n^2 + nm + md)) \) and \( 2O(VT) \), respectively. Thus the time complexity of M3Lcmf is \( O(t(mn(d^2 + d + 2c + 2n + m) + d(4c + m)) + 2n^2c + 2V)) \).

**References**
