A NMF-framework for Unifying Posterior Probabilistic Clustering and Probabilistic Latent Semantic Indexing

Zhong-Yuan Zhang
School of Statistics
Central University of Finance and Economics, P.R.China
zhyuanzh@gmail.com

Tao Li
School of Computing and Information Sciences
Florida International University, U.S.A.

Chris Ding
Department of Computer Science and Engineering
University of Texas at Arlington, U.S.A.

Jie Tang
Department of Computer Science and Technology
Tsinghua University, P.R.China.

Key Words: Posterior Probabilistic Clustering; Probabilistic Latent Semantic Indexing; NMF-framework

ABSTRACT

In document clustering, a document may be assigned to multiple clusters and the probabilities of a document belonging to different clusters are directly normalized. We propose a new Posterior Probabilistic Clustering (PPC) model that has this normalization property. The clustering model is based on Nonnegative Matrix Factorization (NMF) and flexible such that if we use class conditional probability normalization, the model reduces to Probabilistic Latent Semantic Indexing (PLSI) . Systematic comparison and evaluation indicates that PPC is competitive with other state-of-art clustering methods. Furthermore, the results of

PPC are more sparse and orthogonal, both of which are highly desirable.

APPENDIX

1. An Illustration

We use a small example to further illustrate the uncertainty nature of NMF. The term-document data matrix is:

$$X = \begin{pmatrix} 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix},$$

each column of which is a document that is characterized by six terms. An optimal solution (F^*, G^*) is:

$$F^* = \begin{pmatrix} 1.05 & 0.42 \\ 0.48 & 0.70 \\ 0.64 & 0.00 \\ 1.04 & 0.00 \\ 0.00 & 1.22 \\ 0.46 & 1.27 \end{pmatrix}; \quad G^* = \begin{pmatrix} 0.69 & 0.00 \\ 1.10 & 0.24 \\ 0.00 & 0.85 \\ 0.18 & 0.71 \end{pmatrix}.$$

Hence, according to G^*, the first two documents are clustered into one group and the last two documents are clustered into the other one. Consequently we can choose a special diagonal matrix

$$D = \begin{pmatrix} 1 & 0 \\ 0 & 6 \end{pmatrix}.$$

such that

$$F^{**}=F^{*}D=\begin{pmatrix} 1.05 & 2.53 \\ 0.48 & 4.23 \\ 0.64 & 0.00 \\ 1.04 & 0.00 \\ 0.00 & 7.33 \\ 0.46 & 7.60 \end{pmatrix}; \quad G^{**}=G^{*}D^{-1}\begin{pmatrix} 0.69 & 0.00 \\ 1.10 & 0.04 \\ 0.00 & 0.14 \\ 0.18 & 0.12 \end{pmatrix}.$$

is also an optimal solution. According to G^{**}, the third document is clustered into the one group while the other three documents are clustered into another group. One can clearly observe that the cluster assignments on documents have been changed (e.g., the fourth document).

On the other hand, the result of PPC clustering is as follows:

$$\tilde{F}=\begin{pmatrix} 1.08 & 0.37 \\ 0.41 & 0.60 \\ 0.49 & 0.00 \\ 1.02 & 0.00 \\ 0.00 & 1.10 \\ 0.46 & 1.08 \end{pmatrix}; \quad \tilde{G}=\begin{pmatrix} 1.00 & 0.00 \\ 0.92 & 0.08 \\ 0.00 & 1.00 \\ 0.21 & 0.79 \end{pmatrix}.$$

Note that each row of $\tilde{G}$ is summed up to 1 and the posterior probabilities of cluster assignments are directly enforced. As a consequence, the clustering results on $\tilde{G}$ has a very clear and rigorous probabilistic interpretation.

2. A further insight into the PPC model

Here we note that (**See § 3**) the class conditional probability normalization of Eq.(5) is easy to satisfy, because Theorem 1 asserts that for any optimal solution (C^*, S^*, G^*) we can choose D_1, D_2 in Eq.(8) to obtain an equivalent optimal solution (C^{**}, S^{**}, G^{**}) that satisfies the class conditional probability normalization. Unfortunately, the posterior probability normalization is harder to achieve:

Theorem 5 *Suppose the input X is normalized. For a given optimal solution (C^*, S^*, G^*), in general, there exists no equivalent optimal solution (C^{**}, S^{**}, G^{**}) which satisfies the posterior probability normalization of Eq.(6). In particular, there exists no $(d_1, \cdots, d_K)$ such the posterior probability normalization of Eq.(2) is satisfied for all j.*

Proof: We prove the simple case of the normalization in Eq.(2). There are n constraints (equations) and K variables $(d_1, \cdots, d_K)$. In general, K is much less than n. Therefore, these is no solution to this linear equation system. $\square$.

For this reason, posterior probabilistic normalization needs to be imposed as a constraint in the optimization.

3. Proof of Theorem 3

We only prove the convergence of update rule Eq.(15) by constructing an auxiliary function of $L(F, G)$. The convergence of Eq (16) is clearly proved in [5]. If we regard the penalty function L as a function of G, it can be rewritten as:

$$L(G) = \|X - FG^T\|_F^2 + \alpha\|GI - I\|_F^2,$$

where I is a vector of ones whose dimension is omitted since it is clear from the context. It is easy to show that the following function $Z(G, G')$ is an auxiliary function of $L(G)$.

$$\begin{aligned} Z(G, G') = & \sum_{i,j} X_{ij}^2 + \sum_{i,j,a} \frac{(FG'^T)_{ij}}{F_{ia}G'_{ja}} F_{ia}^2 G_{ja}^2 - 2\sum_{i,j,a} F_{ia}G_{ja}X_{ij} \\ & + \alpha n + \alpha \sum_{j,k} \frac{\sum_k G'_{jk}}{G'_{jk}} G_{jk}^2 - 2\alpha \sum_{j,k} G_{jk} \end{aligned}$$

The minimum of $Z(G, G')$ with respect to G is give by

$$\begin{aligned} \frac{\partial Z(G, G')}{\partial G_{ml}} = & 2\sum_i \frac{(FG'^T)_{im}}{G'_{ml}} F_{il}G_{ml} - 2\sum_i F_{il}X_{im} + \\ & + 2\alpha \frac{\sum_k G'_{mk}}{G'_{ml}} G_{ml} - 2\alpha = 0. \end{aligned}$$

since $Z(G, G')$ is convex with respect to G. Solving for G_{ml}, the minimum is:

$$G_{ml} := G_{ml} \frac{(X^T F)_{ml} + \alpha}{(GF^T F)_{ml} + \alpha \sum_k G_{mk}}.$$

which gives the update rule (15).

4. Proof of Theorem 4

The proof of Theorem 4 is based on the following lemma.

Lemma 6 *If $\alpha_1 > \alpha_2$, then $J(G(\alpha_1)) > J((\alpha_2))$, where $G(\alpha_i)$ is the optimal solution of $L_{\alpha_i} = \|X - FG^T\|_F^2 + \alpha_i \|GI - I\|_F^2, \quad i = 1, 2.$*

Proof of Lemma 6: We have

$$L_{\alpha_1}(G(\alpha_2)) \geqslant L_{\alpha_1}(G(\alpha_1)) \tag{20}$$

$$L_{\alpha_2}(G(\alpha_1)) \geqslant L_{\alpha_2}(G(\alpha_2)). \tag{21}$$

Hence

$$\begin{aligned} 0 &\leqslant L_{\alpha_1}(G(\alpha_2)) + L_{\alpha_2}(G(\alpha_1)) - [L_{\alpha_1}(G(\alpha_1)) + L_{\alpha_2}(G(\alpha_2))] \\ &= (\alpha_1 - \alpha_2)(\|G(\alpha_2)I - I\|_F^2 - \|G(\alpha_1)I - I\|_F^2). \end{aligned}$$

Remembering that $\alpha_1 > \alpha_2$ and using (20) and (21), one has

$$\begin{aligned} J(G(\alpha_2)) &\leqslant J(G(\alpha_2)) + \alpha_1 [\|G(\alpha_2)I - I\|_F^2 - \|G(\alpha_1)I - I\|_F^2] \\ &\leqslant J(G(\alpha_1)). \end{aligned}$$

Now, we are ready to prove Theorem 4.

Proof: If this were not the case, one has $\alpha \to +\infty$ and

$$\|G^{(k)}I - I\|_F^2 \geqslant \varepsilon \tag{22}$$

for all k, where $G^{(k)}$ is the result after the kth iteration. But obviously there must exists $\tilde{G} \geqslant 0$ such that

$$\|\tilde{G}I - I\|_F^2 = 0 < \varepsilon/2, \tag{23}$$

hence

$$\begin{aligned}\|X-F\tilde{G}^T\|_F^2+\alpha\|\tilde{G}I-I\|_F^2 &\geqslant \|X-F{G^{(k)}}^T\|_F^2+\alpha\|G^{(k)}I-I\|_F^2\\ &\geqslant \|X-F{G^{(1)}}^T\|_F^2+\alpha\|G^{(k)}I-I\|_F^2,\end{aligned}$$

in other words,

$$\|\tilde{G}I-I\|_F^2-\|G^{(k)}I-I\|_F^2\geqslant\tfrac{1}{\alpha}(\|X-F{G^{(1)}}^T\|_F^2-\|X-F\tilde{G}^T\|_F^2)\to 0$$

since $\alpha\to+\infty$, which contradicts with (22) and (23). The proof is thus completed.