Posterior Probabilistic Clustering using NMF

Chris Ding
CSE Department
University of Texas at Arlington
chqding@uta.edu

Tao Li
School of Computer Science
Florida International. University
taoli@cs.fiu.edu

Dijun Luo
CSE Department
University of Texas at Arlington
dijun.luo@gmail.com

Wei Peng
School of Computer Science
Florida International University
wpeng002@cs.fiu.edu

ABSTRACT

We introduce the posterior probabilistic clustering (PPC), which provides a rigorous posterior probability interpretation for Nonnegative Matrix Factorization (NMF) and removes the uncertainty in clustering assignment. Furthermore, PPC is closely related to probabilistic latent semantic indexing (PLSI).

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Clustering; I.2 [**Artificial Intelligence**]: Learning

General Terms

Algorithms, Experimentation, Measurement, Performance, Theory

Keywords

Sparse, Posterior Probabilistic Clustering, NMF

1. INTRODUCTION

Non-negative Matrix Factorization (NMF) [4] has been successfully applied to document clustering recently [5, 1]. However, in the standard NMF clustering, cluster assignment is rather *ad hoc*. In addition, matrix factors lack clear interpretations.

In this work, we introduce the posterior probabilistic clustering (PPC), which has 3 benefits: (1) It provides a rigorous posterior probability interpretation for both matrix factors F, G in the factorization of input X: $X \simeq FG^T$. (2) It removes the uncertainty in clustering assignment. (3) Furthermore, when we perform simultaneous word and document clustering, the new model has a very close relation to probabilistic latent semantic indexing (PLSI) [3]: in PLSI, F, G are class conditional probabilities; in PPC, F, G are class posterior probabilities.

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

2. STANDARD NMF CLUSTERING

Suppose we have n documents and m words (terms). Let $X = (X_{ij})$ be the word-to-document matrix: $X_{ij} = X(w_i, d_j)$ is the frequency of word w_i in document d_j. Standard NMF optimization is

$$\min_{F\geq 0, G\geq 0} \|X - FG^T\|^2, \quad (1)$$

where X has size $m \times n$, F has size $m \times K$, G has size $n \times K$. Once the solution (F^*, G^*) is computed, standard approach is to assign d_j to the cluster C_k where

$$k = \arg\max(G^*_{j1}, \cdots, G^*_{jK}), \quad (2)$$

i.e., the largest element of j-th row of G.

There is a fundamental problem with this approach. First, the solution to NMF is not unique. For an arbitrary positive diagonal matrix $D = \text{diag}(d_1, \cdots, d_k)$, we have

$$F^* G^{*T} = (F^* D^{-1})(G^* D)^T$$

i.e., $(F^* D^{-1}, G^* D)$ is also an optimal solution. Thus the cluster assignment is modified to

$$k = \arg\max(G^*_{j1} d_1, \cdots, G^*_{jk} d_K). \quad (3)$$

A different choice of D leads to different cluster assignment. An *ad hoc* solution is to choose D such that columns of F have unit length in L_2 norm.

3. POSTERIOR PROBABILITY

In this work, we present a principled way to resolve this problem. This is based on posterior probability interpretation of G. In fact, we can see from Eq.(3) that (roughly speaking)

$$(G^*_{j1} d_1, \cdots, G^*_{jk} d_K),$$

is the posterior probability that d_j belongs to different clusters. Thus we wish to choose D such that

$$G^*_{j1} d_1 + \cdots + G^*_{jk} d_K = 1, \;\; j = 1, \cdots, n$$

This requirement has no solution, because there are n constraints and K variables, but K is much less n. Therefore, in standard NMF, there is no way to enforce posterior probability normalization.

4. POSTERIOR PROBABILISTIC CLUSTERING

In our approach, we enforce the posterior probability normalization directly. The posterior probabilistic clustering is to optimize

$$\min_{F\geq 0, G\geq 0} \|X - FG^T\|^2, \quad s.t. \quad \sum_{k=1}^{K} G_{jk} = 1, \tag{4}$$

Using Lagrangian multipliers to enforce the constraints, we derive the following updating rules to solve this problem

$$G_{ik} \leftarrow G_{ik} \frac{(X^T F)_{ik} + (GF^T FG^T)_{ii}}{(GF^T F)_{ik} + (X^T FG^T)_{ii}} \tag{5}$$

$$F_{ik} \leftarrow F_{ik} \frac{(XG^T)_{ik}}{(FG^T G)_{ik}} \tag{6}$$

The correctness and convergence can be proved rigorously. In the updating process, the constraints should be enforced periodically.

5. SIMULTANEOUS WORD AND DOCUMENT CLUSTERING (SPPC)

We generalize PPC to simultaneous word and document clustering. We use F as the posterior probability for word clustering, and the posterior probability normalization is $\sum_{k=1}^{K} F_{ik} = 1$. The simultaneous PPC (SPPC) problem becomes

$$\min_{F\geq 0, S, G\geq 0} \|X - FSG^T\|^2, \quad s.t. \quad \sum_{k=1}^{K} F_{ik} = 1, \sum_{k=1}^{K} G_{jk} = 1, \tag{7}$$

We derived the updating algorithm as follows. Let $\widetilde{F} = FS$, $\widetilde{G} = GS^T$, the updating algorithm is

$$G_{ik} \leftarrow G_{ik} \frac{(X^T \widetilde{F})_{ik} + (G\widetilde{F}^T \widetilde{F} G^T)_{ii}}{(G\widetilde{F}^T \widetilde{F})_{ik} + (X^T \widetilde{F} G^T)_{ii}} \tag{8}$$

$$F_{ik} \leftarrow F_{ik} \frac{(X\widetilde{G}^T)_{ik} + (F\widetilde{G}^T \widetilde{G} F^T)_{ii}}{(F\widetilde{G}^T \widetilde{G})_{ik} + (X\widetilde{G} F^T)_{ii}} \tag{9}$$

$$S_{kk'} \leftarrow S_{kk'} \frac{(F^T XG)_{kk'}}{(F^T FSG^T G)_{kk'}} \tag{10}$$

We initialize F, G to the K-means clustering results on words F_0 and on documents G_0. where F_0, G_0 are cluster indicators. We set $F = F_0 + 0.2$ and $G = G_0 + 0.2$.

5.1 Relation to PLSI

In PLSI we view the word-document matrix X as the joint probability of word and documents. [We re-scale the term frequency X_{ij} by $X_{ij} \leftarrow X_{ij}/T_w$, where $T_w = \sum_{ij} X_{ij}$. With this, $\sum_{ij} X_{ij} = 1$.] The joint occurrence probability is $p(w_i, d_j) = X_{ij}$. PLSI decompose it as product of class-conditional probabilities:

$$X_{ij} \approx \sum_k P(word_i|class_k)P(class_k)P(doc_j|class_k).$$

Let $F_{ik} = P(word_i|class_k), S_{kk} = P(class_k), G_{jk} = P(doc_j|class_k)$, PLSI optimization problem is:

$$\min_{F\geq 0, S, G\geq 0} Dist(X, FSG^T), \quad s.t. \quad \sum_{i=1}^{m} F_{ik} = 1, \sum_{j=1}^{n} G_{ik} = 1, \tag{11}$$

Therefore, our SPPC is quite similar to PLSI, except SPPC has a different normalization $\sum_{k=1}^{K} G_{jk} = 1, \sum_{k=1}^{K} F_{ik} = 1$. In other words, SPPC treats G_{jk}, F_{ik} as posterior probabilities; PLSI treats G_{jk}, F_{ik} as class-conditional probabilities.

Note that in PLSI, the sum of probabilities of a document belong to different classes, $\sum_{k=1}^{K} G_{jk} = P(doc_j|class_k) \neq 1$. Intuitively for clustering, we would like the total probability adds up to 1. This deficiency is removed in SPPC.

5.2 An Illustrative example

We give a simple example to illustrate the PPC and SPPC results. The data matrix is given bellow. From inspection, first 3 columns belong to one cluster and the last 4 columns belong to another. For rows, first 3 rows belong to one cluster and the last 2 row belong to another. The resulting F, G recover the clustering correctly.

$$X = \begin{pmatrix} 0.185 & 0.326 & 0.761 & 2.799 & 2.375 & 2.970 & 2.585 \\ 0.508 & 0.380 & 0.884 & 2.134 & 2.374 & 2.342 & 2.524 \\ 0.452 & 0.887 & 0.457 & 2.065 & 2.484 & 2.253 & 2.163 \\ 1.486 & 1.843 & 1.858 & 0.566 & 0.103 & 0.417 & 0.269 \\ 1.496 & 1.806 & 1.610 & 0.612 & 0.158 & 0.560 & 0.784 \end{pmatrix}$$

$$F_{\text{PPC}}^T = \begin{pmatrix} 0.068 & 0.059 & 0.056 & 0.000 & 0.005 \\ 0.007 & 0.011 & 0.012 & 0.037 & 0.035 \end{pmatrix}$$

$$G_{\text{PPC}}^T = \begin{pmatrix} 0.003 & 0.001 & 0.077 & 0.763 & 0.835 & 0.836 & 0.795 \\ 0.997 & 0.999 & 0.923 & 0.237 & 0.165 & 0.164 & 0.205 \end{pmatrix}$$

$$F_{\text{SPPC}}^T = \begin{pmatrix} 0.961 & 0.830 & 0.792 & 0.019 & 0.091 \\ 0.039 & 0.170 & 0.208 & 0.981 & 0.909 \end{pmatrix}$$

$$G_{\text{SPPC}}^T = \begin{pmatrix} 0.033 & 0.031 & 0.113 & 0.846 & 0.922 & 0.924 & 0.880 \\ 0.967 & 0.969 & 0.887 & 0.154 & 0.078 & 0.076 & 0.120 \end{pmatrix}$$

6. EXPERIMENTS

We compare the clustering performance of each method on 5 real-life datasets. More details of these datasets can be found in [2]. We use accuracy as the performance measure. The experimental results are shown in Table 1. We see that SPPC performs slightly better than NMF and PLSI.

Datasets	K-Means	NMF	PLSI	SPPC
CSTR	0.4256	0.5713	0.587	0.5945
WebKB4	0.3888	0.4418	0.503	0.4411
Log	0.6876	0.7805	0.778	0.7915
Reuters	0.4448	0.4947	0.4870	0.5648
WebAce	0.4001	0.4761	0.4890	0.4953

Table 1: Clustering Results. Shown are the accuracy results for different methods.

7. REFERENCES

[1] C. Ding, X. He, and H.D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. *Proc. SIAM Data Mining Conf*, 2005.

[2] C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *KDD'06*, pages 126–135, 2006.

[3] T. Hofmann. Probabilistic latent semantic analysis. In *ACM SIGIR-99*, pages 289–296, 1999.

[4] D.D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, volume 13.

[5] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proc. ACM conf. Research and development in IR(SIGIR)*, pages 267–273, Toronto, Canada, 2003.