Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics and Data Analysis 52 (2008) 3913–3927
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

On the equivalence between Non-negative Matrix Factorization and Probabilistic Latent Semantic Indexing

Chris Ding[a], Tao Li[b,*], Wei Peng[b]

[a] *Department of CSE, University of Texas at Arlington, Arlington, TX 76019, United States*
[b] *School of Computer Science, Florida International University, Miami, FL 33199, United States*

Received 30 September 2007; received in revised form 16 January 2008; accepted 18 January 2008
Available online 2 February 2008

Abstract

Non-negative Matrix Factorization (NMF) and Probabilistic Latent Semantic Indexing (PLSI) have been successfully applied to document clustering recently. In this paper, we show that PLSI and NMF (with the I-divergence objective function) optimize the same objective function, although PLSI and NMF are different algorithms as verified by experiments. This provides a theoretical basis for a new hybrid method that runs PLSI and NMF alternatively, each jumping out of the local minima of the other method successively, thus achieving a better final solution. Extensive experiments on five real-life datasets show relations between NMF and PLSI, and indicate that the hybrid method leads to significant improvements over NMF-only or PLSI-only methods. We also show that at first-order approximation, NMF is identical to the χ^2-statistic.

1. Introduction

Document clustering has been widely used as a fundamental and effective tool for efficient document organization, summarization, navigation, and retrieval of large number of documents. Generally document clustering problems are determined by three basic tightly-coupled components: (a) the (physical) representation of the given dataset; (b) the criterion/objective function which the clustering solutions should aim to optimize; and (c) the optimization procedure (Li, 2005).

Among clustering methods, the K-means algorithm has been the most popularly used. A recent development is the Probabilistic Latent Semantic Indexing (PLSI). PLSI is an unsupervised learning method based on statistical latent class models and has been successfully applied to document clustering (Hofmann, 1999). (PLSI has been further developed into a more comprehensive Latent Dirichlet Allocation model (Blei et al., 2003).)

Non-negative Matrix Factorization (NMF) is another recent development in document clustering. The initial work on NMF (Lee and Seung, 1999, 2001) emphasizes that the NMF factors contain coherent parts of the original data (images). Later works (Xu et al., 2003; Pauca et al., 2004) show the usefulness of NMF for clustering with experiments

* Corresponding author.
E-mail address: taoli@cs.fiu.edu (T. Li).

0167-9473/$ - see front matter © 2008 Published by Elsevier B.V.
doi:10.1016/j.csda.2008.01.011

on document collections, and a recent theoretical analysis (Ding et al., 2005) shows the equivalence between NMF and K-means /spectral clustering.

Despite significant research on both NMF and PLSI, few attempts have been made to establish the connections between them while highlighting their differences in the clustering framework. Gaussier and Goutte (2005) made the initial connection between NMF and PLSI, by showing that the iterative update procedures of PLSI and NMF are similar in that the fixed-point equations for the converged solutions are the same. However, we emphasize that NMF and PLSI are different algorithms: they converge to different solutions, even if they start from the same initial condition, as verified by experiments (see later sections).

In this paper, we first show that both NMF (with I-divergence objective) and PLSI optimize the same objective function. This fundamental fact and the L_1-normalization NMF ensure that NMF and PLSI are equivalent. In other words, PLSI is equivalent to NMF with I-divergence objective.

Second, we show, by an example and extensive experiments, that NMF and PLSI are different algorithms and they converge to different local minima. This leads to a new insight: NMF and PLSI are different algorithms for optimizing the same objective function.

Third, we give a detailed analysis about the NMF and PLSI solutions. They are local minima of the same landscape in a very high-dimensional space. We show that PLSI can jump out of the local minima where NMF converges to and vice versa. Based on this, we further propose a hybrid algorithm to run NMF and PLSI alternatively to jump out of a series of local minima and finally reach a much better minimum. Extensive experiments show this hybrid algorithm improves significantly over the standard NMF-only or PLSI-only algorithms.

A preliminary version of this paper appeared in Ding et al. (2006). More theoretical analysis and experiments are included in the journal version. The rest of the paper is organized as follows: Section 2 introduces the data representations of NMF and PLSI, Section 3 presents the equivalence between NMF and PLSI, Section 4 shows that the column normalized NMF is equivalent to the probability factorization, Section 5 uses an example to illustrate the difference between NMF and PLSI, Section 6 gives the empirical comparison results between NMF and PLSI, Section 7 proposes a hybrid algorithm to run NMF and PLSI alternatively and finally Section 8 concludes.

2. Data representations of NMF and PLSI

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

In this paper, we re-scale the term frequency F_{ij} by $F_{ij} \leftarrow F_{ij}/T_w$, where $T_w = \sum_{ij} F_{ij}$ is the total number of words. With this stochastic normalization, $\sum_{ij} F_{ij} = 1$. The joint occurrence probability $p(w_i, d_j) = F_{ij}$.

The general form of NMF is

$$F = CH^{\mathrm{T}}, \tag{1}$$

where the matrices $C = (C_{ik})$, $H = (H_{jk})$ are non-negative matrices. They are determined by minimizing

$$J_{\mathrm{NMF}} = \sum_{i=1}^{m}\sum_{j=1}^{n} F_{ij} \log \frac{F_{ij}}{(CH^{\mathrm{T}})_{ij}} - F_{ij} + (CH^{\mathrm{T}})_{ij}. \tag{2}$$

PLSI maximizes the likelihood

$$\max J_{\mathrm{PLSI}}, \qquad J_{\mathrm{PLSI}} = \sum_{i=1}^{m}\sum_{j=1}^{n} F_{ij} \log P(w_i, d_j) \tag{3}$$

where $P(w_i, d_j)$ is the factorized (i.e., parameterized or approximated) joint occurrence probability

$$\begin{aligned} P(w_i, d_j) &= \sum_k P(w_i, d_j|z_k)P(z_k) \\ &= \sum_k P(w_i|z_k)P(d_j|z_k)P(z_k), \end{aligned} \tag{4}$$