Pattern Recognition Letters 54 (2015) 97–102

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

A stable approach for model order selection in nonnegative matrix factorization[☆]

Meng Sun[a,*], Xiongwei Zhang[a], Hugo Van hamme[b]

[a] Lab of Intelligent Information Processing, College of Command Information Systems, PLA University of Science and Technology, Hai Fu Xiang 1, Nanjing 210007, China

[b] Department of Electrical Engineering (ESAT), Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, Leuven 3001, Belgium

ARTICLE INFO

Article history:
Received 21 March 2014
Available online 9 January 2015

Keywords:
Model order selection
Nonnegative matrix factorization
Minimum entropy regularization
Automatic relevance determination

ABSTRACT

In order to find the correct model order in non-negative matrix factorization (NMF), an algorithm called automatic relevance determination (ARD) is proposed in Tan and Fevotte (2013). The algorithm explores the similarities of the NMF components and removes redundant ones iteratively. However, the algorithm can yield over-parsimonious representations where ground truth patterns can be grouped into one single component to cause superposition. In this paper, mixed entropy regularized NMF (MER-NMF) is proposed to overcome the above problem. In MER-NMF, the objective function of NMF is regularized by minimizing a mixed entropy of the coefficient matrix which is a weighted sum of two parts: the entropy of all the entries and the entropy of the row sums of the coefficient matrix. With the mixed entropy regularization, the algorithm tends to yield sharper activations of the components for each sample. By combining MER-NMF and ARD-NMF, correct number of components can always be selected according to our experiments.

1. Introduction

In the last few years, nonnegative matrix factorization (NMF) has received a lot of attention in machine learning and data mining research [5]. The most attractive property of NMF is that it is able to decompose the input data matrix into nonnegative components, which enables it to find repeated "parts" from the "whole" dataset. Because the number of components is far less than the number of data samples, NMF actually acts as a dimension reduction tool. One of the open problems in NMF is how many components one should choose for the factorization, i.e. the model order selection problem. The problem is actually related to many other concurrent NMF research topics, such as sparse NMF [3], Bayesian NMF [1], and graph regularized NMF [11], as is explained below: (1) If an algorithm could correctly allocate the number of components in the data, the solution would also be optimal for sparse NMF, since fewer components would not be adequate to model the data. (2) With adequate Bayesian priors, NMF would produce solutions which reflect the underlying data structures and imply the correct number of components [1]. (3) In our previous work [11], we observed that with graph regularization NMF tends to be insensitive to the choice of the model order.

Techniques like Bayesian information criterion (BIC) [10] cannot be applied to the problem of NMF model order selection, because the number of parameters is assumed constant in BIC, while actually it scales linearly with the number of data points [12]. Markov chain Monte Carlo (MCMC) methods proposed in [1] and [6] calculate the evidence for each candidate value of the model order, and the model order with highest evidence value is selected. However, the sampling-based methods are computationally intensive [9,12,13]. As a by-product, model order selection can also be achieved by imposing sparsity in NMF as is studied in [7]. Bayesian NMF is applied in [2] with priors which would restrict the NMF algorithm towards better solutions. As the most recent state-of-the-art work, Tan and Fevotte proposed a novel algorithm for automatic relevance determination in NMF in [12] which outperforms the methods using sparse NMF in [7] and Bayesian NMF [2]. The detailed explanation will be given below.

1.1. Automatic relevance determination

In [12], the model order selection problem is converted into an automatic relevance determination (ARD) of the components to be learned. For the NMF with Kullback–Leibler divergence, the entries of each component and its corresponding activations in the coefficient matrix are assumed to be generated from a half-normal distribution with parameter λ. A large λ indicates a strong relevance of this component to model the data, while a small one implies a weak relevance. Furthermore, an inverse-Gamma prior is imposed on the hyper

[☆] This paper has been recommended for acceptance by A. Petrosino.

[*] Corresponding author. Tel.: +86 18602544808.
E-mail address: sunmengccjs@gmail.com (M. Sun).

http://dx.doi.org/10.1016/j.patrec.2015.01.002
0167-8655/© 2015 Elsevier B.V. All rights reserved.

parameter λ. The inverse-Gamma distribution has a shape parameter a and a scale parameter b. With the Bayesian formulation and by maximizing the log posterior, an elegant updating algorithm is derived and has manifested its good performance on three datasets.

As is reported in Section 4 of [12] and also according to our own experiences on this algorithm, smaller values of the shape parameter a typically produce better results, while larger values do not. However, a small a can be dangerous because the algorithm tends to yield over-parsimonious representations where some ground truth patterns can be grouped into one single component to lead to erroneous superposition. So as is concluded in [12], it is still an open problem to find a suitable a.

In this paper, we try to avoid the problem of looking for a proper shape parameter a by introducing entropy regularization for NMF.

1.2. Entropy regularization

In information theory, Shannon entropy is a measure of uncertainty in random variables. The entropy rate of a data source means the average number of bits per symbol needed to encode it. In the problem of model order selection in NMF, we hope to reduce the uncertainty when representing the data using the components. In other words, it is not expected that two or more components behave similarly, i.e. similar components should be suppressed. By doing this, it will reduce the average number of components per data sample for representation. Therefore, we can achieve this by minimizing the entropy of the activation coefficients of the components.

In this paper, NMF with Kullback–Leibler divergence is first briefly introduced where probabilistic constraints are imposed by using the ℓ_1 normalization. The normalization, on one hand endows the NMF model with a probabilistic interpretation, and on the other hand makes the definition of entropy reasonable. Two entropy terms are considered as regularization functions. The first one is the entropy of the coefficient matrix which reflects the flatness of the joint distribution of components and data samples. The second one is the entropy of the row sums of the coefficient matrix which governs the contributions of every component. The lower the two entropy values, the sparser the representations.

1.3. Outline of the paper

The remaining parts of the paper are listed as below. In Section 2, we present the mixed entropy regularized NMF and derive its algorithms. In Section 3, the performance of the algorithms are evaluated on Swimmer dataset and TIDIGITS speech dataset. Conclusions are drawn in Section 4.

2. Nonnegative matrix factorization with mixed entropy regularization

2.1. NMF with Kullback–Leibler divergence

Let V denote the nonnegative data matrix to be factorized, NMF with Kullback–Leiber divergence can be modeled as the following optimization problem.

$$\begin{aligned} \operatorname{argmin}_{W,H} \quad & \mathrm{KLD}(V||WH) \\ \text{s.t.} \quad & \textstyle\sum_i W_{i,k} = 1, \quad W_{i,k} \geq 0 \\ & \textstyle\sum_{k,j} H_{k,j} = 1, \quad H_{k,j} \geq 0. \end{aligned} \tag{1}$$

In Eq. (1), the input data matrix V is normalized by $V_{i,j} \leftarrow V_{i,j}/\sum_{i,j} V_{i,j}$. $\mathrm{KLD}(x||y) := \sum_i x_i \log x_i/y_i$ is the Kullback–Leiber divergence between two discrete distributions x and y. The columns of W is normalized such that $\sum_i W_{i,k} = 1$. The coefficient matrix H is normalized by $H_{i,j} \leftarrow H_{i,j}/\sum_{i,j} H_{i,j}$. The above normalization endows the NMF with a probabilistic interpretation. In the wordings of topic modeling, $V_{i,j}$ corresponds to the joint probability of observing term t_i in document d_j, i.e. $\Pr(t_i, d_j)$. $W_{i,k} = \Pr(t_i|z_k)$ is the conditional probability of term t_i given latent topic z_k. $H_{k,j}$ corresponds to the joint probability of observing topic z_k in document j, i.e. $\Pr(z_k, d_j)$.

2.2. Mixed entropy as a regularizer

In this section, we consider to minimize the entropy of the coefficient matrix H. For a multivariate random variable, low entropy states the preference for a few values, while a high entropy is realized by spreading the probability mass over many values. By minimizing entropy, we expect to “sharpen” the entries in the coefficient matrix H, such that the important components become salient while unimportant ones vanish.

Two types of entropy of H are constructed to regularize the original NMF objective function. The first ones is the entropy of the matrix in Eq. (2). Note that $\sum_{k,j} H_{k,j} = 1$ satisfies the probability rationale. Hence, $\mathcal{R}_1(H) \geq 0$.

$$\mathcal{R}_1(H) := -\sum_{k,j} H_{k,j} \log_2(H_{k,j}). \tag{2}$$

The second term is the entropy of a marginal distribution which shows the total activations of each component through the data.

$$\mathcal{R}_2(H) := -\sum_{k} S_k \log_2(S_k), \tag{3}$$

where $S_k = \sum_j H_{k,j}/\sum_{k,j} H_{k,j}$ refers to the row sums of H. Again it is easy to check that $\sum_k S_k = 1$ and $\mathcal{R}_2(H) \geq 0$.

By minimizing the above two terms, a compact H is expected in which only a few entries in some rows of H have positive activations. The joint cost function for the mixed entropy regularized NMF (MER-NMF) therefore becomes,

$$\mathcal{F} := \mathrm{KLD}(V||WH) + \alpha \mathcal{R}_1(H) + \beta \mathcal{R}_2(H), \tag{4}$$

where α and β are the regularization parameters. The optimization problem is thus,

$$\begin{aligned} \operatorname{argmin}_{W,H} \quad & \mathcal{F}(V||WH) \\ \text{s.t.} \quad & \textstyle\sum_i W_{i,k} = 1, \quad W_{i,k} \geq 0, \\ & \textstyle\sum_{k,j} H_{k,j} = 1, \quad H_{k,j} \geq 0. \end{aligned} \tag{5}$$

2.3. The derivation of the algorithm

The algorithm to solve problem (5) is listed in Table 1. The derivation of the updating of W is the same as in [5] and the normalization in step 3 of Table 1 will not affect the convergence as is explained in [11]. The objective function of H, $\mathcal{F}(H)$, is non-convex. It is difficult to derive updating rules by using an auxiliary function. Traditional optimization techniques can be implemented in this constrained optimization problem, such as gradient descent and active sets. However, multiplicative updates are simple to implement and having good properties of zero-locking and nonnegativity keeping. Therefore, multiplicative updating rules are adopted in this paper where an exponential parameter γ performs as a tunable step size. Let $\nabla_H^+ \mathcal{F}$ and $\nabla_H^- \mathcal{F}$ be the positive and negative part of the derivative with respect to H. The multiplicative update is as follows,

$$H \leftarrow H \odot \left(\frac{\nabla_H^-(\mathcal{F})}{\nabla_H^+(F)} \right)^{\gamma} \tag{6}$$

It is straightforward to see that H is left unchanged when $\nabla_H^+(\mathcal{F}) = \nabla_H^-(F)$, i.e. the gradient is zero. If the gradient is positive, i.e. $\nabla_H^+(\mathcal{F}) > \nabla_H^-(F)$, H will decrease and vice versa if the gradient is negative. γ is a step size parameter that potentially can be tuned to assist convergence. When $\gamma \to 0$ only very small steps in the negative gradient direction are taken. Thus, there is a one-to-one relation between fixed points of the multiplicative update rule and stationary points under gradient descend [8].