Sparse Non-negative Matrix Factorization Algorithm for Independent Feature Learning (Postprint)

Authors: Huang Weichun, Zhao Yang, Xiong Liyan

Date: 2019-01-28T00:00:00+00:00

Abstract

Improving the semantic independence of latent features while preserving the interpretability of Non-negative Matrix Factorization (NMF) is an open research problem. To prevent feature co-adaptation, we propose utilizing cosine similarity to reduce the correlation between latent features, thereby enhancing the independent feature learning capability of NMF. Furthermore, to achieve favorable sparsity in the decomposed matrices, we propose introducing an L_{2,1/2} sparsity constraint into the traditional NMF model, which enhances the algorithm' s local learning capability and robustness. Consequently, the semantic information within the latent features becomes more pronounced, and the representation of the latent space becomes more discriminative. Experimental results on document clustering using the fetch_20newsgroups dataset demonstrate that the proposed INMF algorithm outperforms traditional NMF, SNMF, and other algorithmic models across a series of evaluation metrics.

Full Text

Preamble

Sparse Non-negative Matrix Factorization Algorithm for Independent Feature Learning

Huang Weichun[1], Zhao Yang[1], , Xiong Liyan[2]
([1]School of Software; [2]School of Information Engineering, East China Jiaotong University, Nanchang 330013, China)

Abstract: Improving the semantic independence of latent features while maintaining the interpretability of Non-negative Matrix Factorization (NMF) is an open research problem. To prevent feature co-adaptation, we propose utilizing

cosine similarity to reduce correlation between latent features, thereby enhancing the independent feature learning capability of NMF. Furthermore, to achieve better sparsity in the factorized matrices, we introduce sparse constraints into the traditional NMF model, which enhances the algorithm' s local learning ability and robustness. Consequently, the semantic information in latent features becomes more pronounced, and the representation in the latent space becomes more discriminative. Experimental results on document clustering using the fetch_20newsgroups dataset demonstrate that the proposed INMF algorithm outperforms traditional NMF, SNMF, and other algorithms across a series of evaluation metrics.

Key words: non-negative matrix factorization; sparse; independent feature learning; cosine similarity

0 Introduction

The purpose of NMF is to decompose an original high-dimensional data matrix into two low-dimensional matrices, where the product of these two low-dimensional matrices approximates the original high-dimensional data matrix as closely as possible.

Let $X = [x_1, x_2, \dots, x_M] \in \mathbb{R}_+^{N\times M}$ be an original data matrix. NMF seeks to decompose X into a non-negative basis matrix $U = [u_1, u_2, \dots, u_K] \in \mathbb{R}_+^{N\times K}$ and a non-negative coefficient matrix $V = [v_1, v_2, \dots, v_K] \in \mathbb{R}_+^{K\times M}$, where K is the number of latent features. This can also be written in the equivalent vector form $x_j \approx \sum_{i=1}^{K} u_i v_{ij}$. Typically, $K < \min(M, N)$ is used for dimensionality reduction, where v_j represents the weight coefficient of the original data vector x_j in the columns of U. NMF decomposes the data matrix into a basic linear combination of vectors.

Many researchers have attempted to adopt different cost functions to measure NMF performance, with their focus being on finding factor matrices that minimize the loss function. Currently, numerous methods have been proposed to improve NMF, such as incremental non-negative matrix factorization based on Fisher discriminant analysis [1], multi-view joint non-negative matrix factorization based on Hessian regularization [2], and incremental learning algorithms for sparse constrained graph regularized non-negative matrix factorization [3]. Han et al. [4] proposed imposing L_1 and L_2 norm sparse constraints on the column vectors of the basis matrix during non-negative matrix factorization to first mine the low-dimensional data structure embedded in high-dimensional data and achieve low-dimensional representation, then applied the K-Means algorithm, which performs well in low-dimensional data clustering, to cluster the sparsely reduced data. Sun et al. [5] proposed a multi-constraint non-negative matrix factorization algorithm based on feature fusion, which considers not only label information and sparse constraints for a small number of known samples but also graph regularization, and fuses image features with different sparsity

degrees after decomposition to enhance clustering performance and effectiveness.

However, the inherent correlation between latent features makes traditional NMF algorithms lack discriminative power and increases the difficulty of minimizing the loss function. Moreover, traditional NMF algorithms do not fully utilize matrix sparsity and ignore useful information about correlations between different features.

Sparse feature selection aims to apply various sparse models to achieve feature selection and obtain sparse data representation. Many studies [15–17] have extended the L_p norm ($0 < p < 1$) to achieve better sparsity. In references [18, 19], Xu et al. concluded that when $p = 1/2$, the $L_{1/2}$ norm has the best sparsity. In reference [20], Nie et al. introduced joint $L_{2,1}$-norm minimization for loss functions and regularized feature selection. However, the $L_{2,1}$ norm does not have good sparsity because it is based on the L_1 norm. Recently, Wang et al. [21] proposed extending the $L_{2,p}$-matrix norm to select joint, sparser features, and this model has better robustness than the $L_{2,1}$ norm. Experiments have shown that when $p = 1/2$, the $L_{2,1/2}$-matrix norm achieves the best performance. Therefore, this paper applies the $L_{2,1/2}$-matrix norm model to NMF to achieve sparse constraints.

This paper proposes utilizing cosine similarity to reduce correlation between latent features, thereby improving the independent feature learning capability of NMF. Additionally, we introduce $L_{2,1/2}$ sparse constraints into INMF. Consequently, the semantic information in latent features becomes clearer, and the representation in the latent space becomes more discriminative. We compare these methods with several basic NMF models for document clustering and present experimental results comparing our algorithm with traditional algorithms on real datasets. The main contributions of this work can be summarized as follows:

a) The proposed INMF improves NMF by introducing cosine similarity to prevent feature co-adaptation;

b) Adding $L_{2,1/2}$-norm sparse constraints on the basis matrix in feature space not only achieves sparse data representation and simplifies computation but also improves the algorithm' s local learning ability and robustness.

1 Related Work

Researchers have improved NMF from different perspectives, such as constrained NMF [6–9], structured NMF [10, 11], and generalized NMF [12–14]. The most common constrained NMF is sparse NMF, where sparsity constraints help improve decomposition uniqueness and obtain local-based representations, typically achieved through the L_1 norm [6] for feature selection and sparse data representation. Orthogonal NMF performs well because its results correspond to sparse regions in a unique solution region, learning the most salient features

[7]. Graph regularized NMF (GRNMF) improves performance in clustering tasks such as documents and images by constructing nearest neighbor graphs on scattered data points to model manifold structure [8, 9]. Formula weighting is a common improvement method for learning algorithms that can emphasize the relative importance of different components. Weighted NMF is popular in collaborative filtering and clustering tasks because they incorporate prior knowledge into the loss function based on instance connections [10, 11]. For tasks with complex and heterogeneous information sources, generalized NMF has been proposed, such as Semi-NMF [12], non-negative tensor decomposition [13], and non-negative matrix-set factorization [14].

A special method, dropout NMF [22], can prevent mutual adaptation of hidden units by altering the update process of latent features. Since fixed co-adaptation is broken, hidden units can still learn from others but with less dependency. Inspired by these approaches, this paper proposes a novel NMF that works by minimizing cosine similarity between latent features. We find that NMF can be improved by breaking correlations between latent features. Therefore, in the next chapter, we incorporate cosine similarity into NMF. Additionally, $L_{2,1/2}$ sparse constraints added to the coefficient matrix U are used to select the most discriminative sparse features.

---

2 Methods

The purpose of NMF is to make the product of coefficient matrix U and basis matrix V approximate the original data matrix X as closely as possible. The NMF formula is as follows:

$$X \approx UV$$

2.1 Sparse Non-negative Matrix Factorization for Independent Feature Learning

NMF can be represented as a linear neural network because the input x is represented by a linear combination of basis vectors in U:

$$x \approx \sum_{k=1}^{K} u_k v_k$$

where u_k is the k-th latent feature.

Since latent features in NMF are correlated, co-adaptation refers to the state where updates stop at a saddle point where the loss L can still be further optimized until reaching maximum iterations. Researchers have attempted to avoid

co-adaptation by performing improved NMF under different initialization strategies on the same dataset, which brings significant computational complexity [18]. Therefore, this paper proposes a novel NMF where, in each iteration, features are updated independently by minimizing cosine similarity between latent features, making them gradually uncorrelated.

2.1.1 Sparse Constraint Although NMF can reduce the dimensionality of original data, selecting discriminative features and achieving sparse data representation remains challenging. Therefore, the $L_{2,1/2}$-norm sparse constraint is introduced as an additional condition on the basis matrix U. It can be expressed as:

$$\|U\|_{2,1/2} = \sum_{i=1}^{m} \left(\sum_{t=1}^{K} u_{it}^2 \right)^{1/4}$$

2.1.2 Cosine Similarity Measure Theodoridis et al. [24] defined cosine similarity measure as:

$$\text{cosine}(x, y) = \frac{x^T y}{\|x\|\|y\|}$$

where $\|x\| = \sqrt{\sum_{i=1}^{l} x_i^2}$ and $\|y\| = \sqrt{\sum_{i=1}^{l} y_i^2}$ are the lengths of vectors x and y, respectively; both x and y are l-dimensional vectors. Since cosine similarity is easy to interpret and simple to compute for sparse vectors, it is widely used in text mining and information retrieval [25].

2.1.3 Iterative Update Rules Considering the above factors, we define the objective function for Sparse Non-negative Matrix Factorization for Independent Feature Learning as follows:

$$L = \|X - UV\|_F^2 + \alpha\|V\|_F^2 + \beta\|U\|_{2,1/2} + \theta \sum_{i,j} \cos^2(u_i, u_j)$$

where α, β, and θ are non-negative parameters that balance the weights of the reconstruction error terms, with α and β being sparsity parameters and θ being the cosine similarity parameter.

To optimize this objective function, we can transform the objective function in equation (7) into:

$$L = \text{Tr}(XX^T) - 2\text{Tr}(XVU^T) + \text{Tr}(UVVU^T) + \alpha\text{Tr}(VV^T) + \beta\text{Tr}(U^TQU) + \theta\text{Tr}(U^TSU)$$

where Q is a diagonal matrix whose i-th diagonal element can be calculated as:

$$q_{ij} = \frac{3/2}{4\sqrt{\|u_i\|_2}}$$

To avoid overflow, a sufficiently small constant ε is added when defining matrix Q, so equation (8) can also be written as:

$$q_{ij} = \frac{3/2}{4\max(\|u_i\|_2, \varepsilon)}$$

In this paper, we adopt the squared Euclidean distance loss function:

$$L = \|X - UV\|_F^2$$

The iterative update rules using gradient descent algorithm are as follows:

$$U \leftarrow U \odot \frac{XV^T + \theta SU}{UVV^T + \beta QU}$$

$$V \leftarrow V \odot \frac{U^T X}{U^T UV + \alpha V}$$

where $\odot$ denotes element-wise multiplication. The elements of matrix S can be calculated as:

$$S_{ij} = \sum_k \cos(u_{ik}, u_{jk})$$

To obtain the iterative update rules for basis matrix U and coefficient matrix V, we take the partial derivatives of L:

$$\frac{\partial L}{\partial U} = -2XV^T + 2UVV^T + 2\beta QU + 2\theta SU$$

$$\frac{\partial L}{\partial V} = -2U^T X + 2U^T UV + 2\alpha V$$

The iterative update rules for basis matrix U and coefficient matrix V are shown as follows:

$$U_{mk} \leftarrow U_{mk} \frac{(XV^T)_{mk}}{(UVV^T + \beta QU + \theta SU)_{mk}}$$

$$V_{kn} \leftarrow V_{kn} \frac{(U^T X)_{kn}}{(U^T U V + \alpha V)_{kn}}$$

2.2 Convergence Analysis

This section analyzes the convergence of the algorithm, proving that the objective function in equation (7) monotonically decreases under the iterative update rules in equations (14)-(15). We first analyze the convergence of the iterative update rule in equation (14).

Lemma 2.1 [26]: By solving equations (22) and (23) simultaneously, we know that h^* is a local minimum of equation (23), and h^* is the corresponding local minimum point.

Proof: From Lemma 2.1, we have:

$$G(h, h) = F(h)$$

We define a function as follows:

$$G(h, h') = F(h') + (h - h')^T \nabla F(h') + \frac{1}{2}(h - h')^T K(h')(h - h')$$

From this, we can obtain inequality (19):

$$F(h^{t+1}) \leq G(h^{t+1}, h^t) \leq G(h^t, h^t) = F(h^t)$$

Combining equation (15) and inequality (18), we can obtain:

$$F(h^{t+1}) \leq F(h^t)$$

Therefore, the objective function in equation (6) monotonically decreases under the iterative update rule in equation (14). Next, we continue to analyze the convergence of the iterative update rule in equation (15).

Definition 2.1: $G(h, h')$ is an auxiliary function of $F(h)$ if it satisfies:

$$G(h, h') \geq F(h), \quad G(h, h) = F(h)$$

Lemma 2.2: If G is an auxiliary function, then F is non-increasing under the update:

$$h^{t+1} = \arg\min_h G(h, h^t)$$

Lemma 2.3: The function $G(V, V^t)$ is convergent.

Proof: We can clearly prove that:

$$G(V, V^t) \geq F(V)$$

and

$$G(V^t, V^t) = F(V^t)$$

Therefore, $F(V)$ is monotonically decreasing under equation (15). The function $G(V, V^t)$ is an auxiliary function, so F decreases monotonically through equation (15).

---

3 Experiments

3.1 Dataset

We select the publicly available corpus fetch_20newsgroups dataset for experiments. This dataset contains approximately 20,000 news documents uniformly divided into 20 different topic categories. Originally collected by Lang, it contains 18,846 documents, and after preprocessing in this paper, only 1,000 terms are retained. Each document is converted into a vector $x \in \mathbb{R}^{1000}$. We apply standard NMF, sparse NMF (SNMF), and our new algorithm to this data matrix $X \in \mathbb{R}^{18846 \times 1000}$ for comparative study of algorithm accuracy and recall.

3.2 Experimental Setup

Evaluation Metrics: In this paper, we use three evaluation metrics—precision, recall, and F-measure (F1-score)—to assess the clustering performance of the aforementioned algorithms. The formulas are as follows:

$$\text{precision} = \frac{n_{ij}}{n_j}$$

$$\text{recall} = \frac{n_{ij}}{n_i}$$

$$F\text{-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

where n_{ij} is the number of documents from known category i that belong to cluster j; n_i is the number of documents in known category i; and n_j is the number of documents in cluster j.

The F-measure simultaneously considers both precision P and recall r to compute its score. It is a composite metric where higher values indicate better clustering performance.

Parameter Settings: For the fetch_20newsgroups dataset, the parameters α, β, and θ for the new NMF are selected from the range {0.01~0.1}, and the regularization balance parameters for SNMF are the same as above. To verify algorithm performance under different numbers of topics, K is set to {5, 10, 15, 20} respectively. We compare the INMF algorithm with traditional NMF and SNMF algorithms, with detailed experimental results analyzed in the next section.

3.3 Clustering Results

Tables 1-4 show the clustering results of the two baseline methods compared with our method on the fetch_20newsgroups dataset, comparing precision, recall, and F1-score metrics under different K values. The best performance among the three methods is shown in bold. The results indicate that regardless of the K value, our proposed new method consistently outperforms traditional methods, demonstrating the effectiveness of our approach in preventing co-adaptation of latent features.

Results on k=5 datasets
Results on k=10 datasets
Results on k=15 datasets
Results on k=20 datasets

Figures 1-3 intuitively show the comparison of clustering results between our proposed INMF method and the other two methods as K values change, evaluated from precision, recall, and F-measure respectively.

As shown in Figure 1, when k takes values of 5, 10, 15, and 20, INMF is compared with traditional NMF and SNMF in terms of precision. Through bar chart comparison, we can more intuitively see that under different K values, the precision of our INMF algorithm is superior to the other two algorithms.

[Figure 1: see original paper] Accuracy comparison of NMF, SNMF and INMF under each K values

As shown in Figure 2, when k takes different values, INMF is compared with traditional NMF and SNMF in terms of recall. The bar chart comparison clearly shows that the recall rate of our INMF algorithm is better than the other two algorithms, and the advantage becomes more pronounced as K increases.

[Figure 2: see original paper] Comparison of recall rates of NMF, SNMF and INMF under each K values

As shown in Figure 3, when k takes different values, INMF is compared with traditional NMF and SNMF in terms of F-measure. The bar chart comparison shows that the F-measure of our INMF algorithm is superior to the other two

algorithms, but the overall performance decreases as K increases because larger datasets with more topics are more difficult for clustering.

[Figure 3: see original paper] Comparison of F-measure of NMF, SNMF and INMF under each K values

3.4 Parameter Selection and Convergence Analysis

When p is set to $\{0.1, 0.3, 0.5, 0.7, 0.9\}$ respectively, we compare the convergence curves of traditional NMF, SNMF, and our algorithm on the dataset at $k = 5$.

[Figure 4: see original paper] Convergence Curves of NMF, SNMF and INMF on fetch_20newsgroups

Considering both clustering performance and time consumption, we set the most appropriate parameters. The comparative analysis shows that all algorithms tend to converge within 20 iterations. On the fetch_20newsgroups dataset, NMF and SNMF require approximately 18 and 19 iterations respectively for the objective function to stabilize, while INMF can converge at 15 iterations. This is because after our improvements, more latent features become uncorrelated after each iteration, and the cumulative effect leads to better performance.

---

4 Conclusion

This paper analyzes the impact of correlations between latent features in traditional NMF on algorithm performance and proposes an Independent feature learning sparse Non-negative Matrix Factorization method (INMF). This method can not only utilize cosine similarity of vectors but also effectively learn local information of targets. In INMF, latent features are updated by minimizing cosine similarity in each iteration. The proposed algorithm effectively prevents feature co-adaptation, making latent features more explicit and discriminative. Additionally, the $L_{2,1/2}$ sparse constraint as a condition in NMF can make the decomposed matrices have good sparsity, enhancing the algorithm' s local learning ability and robustness. From the experimental results above, we can conclude that the proposed algorithm achieves good performance in both precision and recall. In the future, we will explore NMF with other loss functions and constraints, and apply the proposed new method to different domains such as computer vision, speech recognition, network security, and biomedical engineering.

References

[1] Cai Jing, Wang Wanliang, Zheng Jianbiao, et al. Incremental nonnegative matrix factorization algorithm based on fisher discriminant analysis [J]. Pattern Recognition & Artificial Intelligence, 2018, 31(6): 505-515.

[2] Wang Chaofeng, Shi Jun, Wu Jinjie, et al. Multi-view joint non-negative matrix factorization algorithm based on Hessian regularization [J]. Computer Engineering, 2017, 43(11): 134-139.

[3] Wang Jintao, Cao Yudong, Sun Fuming. Incremental learning algorithm for regular non-negative matrix partition decomposition of sparse constrained graphs [J]. Journal of Computer Applications, 2017, 37(4): 1071-1074.

[4] Han Suqing, Jia Ru. K-Means clustering algorithm based on sparse constrained nonnegative matrix factorization [J]. Journal of Data Acquisition & Processing, 2017, 32(6): 1216-1222.

[5] Sun Jing, Cai Xixi, Sun Fuming. Multi-constraint nonnegative matrix factorization algorithm based on feature fusion [J]. Journal of Computer Applications, 2017, 37(10): 2834-2840.

[6] Mohammadiha N, Leijon A. Nonnegative matrix factorization using projected gradient algorithms with sparseness constraints [C]// Proc of IEEE International Symposium on Signal Processing and Information Technology. New York: IEEE Conference Proceedings, 2009: 418-423.

[7] Ding Chris, Li Tao, Peng Wei, et al. Orthogonal nonnegative matrix factorizations for clustering [C]// Proc of ACM Sigkdd International Conference on Knowledge Discovery & Data Mining. USA: Philadelphia, PA, 2006.

[8] Cai Deng, He Xiaofei, Han Jiawei, et al. Graph regularized non-negative matrix factorization for data representation [J]. IEEE Trans Pattern Anal Mach Intell, 2011, 33(8): 1548-1560.

[9] Zhi Ruicong, Flierl Markus, Ruan Qiuqi, et al. Facial expression recognition based on graph-preserving sparse non-negative matrix factorization [C]// Proc of IEEE International Conference on Image Processing. [S. l.]: IEEE Signal Processing Society, 2010: 3293-3296.

[10] Mao Yun, Saul L K. Modeling distances in large-scale networks by matrix factorization [C]// Proc of the 4th ACM SIGCOMM Conference on Internet Measurement. New York: ACM Press, 2004: 278-287.

[11] Kim Y D, Choi S. Weighted nonnegative matrix factorization [C]// Proc of IEEE International Conference on Acoustics, Speech and Signal Processing. Washington DC: IEEE Computer Society, 2009: 1541-1544.

[12] Ding C, Li Tao, Jordan M I. Convex and semi-nonnegative matrix factorizations [J]. IEEE Transa on Pattern Analysis & Machine Intelligence, 2010, 32(1): 45-55.

[13] Mørup M, Hansen L K, Arnfred S M. Algorithms for sparse nonnegative Tucker decompositions [J]. Neural Computation, 2008, 20(8): 2112-2131.

[14] Li Le, Zhang Yujin. Non-negative matrix-set factorization [C]// Proc of Fourth International Conference on Image and Graphics. Piscataway, NJ: IEEE Press, 2007: 564-569.

[15] Foucart S, Lai Minjun. The sparest solutions of underdetermined linear system by Lq-minimization for $0 < q \leq 1$ [J]. Applied & Computational Harmonic Analysis, 2009, 26(3): 395-407.

[16] Chartrand R. Exact reconstruction of sparse signals via nonconvex minimization [J]. IEEE Signal Processing Letters, 2007, 14(10): 707-710.

[17] Chartrand R. Fast algorithms for nonconvex compressive sensing: MRI reconstruction from very few data [C]// Proc of IEEE International Conference on Symposium on Biomedical Imaging. 2009: 262-265.

[18] Xu Zongben, Chang Xiangyu, Xu Fengmin, et al. L1/2 regularization: a thresholding representation theory and a fast solver [J]. IEEE Trans on Neural Networks & Learning Systems, 2012, 23(7): 1013-1027.

[19] ZongBen Xu, Hai Zhang, Yao Wang, et al. L1/2regularization [J]. Science in China Series F: Information Science, 2010, 53(6): 1159-1169.

[20] Nie Feiping, Huang Heng, Cai Xiao, et al. Efficient and robust feature selection via joint L2, 1-norms minimization [C]// Proc of International Conference on Neural Information Processing Systems. [S.l.]: Curran Associates Inc, 2010: 1813-1821.

[21] Wang Liping, Chen Songcan. l2, p-matrix norm and its application in feature selection [EB/OL]. (2013). Available: http://arxiv.org/abs/1303.3865.

[22] He Zhicheng, Liu Jie, Liu Caihua, et al. Dropout non-negative matrix factorization for independent feature learning [M]// Natural Language Understanding and Intelligent Applications. [S.l.]: Springer International Publishing, 2016: 201-212.

[23] Langville A N, Meyer C D, Albright R, et al. Algorithms, initializations, and convergence for the nonnegative matrix factorization [J]. Eprint Arxiv, 2014.

[24] Theodoridis S, Koutroumbas K. Pattern recognition (fourth edition) [M]. [S.l.]: Academic Press, 2008.

[25] Dhillon I S, Modha D S. Concept decompositions for large sparse text data using clustering [J]. Machine Learning, 2001, 42(1-2): 143-175.

[26] Shi Caijuan, Ruan Qiuiqi, An Gaoyun, et al. Hessian semi-supervised sparse feature selection based on l2, 1/2-matrix norm [J]. IEEE Trans on Multimedia, 2014, 17(1): 16-28.

Note: Figure translations are in progress. See original paper for figures.