Robust and Fast Non-Negative Matrix Factorization for Large Scale Data Clustering

De Wang, Feiping Nie, Heng Huang
Department of Computer Science and Engineering, University of Texas at Arlington

Abstract

Human action recognition is important in improving human life in various aspects. However, the outliers and noise in data often bother the clustering tasks. Therefore, there is a great need for the robust data clustering techniques. Non-negative matrix factorization (NMF) and Non-negative Matrix Tri-Factorization (NMTF) methods have been widely researched these years and applied to many data clustering applications. With the presence of outliers, most previous NMF/NMTF models fail to achieve the optimal clustering performance. To address this challenge, in this paper, we propose three new NMF and NMTF models which are robust to outliers. Efficient algorithms are derived, which converge much faster than previous NMF methods and as fast as K-means algorithm, and scalable to large-scale data sets. Experimental results on both synthetic and real world data sets show that our methods outperform other NMF and NMTF methods in most cases, and in the meanwhile, take much less computational time.

Motivation

Typical NMF model solves the following objective functions, subjecting to different kinds of constraints.

$$\min_{F,G} \left\|X - FG^T\right\|_F^2 \qquad s.t. \quad F \geq 0, G \geq 0 \tag{1}$$

where $X \in \mathbb{R}_+^{d\times n}$ is a data matrix with d features and n samples. This model has close relationshi pwith k-means algorithm: $F \in \mathbb{R}_+^{d\times c}$ can be viewed as cluster centroids, and $G \in \mathbb{R}_+^{n\times c}$ can be viewed as clustering indicator matrix. There are several drawbacks of typical NMF methods:

- Converges slowly: usually takes hundreds of iterations
- High computational cost: involves large matrix multiplication in each iteration
- Soft clustering: need post processing step to get the final clustering results.
- Not robust to outliers

Fast Robust NMF Models

To make the NMF/NMTF models robust to outliers, instead of using Frobenius norm, we propose to use the $\ell_{2,1}$-norm and ℓ_1-norm as loss measurements. The new robust and fast NMF models aim to minimize the following objective functions:

$$\min_{F\geq 0, G\in Ind} \left\|X - FG^T\right\|_1 \tag{2}$$

$$\min_{F\geq 0, G\in Ind} \left\|X - FG^T\right\|_{2,1} \tag{3}$$

$$\min_{F\in Ind, G\in Ind, S\geq 0} \left\|X - FSG^T\right\|_1 \tag{4}$$

where $G \in Ind$ or $F \in Ind$ indicates that G and F are indicator matrices, *i.e.* $g_{ij} = 1$ if x_i belongs to class j, and $g_{ij} = 0$ otherwise. There is only one element can be non-zero in each row of binary indicator matrix. Since our methods are robust and fast NMF/NMTF models, we call the three models as RFNMF_L1, RFNMF, and RFNMTF, respectively. The proposed methods have several advantages compared to typical NMF models:

- Converges fast
- Light computation in each iteration: simple meadian finding plus label assignment
- Hard clustring: no post processing step
- Robust to outliers using $\ell_{2,1}/\ell_1$ loss functions

Optimization for RFNMF_L1: 1) When G fixed:

$$\min_{F\geq 0} \left\|X - FG^T\right\|_1 \tag{5}$$

$$\Rightarrow \quad \min_{F\geq 0} \sum_i \left\|X_{i\cdot} - \sum_k F_{ik} G^T_{\cdot k}\right\|_1$$

$$\Rightarrow \quad \min_{F\geq 0} \sum_i \sum_k \sum_{G_{jk}=1} |X_{ij} - F_{ik}|$$

The optimal solution of F_{ik} can be efficiently obtained by finding the median values of samples belong to the k-th cluster.

2) When F fixed, optimal solution of G is:

$$g_{ij} = \begin{cases} 1 & j = arg\min_k \|X_{\cdot i} - F_{\cdot k}\|_1 \\ 0 & otherwise \end{cases} \tag{6}$$

Experimental Results

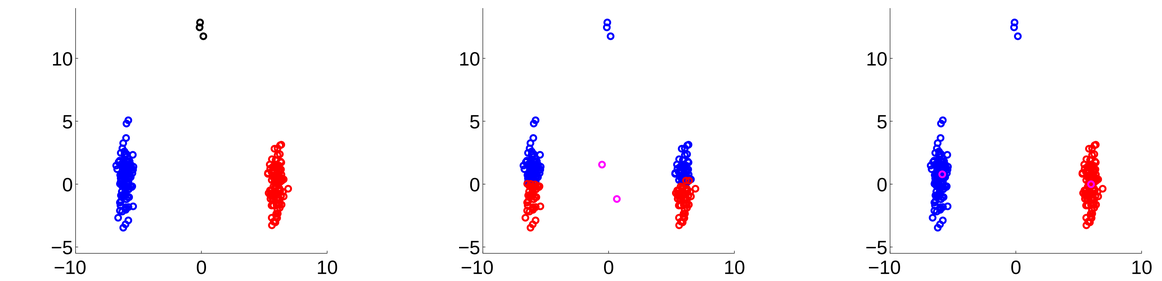

Figure 1: Clustering performance on synthetic data. Blue points and red points are normal data drawn from two gaussian distributions. Black points are outliers. Magenta points are computed cluster centroids. Left: Original data points; Center: Clustering results using typical NMF model, which is not robust to outliers; Right: Clustering results using proposed NMF models.

	normal data	outliers	all data
NMF	6.02	10.82	6.09
Our methods	1.27	12.96	1.45

Table 1: Average distance from the centroids for normal data, outliers, and all data in the synthetic data set. The distance of each points to the corresponding centroids for our method is smaller than typical NMF method.

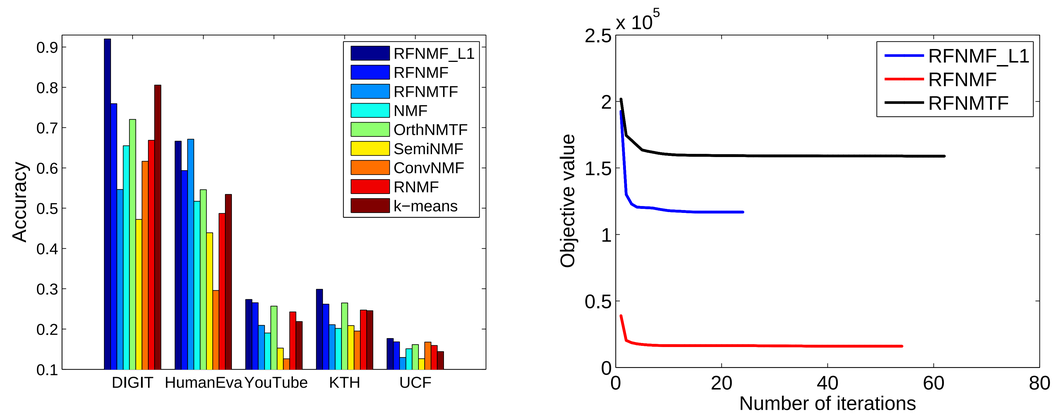

Figure 2: Left: Clustering accuracy comparison; Our methods outperform other comparison methods. Right: Convergence of objective value for the proposed NMF/NMTF models. Our methods usually take less than 20 iterations to converge.

	RFNMF_L1	RFNMF	RFNMTF	NMF	OrthNMTF	SemiNMF	ConvNMF	RNMF	K-means
DIGIT	**3.57**	5.98	12.13	54.89	243.27	46.98	55.47	264.22	**1.72**
HumanEva	**5.79**	49.32	23.63	67.64	2181.26	12.60	775.23	1626.69	**2.23**
YouTube	**3.70**	**7.86**	20.32	366.91	553.55	43.89	37.68	203.30	28.25
KTH	**8.04**	**12.51**	26.41	300.72	676.98	64.71	65.04	499.08	27.83
UCF	**246.19**	251.20	278.06	1974.57	3891.82	349.75	639.00	1580.65	**163.20**

Table 2: Computational time (in seconds) comparison. Averaged over 10 repetitions. Our methods are as fast as k-means, and take much less time to converge than other comparison methods.