



Robust Multi-view Features Fusion Method Based on CNMF

Bangjun Wang¹(✉), Liu Yang², Li Zhang¹, and Fanzhang Li¹

¹ School of Computer Science and Technology & Joint International Research Laboratory of Machine Learning and Neuromorphic Computing, Soochow University, Suzhou, China
wangbangjun@suda.edu.cn

² School of Computer Science and Technology, Tianjin University, Tianjin, China

Abstract. Multi-view feature fusion should be expected to mine implicit nature relationships among multiple views and effectively combine the data presented by multiple views to obtain the new feature representation of the object using a right model. In practical applications, Collective Matrix Factorization (CMF) has good effects on the fusion of multi-view data, but for noise-containing situations, the generalization ability is poor. Based on this, the paper came up with a Robust Collective Non-negative Matrix Factorization (RCNMF) model which can learn the shared feature representation of multi-view data and denoise at the same time. Based on several public data sets, experimental results fully demonstrate the effectiveness of the proposed method.

Keywords: Multi-view · CMF · Future fusion

1 Introduction

Due to the diversity and convenience of data acquisition channels, a large amount of multi-view data has been accumulated, and its rational use has attracted more and more attention in machine learning and pattern recognition. Often, multi-view data is characterized by different features which are homogenous or heterogeneous. For example, multi-view features in human body object recognition such as face image, fingerprint information, sound information and signature information belong to heterogeneous data sets. Although multi-view data has different feature descriptions, the semantics represented under certain conditions are consistent. It can therefore be assumed that they share the same implicit high-level semantic space [4, 8, 9, 15]. Based on which, there have been many researches, CCA (Canonical Correlation Analysis) [10] is a multivariate statistical analysis method studied the correlation between two groups of variables and extended by Chaudhuri et al. [4] to multiple views and obtained a multi-datasets canonical correlation analysis (Multiset CCA). Considering the nonlinear situation, Haroon et al. [9] further expanded this method to KCCA

(Kernel CCA) by adding kernels. Guo et al. [8] proposed a convex subspace representation learning method for unsupervised multi-view clustering. Li et al. [14] proposed a discriminative multi-view interactive image rearrangement algorithm that integrated users’ feedbacks and intents to fully describe multiple features of an image. Zhang et al. [19] proposed a multi-view dimension collaboration reduction approach considering the complementary of different views and similarities among the data points. The method enhances the correlation between different views and restrains the inconsistencies simultaneously with the kernel matching constraint based on the Hilbert-Schmidt independence criterion. Hou et al. [11] proposed a multi-view unsupervised feature selection algorithm using adaptive similarity and views weighting to overcome the problem of obtaining markup data in multi-view feature selection.

Non-negative matrix factorization (NMF) [13] shows good performance in single-view subspace clustering. Liu et al. [1, 15] applied it to multi-view datasets. Zhao et al. [21] proposed a deep matrix factorization framework for multi-view clustering, using nonnegative matrix factorization to learn the hierarchical semantics of multi-view data in a hierarchical way. In recent years, the Collective Matrix Factorization (CMF) method [17] has become an important method for multi-view learning. This method can be used for finding the shared subspace of multiple views, so as to achieve the purpose of dimensionality reduction and feature fusion. CMF factorizes multiple matrices at the same time, and shares the subspace representation in the process of factorization, where \mathbf{X}^1 and \mathbf{X}^2 respectively denote the data matrix of the image and text, \mathbf{W}^1 and \mathbf{W}^2 are respectively the mapping matrix of the image and text, and \mathbf{H} is the representation of the image and text data in subspaces. CMF has been successfully applied to many applications [16, 18, 22]. However, those CMF-based approaches did not consider scenes that the data contains noise in multi-view learning. In many practical applications, the data usually contain a lot of noise. In Fig. 1, for example, it is clear that the parts surrounded by the solid red boxes does not belong to the same class as the others, which are not helpful for the final clustering task. So the two parts can be regarded as noise.

For the NMF model in a single view, many researchers have considered the problem of how to remove the noise. These methods fall broadly into two categories. One kind of them draws on the idea of the robust PCA [3], introducing the error matrix, and adding sparse constraints to simulate the sparse noise in the data. Zhang et al. [20] applied this idea to the traditional NMF method. Such methods assume that noise is sparsely presented in the data so it’s not suitable for processing data that contain dense noise points. The second category uses a different norm than Frobenius as the criterion for sample error in order to reduce the effect of noise on overall performance. The often criteria used are: ℓ_{21} distance [12], Manhattan distance [7], related entropy induction metric function [6], Alpha-Beta Divergence [5] and so on. Among them, the methods based on ℓ_{21} distance achieved good results. As shown in Fig. 2, the method like ℓ_{21} -NMF can weaken the effect of noise point on subspace learning well. Also, ℓ_{21} distances can easily be expanded from single view to multiple views. Yang et al. [18] used

ℓ_{21} distance to solve the problem of noise in migration learning, using two independent robust NMF models for the source and target respectively. While the limitation of this model is that it can only deal with isomorphic data, multi-view data is usually heterogeneous, so the model can not be applied to heterogeneous multi-view data.

To solve this noise issue, this paper proposes a Robust Collective Non-negative Matrix Factorization (RCNMF) method based on nonnegative matrix factorization. RCNMF introduces the ℓ_{21} norm for the error of each view simultaneously and also for the new representation of multi-view data in the shared space which weakens the impact of noise on the overall performance. An iterative method is used to solve the objective function of RCNMF. Clustering tasks are performed to verify the performance of subspace fusion methods. By comparing with some existing methods on the real-world data-sets, the proposed method can effectively fuse the multi-view data and solve the noise problem.



Fig. 1. Display of noise in data

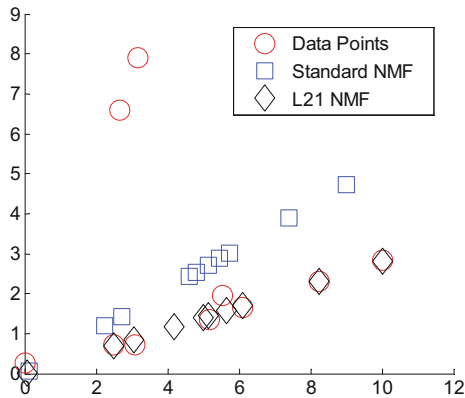


Fig. 2. The result of ℓ_{21} NMF model

The rest of this paper is organized as follows. Section 2 introduces the model of RCNFM proposed in this paper. Section 3 describes the related experiment and the result analysis. Section 4 makes a conclusion about this paper.

2 Robust Multi-view Subspace Fusion Method

This section introduces the traditional method of collective matrix factorization, then explains the robust collective non-negative matrix factorization model proposed in this paper, and finally shows the solution process.

Let $\{\{\mathbf{x}_i^j\}_{i=1}^n\}_{j=1}^m$ be the set of multi-view data, where $\mathbf{x}_i^j \in \mathbb{R}^{d_j}$ is a d_j dimension vector of the j th view, n is the number of samples, and m is the number of views. The task is to cluster the n samples into different groups. For simplicity, the data of each view is represented as a matrix $\mathbf{X}^j = [\mathbf{x}_1^j \cdots \mathbf{x}_n^j] \in \mathbb{R}^{d_j \times n}$, then the data of multi-view is $\{\mathbf{X}^j\}_{j=1}^m$. Since much of the data in multi-view is nonnegative, such as text and image data in the form of bag of words, so this paper assumes that each element is nonnegative.

2.1 CMF

CMF assumes that the representation of different views' data should be consistent in the new shared space, in which the same data matrix $\mathbf{H} \in \mathbb{R}^{r \times n}$ is shared where r is the feature dimension of the new space. Then we can learn the mapping matrix of each view to the shared space $\{\mathbf{W}^j\}_{j=1}^m \in \mathbb{R}^{d_j \times r}$. The related objective function is as follows:

$$\min_{\{\mathbf{W}^j\}_{j=1}^m, \mathbf{H}} \sum_{j=1}^m \lambda_j \|\mathbf{X}^j - \mathbf{W}^j \mathbf{H}\|_F^2 \quad (1)$$

where λ_j is a parameter that represents the coefficient of each view. By optimizing the objective function (1), we can get the mapping matrix $\{\mathbf{W}^j\}_{j=1}^m$ and the new representation matrix \mathbf{H} in the subspace.

Since this paper considers only nonnegative data, we can add non-negative constraints on the basis of CMF and call it CNMF. The collective non-negative matrix factorization model not only has the advantage of finding the essential components of the data, but also can reduce the solution space of the matrix factorization. The objective function becomes:

$$\begin{aligned} \min_{\{\mathbf{W}^j\}_{j=1}^m, \mathbf{H}} \sum_{j=1}^m \lambda_j \|\mathbf{X}^j - \mathbf{W}^j \mathbf{H}\|_F^2 \\ \text{subject to } \mathbf{W}^j, \mathbf{H} \geq 0, \quad j = 1, \dots, m \end{aligned} \quad (2)$$

Note that both the CMF and CNMF models assume that the high-level semantics of multi-view data are consistent, both of them map multi-view data to the same shared subspace, but do not consider the scene of noise.

2.2 RCNMF

In this section, we introduce a robust collective matrix factorization approach that is used primarily to reduce the effects of noise in data and to obtain more accurate subspace representation of features at the same time. The objective function is as follows:

$$\begin{aligned} \min_{\{\mathbf{W}^j\}_{j=1}^m, \mathbf{H}} \sum_{j=1}^m \lambda_j \|\mathbf{X}^j - \mathbf{W}^j \mathbf{H}\|_{21} + \alpha \|\mathbf{H}\|_{21} \\ \text{subject to } \mathbf{W}^j, \mathbf{H} \geq 0, \quad j = 1, \dots, m \end{aligned} \quad (3)$$

where the parameter λ_j means the j th view's weight and α is the regularization coefficient. The norm ℓ_{21} of matrix \mathbf{H} is defined as:

$$\|\mathbf{H}\|_{21} = \sum_{i=1}^n \sqrt{\sum_{k=1}^r \mathbf{H}_{ki}^2} = \sum_{i=1}^n \|\mathbf{h}_i\|_2$$

where \mathbf{h}_i is the i th column of \mathbf{H} . The norm ℓ_{21} of matrix $(\mathbf{X}^j - \mathbf{W}^j \mathbf{H})$ is defined as:

$$\|\mathbf{X}^j - \mathbf{W}^j \mathbf{H}\|_{21} = \sum_{i=1}^n \sqrt{\sum_{p=1}^{d_j} (\mathbf{X}^j - \mathbf{W}^j \mathbf{H})_{pi}^2} = \sum_{i=1}^n \|\mathbf{X}_i^j - \mathbf{W}^j \mathbf{H}_i\|_2$$

and $(\mathbf{X}^j - \mathbf{W}^j \mathbf{H})$ takes the norm ℓ_{21} constraining errors.

The square term is no longer used to calculate each data point's error in this paper, which mainly hope to weak the impact of the larger error of the noise point on the entire data set. In extreme cases, if $\|\mathbf{X}_i^j - \mathbf{W}^j \mathbf{H}_i\|_2 = 0$, then the reconstructed i th sample and the original data are exactly the same. Thus the i th sample is less likely to be noise point. If the value of $\|\mathbf{X}_i^j - \mathbf{W}^j \mathbf{H}_i\|_2$ is large, then it means the reconstruction error is large and this sample is probably the noise point. Then we should weaken the effect of this sample while learning \mathbf{W} and \mathbf{H} . Similarly, taking the ℓ_{21} norm for \mathbf{H} is also expected weakening the effect of noise points. In addition, the norm ℓ_{21} regularization matrix performs the noise processing in a batch manner, so the mutual influence among the samples is considered in the denoising process at the same time.

2.3 Solution to RCNMF

Solving the RCNMF model (3) is a non-convex optimization problem, so we solved it iteratively. For each subproblems, we approximate it using an imprecise method. Existing methods for finding inaccurate solution include multiplicative update rule, projection ALS method and cyclic block coordinate gradient projection method. Among them, the multiplicative update rule method is relatively simple to calculate and widely used. Therefore, this paper uses the

multiplicative update rule method to solve the objective function (3). Specific steps are as follows:

1. Update \mathbf{W}^j :

Fixing $\mathbf{H} = \mathbf{H}^\tau$ (where τ is the current iteration number), then \mathbf{W}^j can be obtained by solving

$$\begin{aligned} & \min_{\mathbf{W}^j} \|\mathbf{X}^j - \mathbf{W}^j \mathbf{H}\|_{21} \\ & \text{subject to } \mathbf{W}^j \geq 0, \quad j = 1, \dots, m \end{aligned} \quad (4)$$

The updating formula of \mathbf{W}^j is:

$$(\mathbf{W}^j)_{pk}^{\tau+1} := (\mathbf{W}^j)_{pk}^\tau \frac{(\mathbf{X}^j \mathbf{D}^j \mathbf{H}^T)_{pk}}{((\mathbf{W}^j)^\tau \mathbf{H} \mathbf{D}^j \mathbf{H}^T)_{pk}} \quad (5)$$

where \mathbf{D}^j is a diagonal matrix whose diagonal elements are $D_{ii}^j = 1/\sqrt{\sum_{k=1}^{d_j} (\mathbf{X}^j - \mathbf{W}^j \mathbf{H})_{ki}^2}$. Since \mathbf{X}^j and \mathbf{H} are known when solving \mathbf{W}^j and the data of each views is independent, so all views' \mathbf{W}^j can be calculated in parallel.

2. Update \mathbf{H} :

Fixing $(\mathbf{W}^j) = (\mathbf{W}^j)^\tau$, \mathbf{H} can be solved by:

$$\begin{aligned} & \min_{\mathbf{H}} \sum_{j=1}^m \lambda_j \|\mathbf{X}^j - \mathbf{W}^j \mathbf{H}\|_{21} + \alpha \|\mathbf{H}\|_{21} \\ & \text{subject to } \mathbf{H} \geq 0 \end{aligned} \quad (6)$$

The updating formula of \mathbf{H} is:

$$\mathbf{H}_{kn}^{\tau+1} := \mathbf{H}_{kn}^\tau \frac{\sum_{j=1}^m (\lambda_j (\mathbf{W}^j)^T \mathbf{X}^j \mathbf{D}^j)_{kn}}{(\sum_{j=1}^m (\lambda_j (\mathbf{W}^j)^T \mathbf{W}^j \mathbf{H}^T \mathbf{D}^j) + \alpha \mathbf{H}^\tau \mathbf{G})_{kn}} \quad (7)$$

where \mathbf{G} is a diagonal matrix whose diagonal elements are $G_{ii} = 1/\sqrt{\sum_{k=1}^r H_{ki}^2}$. The objective function (3) can converge by iteratively applying formulas (5) and (7), similar to ℓ_{21} -NMF on the traditional single view whose convergence has been analyzed by Kong et al. [12].

2.4 Algorithm and Its Complexity Analysis

Algorithm 1 shows the algorithm of RCNMF. From the model (3), we can get the representation \mathbf{H} of multi-view data in the subspace, which can be further used by all related tasks, such as clustering.

For each iteration, the computational complexity of updating \mathbf{W}^j in (5) is $O(n^2 d_j + n^2 r)$. In general, $r \ll d_j$, so the complexity of updating \mathbf{W}^j is about $O(n^2 d_j)$. Similarly, the computational complexity of updating \mathbf{H} in (7) is $O(mn^2 d_j)$. Without loss of generality, let t be the number of iterations. Totally, the complexity of RCNMF is $O(tn^2 m \sum_{j=1}^m d_j)$, where m is the number of views.

Algorithm 1. RCNMF

Inputs: A multi-view data $\{\mathbf{X}^j\}_{j=1}^m$, features dimension r of latent space, a threshold value σ , and parameters λ_j and α .

Outputs: \mathbf{H}

1. **while** $\frac{\text{Obj}^{\tau-1} - \text{Obj}^{\tau}}{\text{Obj}^{\tau-1}} \geq \sigma$.
 2. **for** $j = 1$ to m **do**
 3. using formula (5) updating \mathbf{W}^j
 4. **end for**
 5. using formula (7) updating \mathbf{H}
 6. **end while**
-

3 Experimental Results and Analysis

This section introduces experiments on the real-word data-sets: Berkeley Drosophila Genome Project *BDGP* [2], *WebKB* and *Yale* and validate the performance of RCNMF on multi-view clustering. The data-sets used in experiments are described as:

1. *BDGP*: The Berkeley Drosophila Genome Project dataset consists of 2,500 embryo images of drosophila which are five categories, each corresponds to a stage of gene growth. Each image has two views: visual features (1750 dimensions) and text features (79 dimensions).
2. *WebKB*: This dataset includes 5 categories of documents: Course, Faculty, Student, Project and Staff. We select the website link collections of Cornell University, the sample number is 195. Each sample has two views. One is the property view, the feature number is 1703; the other is the relationship diagram between the samples, it is a matrix of 195×195 .
3. *Yale*: The dataset is made up of 15 people’s face images, each has 11 pictures, including different expressions or different perspectives. Totally, there are 165 images. Each picture has three views described by three types of features: intensity, LBP and Gabor, whose dimensions are 1024, 3304 and 6750 respectively.

3.1 Compared Methods

RCNMF proposed in this paper is to learn the fusion representation of multi-view data in latent space, based on which we can complete clustering tasks. The classic unsupervised learning method (K-means) is used as a benchmark. Moreover, we compare RCNMF with three re-representation methods, NMF [13], CCA and HTLIC [22], where NMF is tested on single view, and CCA maps two views’ feature sets. The comparison algorithms are described as follows:

1. **K-means-best**: K-means-best means that we perform clustering directly on the data of each view and then pick up the best performance as its result.

2. **NMF-best:** NMF-best means that we reduce the dimension of each view with NMF separately, and then complete the clustering on the reduced data separately. The best performance is also taken as its final result.
3. **CCA:** CCA is to use canonical correlation analysis to learn shared subspaces, based on which we complete the clustering task. This method only applies to two views.
4. **HTLIC:** HTLIC is to use collective matrix factorization adding non-negative constraints and to learn the high-level feature subspace in which the clustering task is performed.

Among them, the K-means-best method works directly on the original data, NMF, CCA, HTLIC and RCNMF are all used to re-represent the data, and then K-means is used to cluster the newly represented data set. All the parameters in these methods are adjusted and we record the best result. The initial value of each variable are randomly selected because random initialization is relatively simple, and easy to calculate. In order to weaken the impact of random initialization on the final clustering performance, each parameter in each method is randomly initialized 10 times and the average result is recorded. The termination criteria for all methods is:

$$\frac{\text{Obj}^{\tau-1} - \text{Obj}^{\tau}}{\text{Obj}^{\tau-1}} < \sigma$$

where Obj^{τ} is the objective function value in the τ th loop, and σ is the threshold value. $\sigma = 10^{-4}$ in the following experiments. The performance criteria of clustering are ACC, NMI, AR, F-score, Precision and Recall. The larger the value, the better the clustering performance for all the criteria.

3.2 Effect of Parameters

This experiment takes the data set of *BDGP* to show the effect of parameters on the proposed model. *BDGP* has two views in total. View 1 is visual feature, view 2 is textual features.

1. The effect of subspace dimension r : We set the subspace dimension r to be a integer between 5 and 30 in steps of 5, Fig. 3 gives the corresponding clustering performance diagram when subspace dimension r changing in the set range. It can be seen that at $r = 20$, most of the indices of RCNMF (except for Recall) achieved the best performance in-scope. Which shows better clustering results can be obtained when the original visual features (500 dimensions) and the text features (1,000 dimensions) are fused and the dimensions are reduced to a lower level. In addition the complexity of the algorithm would also be reduced when the clustering is performed in the new space.

2. The effect of view factor λ : Fig. 4 shows the effect of the view factor λ on clustering performance. A small λ indicates that visual features are more important, a large λ indicates that text features are more important. It can be seen from Fig. 4 that the clustering performance is good when λ is close to the

point of 0.8. This shows that the contribution of the text features is greater than the image.

3. The effect of regular parameter α : Fig. 5 shows the effect of the regular parameter α on the clustering result. The larger the value of α is, the greater the proportion of $\|\mathbf{H}\|_{21}$. It is observed that when α is 0.1, the clustering result is much better. It shows that it is necessary to add robust constraints on H .

To show the convergence performance of RCNMF, Fig. 6 gives the curve of the objective function (3) vs. iterations. Figure 6 demonstrates that the proposed model can converge to a local optimal value after several iterations on the *BDGP* dataset.

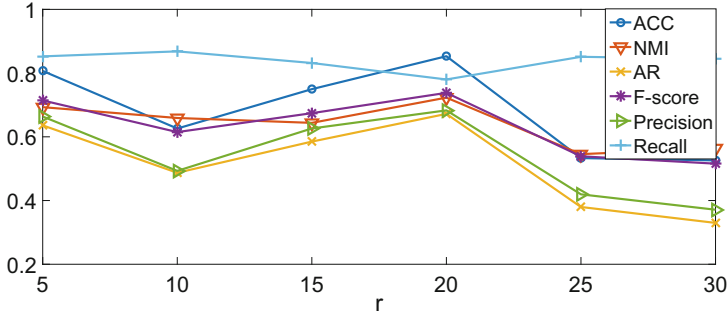


Fig. 3. Clustering performance vs. dimension of subspace r on *BDGP*

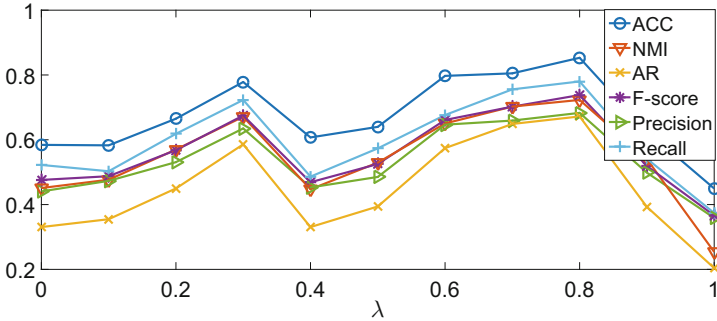


Fig. 4. Clustering performance vs. view coefficient λ on *BDGP*

3.3 Cluster Results Analysis

Tables 1, 2 and 3 show the clustering results on the selected three data-sets with the best results highlighted in black. Experimental results show that methods based on multi-view fusion are better than single-view clustering method (K-means-best and NMF-best). The main reason is that the fusion of multi-view data can obtain more information.

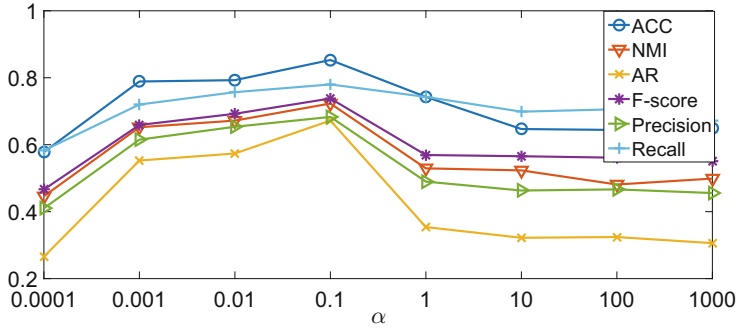


Fig. 5. Clustering performance vs. regularization parameter α on *BDGP*

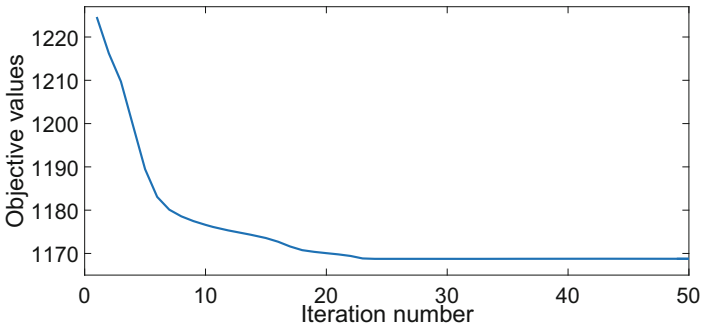


Fig. 6. Curve of convergence for (3) on *BDGP*

Table 1. Cluster results on the *BDGP* dataset

Method	K-means-best	NMF-best	CCA	HTLIC	RCNMF
ACC	0.526	0.624	0.657	0.772	0.853
NMI	0.480	0.511	0.719	0.705	0.723
AR	0.235	0.420	0.559	0.601	0.672
F-score	0.436	0.543	0.668	0.702	0.738
Precision	0.336	0.517	0.559	0.618	0.683
Recall	0.646	0.577	0.852	0.713	0.780

Table 2. Cluster results on the Cornell dataset

Method	K-means-best	NMF-best	CCA	HTLIC	RCNMF
ACC	0.564	0.564	0.567	0.583	0.596
NMI	0.405	0.354	0.366	0.385	0.420
AR	0.290	0.309	0.301	0.309	0.334
F-score	0.474	0.481	0.489	0.505	0.510
Precision	0.505	0.521	0.512	0.523	0.531
Recall	0.466	0.448	0.471	0.488	0.504

Table 3. Cluster results on the Yale dataset

Method	K-means-best	NMF-best	CCA	HTLIC	RCNMF
ACC	0.524	0.572	0.542	0.622	0.642
NMI	0.630	0.614	0.621	0.658	0.669
AR	0.371	0.362	0.402	0.420	0.435
F-score	0.415	0.405	0.415	0.460	0.472
Precision	0.364	0.366	0.380	0.414	0.426
Recall	0.484	0.454	0.490	0.520	0.522

In addition, the performances of RCNMF and HTLIC methods are better than that of CCA, this is because CCA treats all views equally and has orthogonal constraints. While RCNMF and HTLIC consider the different importance of different views, and relax the strong orthogonal constraint. Thus, the two methods can get a better performance. Since HTLIC does not consider the noise in multi-view data and RCNMF introduces the ℓ_{21} norm to effectively deal with the noise, RCNMF can weaken the influence of noise on the learning of subspace and makes the fused subspace more robust.

4 Conclusion

This paper presents a robust multi-view subspace fusion method RCNMF and applies it to clustering. In the proposed model, the multi-view features fusion is achieved by mapping multi-view features to a shared latent space. The proposed method adds the ℓ_{21} norm to the matrix factorization error and re-representation matrix based on CMF to eliminate the influence of noise. Clustering is completed on the re-representation of data in subspaces of all views. RCNMF is solved by an iteratively updated algorithm. After more than once iterations, RCNMF can converge. The performance of RCNMF is validated on three real data-sets. Through the clustering results, we can see that the proposed model can process the noise contained in the views while merging the features of multiple views into a shared subspace.

References

1. Akata, Z., Thurau, C., Bauckhage, C.: Non-negative matrix factorization in multimodality data for segmentation and label prediction. In: The 16th Computer Vision Winter Workshop (2011)
2. Cai, X., Wang, H., Huang, H., Ding, C.: Joint stage recognition and anatomical annotation of drosophila gene expression patterns. *Bioinformatics* **28**(12), 116–124 (2012)
3. Candès, E.J., Li, X., Ma, Y., Wright, J.: Robust principal component analysis. *J. ACM (JACM)* **58**(3), 11 (2011)
4. Chaudhuri, K., Kakade, S.M., Livescu, K., Sridharan, K.: Multi-view clustering via canonical correlation analysis. In: The 26th Annual International Conference on Machine Learning, pp. 129–136. ACM (2009)
5. Cichocki, A., Cruces, S., Amari, S.: Generalized alpha-beta divergence and their application to robust nonnegative matrix factorization. *Entropy* **13**, 134–170 (2011)
6. Du, L., Li, X., Shen, Y.: Robust nonnegative matrix factorization via half-quadratic minimization. In: The IEEE International Conference on Data Mining, pp. 201–210 (2012)
7. Guan, N., Tao, D., Luo, Z., Shawe-Taylor, J.: Mahnmf: manhattan non-negative matrix factorization. *ArXiv Preprint ArXiv:1207.3438* (2012)
8. Guo, Y.: Convex subspace representation learning from multi-view data. In: The Twenty-Seventh AAAI Conference on Artificial Intelligence, pp. 387–393 (2013)
9. Hardoon, D., Shawe-taylor, J.: Convergence analysis of kernel canonical correlation analysis: theory and practice. *Mach. Learn.* **74**(22), 23–38 (2009)
10. Hotelling, H.: Relations between two sets of variates. *Biometrika* **28**, 321–377 (1936)
11. Hou, C., Nie, F., Tao, H., Yi, D.: Multi-view unsupervised feature selection with adaptive similarity and view weight. *IEEE Trans. Knowl. Data Eng.* **29**(9), 1998–2011 (2017)
12. Kong, D., Ding, C., Huang, H.: Robust nonnegative matrix factorization using l_{21} -norm. In: The International Conference on Information and Knowledge Management, pp. 673–682 (2011)
13. Lee, D., Seung, H.: Learning the parts of objects by non-negative matrix factorization. *Nature* **401**(6755), 788–791 (1999)
14. Li, J., Xu, C., Yang, W., Sun, C., Tao, D.: Discriminative multi-view interactive image re-ranking. *IEEE Trans. Image Process.* **26**(7), 3113–3127 (2017)
15. Liu, J., Wang, C., Gao, J., Han, J.: Multi-view clustering via joint nonnegative matrix factorization. In: The 13th SIAM International Conference on Data Mining, pp. 252–260 (2013)
16. Pan, W., Yang, Q.: Transfer learning in heterogeneous collaborative filtering domains. *Artif. Intell.* **197**, 39–55 (2013)
17. Singh, A., Gordon, G.: Relational learning via collective matrix factorization. In: The International Conference on Knowledge Discovery and Data Mining, pp. 650–658 (2008)
18. Yang, S., Hou, C., Zhang, C., Wu, Y.: Robust non-negative matrix factorization via joint sparse and graph regularization for transfer learning. *Neural Comput. Appl.* **23**(2), 541–559 (2013)
19. Zhang, C., Fu, H., Hu, Q., Zhu, P., Cao, X.: Flexible multi-view dimensionality co-reduction. *IEEE Trans. Image Process.* **26**(2), 648–659 (2016)

20. Zhang, L., Chen, Z., Zheng, M., He, X.: Robust non-negative matrix factorization. *Front. Electr. Electron. Eng.* **6**(2), 192–200 (2011)
21. Zhao, H., Ding, Z., Fu, Y.: Multi-view clustering via deep matrix factorization. In: *The AAAI Conference on Artificial Intelligence*, pp. 2921–2927 (2017)
22. Zhu, Y., et al.: Heterogeneous transfer learning for image classification. In: *The AAAI Conference on Artificial Intelligence*, pp. 1–6 (2011)