Exploiting Correlation Consensus: Towards Subspace Clustering for Multi-modal Data

Yang Wang[†‡], Xuemin Lin[†], Lin Wu[†], Wenjie Zhang[†] and Qing Zhang[‡†]
[†]The University of New South Wales, Sydney, Australia
[‡]Australia E-Health Research Centre
{wangy,lxue,linw,zhangw}@cse.unsw.edu.au, qing.zhang@csiro.au

ABSTRACT

Often, a data object described by many features can be decomposed as multi-modalities, which always provide complementary information to each other. In this paper, we study subspace clustering for multi-modal data by effectively exploiting data correlation consensus across modalities, while keeping individual modalities well encapsulated. Our technique can yield a more ideal data similarity matrix, which encodes strong data correlations for the cross-modal data objects in the same subspace.

To these ends, we propose a novel angular based regularizer coupled with our objective function, which is aided by trace lasso and minimized to yield sparse representation vectors encoding data correlations in multiple modalities. As a result, the sparse code vectors of the same cross-modal data have small angular difference so as to achieve the data correlation consensus simultaneously. This can generate a compatible data similarity matrix for multi-modal data. The final subspace clustering result is obtained by applying spectral clustering on such data similarity matrix.

Categories and Subject Descriptors

I.5.3 [**Clustering**]: Similarity measures

Keywords

Correlation Consensus; Angular based Regularizer; Multi-modal Data

1. INTRODUCTION

The nature of visual data, in practice, is multi-modality, *e.g.,* an image can be described by a color modality or a shape modality. These multiple modalities often encode compatible and complementary information, which naturally motivates one to leverage them to obtain better performance than the result yielded by single modality, or simply concatenating all modalities into a monolithic one.

MM'14, November 03–07 2014, Orlando, FL, USA.

ACM 978-1-4503-3063-3/14/11 ...$15.00.
http://dx.doi.org/10.1145/2647868.2654999.

Among the studies of leveraging complementary information from multi-modal data, unsupervised clustering on multi-modal data is the most practical, due to the largely available unlabeled multi-modal data in real life. As shown in [6, 14, 13], data objects with high dimensions are always drawn from a union of subspaces, motivating numerous approaches of subspace clustering for multi-modal data.

Kumar *et al.* [10] presented a co-training [4, 16] based method to achieve a compatible multi-view similarity matrix in eigen-subspaces spanned by Laplacian matrix, such that the similarity matrix in one view is affected by that from another view. However, they simply calculate similarity matrix in a K-nearest neighbors manner, which degrades the performance when data points are nearby the intersection of two distinction subspaces. That is, the neighborhoods of a data point may cover the data points from different subspaces. The same problem holds in [11] as well. Chaudhuri *et al.* [5] proposed to project the multi-modal data into one common subspace, then apply the clustering algorithm, *e.g.,* K-means, to yield the subspace clustering results. Such method, however, is sensitive to data initialization. Specifically, it requires the data initialization strictly follow the Gaussian distribution while keeping different groups of data objects separated. Besides, the number of dimension for the projected subspace needs to be known in advance. Matrix factorization is also utilized to perform subspace clustering for multi-modal data, such as [1, 9], where the essential idea is to first concatenate the features of heterogeneous modalities into a single-long feature representation, then non-negative matrix factorization is applied to get subspace clustering results. One limitation of such one-combo-fits-all strategy is that the data correlation information in each original view-specific feature space is not well exploited. To combat this limitation, Liu *et al.* [12] proposed a joint non-negative matrix factorization paradigm on each individual modality to compute distinct coefficient matrices, which are then regularized towards a common consensus that represents the clustering structure shared by all modalities. This method, however, suffers from the drawback that the dimension number of latent reduced subspace needs to be manually parameterized, rather than automatically.

In light of above raised limitations, we propose a novel technique to address these limitations. The basic idea is we learn a data similarity matrix by aggregating data correlations from multiple modalities, and these correlations are consensus to each other whilst their individualities are well-encapsulated. Then the spectral clustering is applied to this data similarity matrix to obtain the subspace clustering re-

sults. Specifically, we propose a novel angular based regularizer to regularize our objective function, which is facilitated by trace lasso through its fantastic grouping effect in subspaces [13]. The objective function is minimized to sort out data correlations, encoded as the sparse representation vectors for data points under each modality. Meanwhile, sparse code vectors towards the same data object across modalities exhibit a small-angled difference.

The intuition of minimizing the angular based regularizer comes from the fact that if two distinct sparse code vectors encode the similar data correlations, then there will be a small angle between them. Now a natural question arises: whether the two distinct sparse codes of the same data object across modalities are similar? The answer is "yes". A common assumption holds in most multi-view based clustering methods, *e.g.,* [12], is that the same data set under different modalities should characterize the similar cluster structure, and clustering is largely determined by the data correlations representing similarities under various modalities. This observation has been validated by extensive experiments on real-world image datasets.

Our major contributions are summarized as follows.

- To the best of our knowledge, we are the first to consider the application of trace lasso on subspace clustering (See Section 2.2) for multi-modal data.
- To exploit the data correlation consensus across modalities, we propose a novel angular based regularizer over the data sparse codes in multi-modalities (See Section 2.3). The objective function is minimized and regularized by the new regularizer to achieve the goal.
- An efficient optimizing process is designed to optimize the proposed objective function (See Section 3). Experiments on real-world image datasets are conducted to demonstrate the effectiveness of our algorithm.

2. PROPOSED TECHNIQUE

2.1 Problem Definition

Let $X = \{x_k\}_{k=1}^n$ be data point set with n data points. Suppose each data object has V modalities, without loss of generality, for i^{th} modality, $X^i = \{x_k^i\}_{k=1}^n$, $(i = 1, \ldots, V)$, x_k^i is the representation of x_k under i^{th} modality, and s_k^i is the sparse representation vector of x_k^i based on X^i. The trace lasso [8] is defined as $\|X^i Diag(s_k^i)\|_*$, where $Diag(s_k^i)$ represents diagonal matrix with all diagonal elements corresponding to the entries of s_k^i, $\|A\|_*$ denotes the nuclear norm (the sum of all the singular value) of a matrix A. The norms of $\|a\|_1$ and $\|a\|_2$ denote the ℓ_1-norm and ℓ_2-norm of a vector a.

With the above notations defined, we aim to learn the sparse representations s_k^i of each x_k^i based on X^i $(i = 1, \ldots, V)$. These sparse codes are then used to construct a compatible similarity matrix W for multi-modal data. Finally, the spectral clustering is applied upon W to perform subspace clustering for multi-modal data.

2.2 Modeling data correlations in single modality

We propose to deploy trace lasso to model data correlations. That is, high correlations hold for the data points from the same subspace, no (weak) correlations for those from distinct subspaces. As shown in [13], trace lasso is more adaptive than ℓ_1 or ℓ_2 norms, and it is equal to ℓ_1-norm or ℓ_2-norm if data points are uncorrelated (orthogonal) or highly correlated. Thereby, we have $\|s_k^i\|_2 \leqslant \|X^i Diag(s_k^i)\|_* \leqslant \|s_k^i\|_1$.The sparse representation s_k^i of x_k^i can well reflects the correlations between x_k^i and other data points under i^{th} modality. Thus, we formulate the problem of learning sparse code vectors of each data point in i^{th} modality as:

$$\min_{s_k^i} \frac{1}{2}\|x_k^i - X_k^i s_k^i\|_2^2 + \lambda \|X_k^i Diag(s_k^i)\|_*, \tag{1}$$

where X_k^i represents the data set with x_k^i excluded. The parameter λ controls the effect from trace lasso term. Through trace lasso, s_k^i is composed of approximately the equal coefficients yet large weights on a few data points, implying their strong correlations with x_k^i. Meanwhile, the coefficients of data having no (weak) correlation with x_k^i are set to be 0. The claims also apply on $j^{th}(j \neq i)$ modality.

Eq. (1) is made on single modality. How to extend it to the context of multi-modalities? The answer, however, is non-trivial partially due to the fact that many multi-view learning methods, *e.g.,* [15], are aided by common label spaces across modalities. Hence, without label information, it becomes more challenging to exploit their consensus property shared by modalities. To this end, below, we propose to exploit correlation consensus across modalities.

2.3 Exploiting correlation consensus in multiple modalities

Inspired by multi-view clustering [12]: the true underlying clustering would assign corresponding data objects across modalities into the same cluster, we propose to exploit the data correlation consensus w.r.t. modalities, which critically determines the subspace clustering on multi-modal data.

Basic idea: Why angular based similarity? We are expected to effectively exploit the correlations across modalities but keep their individuality well-encapsulated. The question is how to quantify the similarity between the data correlations from different modalities? One may directly use a distance metric, *e.g.,* Euclidean distance, to measure data objects across modalities. It is, however, infeasible because data points in different views might not be comparable in the same scale. Fortunately, the sparse code of one point, *e.g.,* s_k^i for x_k^i, can well reflect the correlations between x_k^i and other data points under i^{th} modality. Another nice property is that sparse codes for different modalities have the same dimension. These properties drive us to quantify the **angular based similarity** over sparse codes of multi-modal data. This is equivalent to quantifying the data correlations in X under different modalities. This enables us to propose a novel regularizer term, $\pi(s_k)$, to encode the cosine similarity among s_k^i regarding x_k^i for multi-modality.

Mathematically, $\pi(s_k)$ can be defined as

$$\pi(s_k) = -\sum_{i,j\neq i}^{V} \frac{(s_k^i)^T \cdot s_k^j}{\|s_k^i\|_2 \|s_k^j\|_2}. \tag{2}$$

The objective function in Eq. (3) needs to be minimized, thus, we add minus sign "−" in Eq. (2).

Combining Eq. (1) with Eq. (2), we confirm our objective function as

$$\min_{s_k^i} \sum_i^V \beta_i \left(\frac{1}{2} \|x_k^i - X_k^i s_k^i\|_2^2 + \lambda \|X_k^i Diag(s_k^i)\|_* \right) + \gamma \pi(s_k), \tag{3}$$

where the parameter β_i controls the contribution from i^{th} modality and $\sum_{i=1}^V \beta_i = 1$. γ is a weight parameter on $\pi(s_k)$, which is able to correlate modalities by essentially regulating the structural consensus over dictionaries. In summary, we aim to learn sparse codes s_k^i of x_k^i for each modality i by optimizing Eq. (3), which will be discussed in Section 3.

3. OPTIMIZATION STRATEGY

The difficulty of optimizing Eq. (3) lies in its non-joint-convex for M_k^i and s_k^i, along with non-smoothness for trace lasso. We alternatively optimize each variable by fixing others. We initialize each s_k^i by optimizing the Eq. (1) via Alternating Direction Method (ADM) in [13], by setting λ to be 0.15. We derive an equivalent variational formulation of the trace norm [2]. Assume $M \in \mathbb{R}^{n \times m}$, then the trace norm of M is equal to

$$\|M\|_* = \frac{1}{2} \inf_{S \succeq 0} tr(M^T S^{-1} M) + tr(S), \tag{4}$$

where the infimum is achieved when $S = (MM^T)^{1/2}$, then we recast Eq. (3) to be

$$\begin{aligned} &\min_{s_k^i} \inf_{M_k^i \succeq 0} \sum_i^V \frac{\beta_i}{2} \|x_k^i - X_k^i s_k^i\|_2^2 \\ &+ \frac{\lambda \beta_i}{2} \left(tr\left((S_k^i)^2 (X_k^i)^T (M_k^i)^{-1} X_k^i \right) + tr(M_k^i) \right) + \gamma \pi(s_k), \end{aligned} \tag{5}$$

where $S_k^i = Diag(s_k^i)$.

Updating M_k^i with fixed s_k^i. The optimization problem in Eq. (5) is convex in M_k^i. We conduct coordinate descent procedure to optimize M_k^i for each x_k^i. Given s_k^i, we have the closed form solution of M_k^i as

$$M_k^i = (X_k^i (S_k^i)^2 (X_k^i)^T)^{1/2} \tag{6}$$

Updating s_k^i with fixed M_k^i. We propose to adopt GIST [7] algorithm to optimize s_k^i, due to its effectiveness in addressing non-convex and non-smoothness optimization problems, and this strategy is **well converged** as proved by [7]. Denote by $F(s_k^i)$ the objective function over s_k^i, that is,

$$\begin{aligned} F(s_k^i) = &\frac{\beta_i}{2} \|x_k^i - X_k^i s_k^i\|_2^2 + \frac{\lambda \beta_i}{2} tr\left((S_k^i)^2 (X_k^i)^T (M_k^i)^{-1} X_k^i \right) \\ &- \gamma \sum_{l \neq i}^V \frac{(s_k^l)^T \cdot s_k^j}{\|s_k^l\|_2 \|s_k^j\|_2}. \end{aligned}$$

Let $s_k^i(j)$ be the j^{th} entry of $s_k^i \in \mathbb{R}^m$, and $s_k^i \geq 0$ (each entry is large or equal to 0). It is common to set $\|s_k^i\|_2 = 1$ for any i, then the subgradient of $F(s_k^i)$ at $s_k^i(j)$ is

$$\begin{aligned} \eta_j = &\frac{\partial F(s_k^i)}{\partial s_k^i(j)} = \beta_i X_k^i(\cdot, j)^T (x_k^i - \sum_{l \neq j} X_k^i(\cdot, l) s_k^i(l) - X_k^i(\cdot, j) s_k^i(j)) \\ &+ \lambda \beta_i s_k^i(j) K_i(j, j) - \gamma \sum_{l \neq i}^V s_k^l(j), \end{aligned}$$

where $K_i = (X_k^i)^T (M_k^i)^{-1} X_k^i$, $X_k^i(\cdot, j)$ denotes the j^{th} column of X_k^i. We assume the sub-gradient of $F(s_k^i)$ at point s_k^i can be computed as $g = (\eta_1, \ldots, \eta_m)$. Consider a series of updated sparse codes $s_k^i[t]$ and the associated subgradients $g[t]$ $(1 \leq t \leq p)$ at time stamp t, then we have

$$F(s_k^i[t]) \geq F_p(s_k^i[t]) = \max_{1 \leq t \leq p} \{F(s_k^i[t-1]) + (s_k^i[t] - s_k^i[t-1]) g[t-1]\}. \tag{7}$$

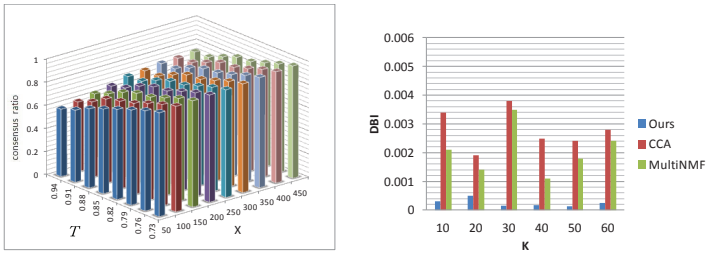
Given the previous prox-center point $s_k^i[t-1]$, it seeks the next potential candidate point by minimizing the piecewise linear lower bound:

$$s_k^i[t] = \underset{\|s_k^i\|=1, s_k^i \geq 0}{\operatorname{argmin}} \; F_p(s_k^i) + \frac{\omega_p}{2} \|s_k^i[t] - s_k^i[t-1]\|^2. \tag{8}$$

We stop the iteration and report the optimal s_k^i if the approximation error residual for s_k^i, measured by $F(s_k^i[t]) - F(s_k^i[t-1])$, is less than a predefined threshold ϵ (In our experiments, ϵ is set to be 10^{-2}). Otherwise, we conduct a line search between $s_k^i[t-1]$ and $s_k^i[t]$ to reproduce a new candidate and repeat the above process until convergence. We follow [3] to initialize and update the step size ω_p.

Remark. The convergence for optimizing Eq. (3) is determined by optimizing s_k^i, as M_k^i enjoys a closed form at each step. It is convergent for updating s_k^i due to the convergence property of GIST algorithm [7]. We calculate the similarity between k^{th} and l^{th} data objects under i^{th} modality, such as $W_i(k,l) = \frac{s_k^i(l) + s_l^i(k)}{2}$. The final affinity matrix regarding all modalities is calculated as $W = \sum_i^V W_i$, which is further utilized for subspace clustering.

4. EXPERIMENTAL RESULTS



(a) (b)

Figure 1: (a) The values of consensus ratios versus number of training data and thresholds for our method. (b) The evaluation on cluster qualities of different methods.

4.1 Datasets and Competitors

Datasets. Three real world image datasets are used in our experiment.

• **UCI Handwritten Digit Dataset** [1]: This handwritten digits (0-9) dataset consists of 2,000 examples. We construct two modalities, with first modality being the 76 Fourier coefficients and second modality being the 240 pixel averages in 2×3 windows.

• **CMU PIE Face Database** [2]: This dataset contains 68 subjects with 41,368 face images. Each image is 32×32, and we use four kinds of features as four modalities: LBP (256-dim), HOG (100-dim), and grey levels (128-dim).

• **PASCAL VOC 2010 Database**: This dataset contains 10,103 images from 20 classes. We adopt four types of

[1] http://archive.ics.uci.edu/ml/datasets.html
[2] http://www.ri.cmu.edu/projects/project_418.html

Method	Accuracy(%)			Normalized Mutual Information(%)		
	UCI	CMU-PIE	PASCAL	UCI	CMU-PIE	PASCAL
BSV-SSC	69.028±0.010	70.255±0.013	62.381±0.009	52.622±0.010	51.457±0.009	56.403±0.009
ConcatSSC	70.832±0.007	65.735±0.011	60.228±0.006	54.055±0.011	48.829±0.008	51.520±0.009
CCA	76.114±0.004	77.482±0.006	70.010±0.006	61.671±0.007	60.227±0.006	58.731±0.008
Co-train-SC	84.642±0.002	83.070±0.007	76.125±0.004	77.011±0.005	71.420±0.004	66.592±0.006
MultiNMF	88.014±0.003	86.807±0.004	80.227±0.007	80.257±0.010	77.561±0.011	69.882±0.011
Ours	**93.149±0.001**	**90.006±0.003**	**87.540±0.004**	**86.553±0.009**	**82.716±0.010**	**74.660±0.010**

Table 1: Clustering performance on three real image datasets.

features corresponding four modalities: color moments (255-dim), color histogram (64-dim), edge distribution (73-dim), and wavelet texture (128-dim).

We consider five algorithms as competitors in our experiments: (1) The best single view on Sparse Subspace Clustering (**BSV-SSC**), (2) Concatenating the features of each view on Sparse Subspace Clustering (**ConcatSSC**), (3) multi-view clustering via Canonical Correlation Analysis (**CCA**) [5], (4) Co-training based multi-view Spectral Clustering (**Co-train-SC**) [10], (5) Multi-view NMF (**MultiNMF**) [12]. Please refer to the section 1 for the discussions on [5], [10] and [12]. We omit the details here due to limited spaces. The clustering results are evaluated by validating the obtained label of each data point with the label provided by the dataset. Two widely adopted metrics, the accuracy (AC) and normalized mutual information (NMI) are used to measure the clustering performance. Please refer to [10] for detailed definitions. The sparsity parameter λ is set 0.15 in all experiments by cross-validation, β_i for each modality equally to be $\frac{1}{V}$. Parameter γ, in Eq. (3), controlling angular based regularization is set by grid search on $\{0.1, \ldots, 0.9\}$.

4.2 Clustering results

Table 1 shows the clustering performances of different algorithms on the three datasets. We can observe that our algorithm outperforms the second best counterpart **MultiNMF** by a margin of 5.8%(7.8%) on UCI, 3.7%(6.6%) on CMU-PIE, 9.1%(6.8%) on PASCAL, in terms of accuracy(NMI). One reason is that our method can automatically learn a good similarity matrix by aggregating consensus data correlations across modalities instead of manually setting the dimension number of reduced subspaces. Moreover, the improvement gain is significantly higher between our method and other alternatives, which are **CCA**, **Co-train-SC**, **ConcatSSC**, and **BSV-SSC**.

We demonstrate the stableness and robustness by comparing with **CCA**, and **MultiNMF**. Both need to manually set the dimension number of reduced subspaces, *i.e.,* number of clusters. To evaluate the quality of clusters $\{U_l\}_{l=1}^K$, we use Davies-Bouldin Index (DBI) to measure the uniqueness of clusters w.r.t. the unified similarity measure.

$$DBI(\{U_l\}_{l=1}^K) = \frac{1}{K}\sum_{i=1}^{K}\max_{j\neq i}\frac{d(c_i,c_j)}{\sigma_i+\sigma_j}, \qquad (9)$$

where c_x is the centroid of U_x, $d(c_i,c_j)$ is the similarity between c_i and c_j, σ_x is the average similarity of vertices in U_x to c_x. Fig.1 (b) shows the DBI comparison on CMU-PIE with various K values. Our method has the lowest DBI, while other methods have higher DBIs. This indicates that our parameter-free algorithm can discover true clusterings more robustly through adaptively correlating data consensus across modalities. As a comparison, manually set up the values of K makes **CCA** and **MultiNMF** sensitive to parameters and less effective in clustering quality.

4.3 Consensus study

To show the correlation consensus propertied by angular term $\pi(s_k)$, we conduct another experiment on PASCAL VOC 2010 wherein a training set X is composed of randomly selected images numbering from 50 to 450. The rest are taken as test samples. The consensus threshold is T, whose value is discretely set from 0.73 to 0.94. We employ the consensus ratio as the evaluation metric, defined as calculating the number of test samples whose values of Eq. (2) are larger or equal to T. The results are shown in Fig.1 (a). It can be seen that as more training samples get involved, we can have larger values of consensus ratio, implying learned sparse codes are more consensus. However, the consensus ratios naturally decrease when cosine values become higher, implying a more restrict consensus. Overall, our method shows a good performance even with a relatively small training set and large values of T.

5. CONCLUSIONS

In this paper, we propose a novel approach towards subspace clustering on multi-modal data. Unlike existing methods, our proposed method can well achieve the consensus of data correlations across modalities, encoded by learned sparse code vectors w.r.t. each modality. This method is able to automatically uncover true underlying clusters. Experimental results validate the effectiveness of our method.

6. REFERENCES

[1] Z. Akata, C. Thurau, and C. Bauckhage. Non-negative matrix factorization in multimodality data for segmentation and label prediction. In *ICCV Winter-workshop*, 2011.

[2] A. Argyriou, T.Evgeniou, and M. Pontil. Multi-task feature learning. In *NIPS*, 2007.

[3] J. R. BlRGE, L. Qi, and Z. Wei. Convergence analysis of some methods for minimizing a nonsmooth convex function. *Journal of optimization theory and applications*, 97:105–122, 1990.

[4] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, 1998.

[5] K. Chaudhuri, S. Kakade, K. Livescu, and K. Sridharan. Multi-view clustering via canonical correlation analysis. In *ICML*, 2009.

[6] E. Elhamifar and R. Vidal. Sparse subspace clustering. In *CVPR*, 2009.

[7] P. Gong, C. Zhang, Z. Lu, jianhua Z. Huang, and J. Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *ICML*, 2013.

[8] E. Grave, G. Obozinski, and F. Bach. Trace lasso: a trace norm regularization for correlated designs. In *NIPS*, 2011.

[9] D. Greene and P. Cunningham. A matrix factorization approach for integrating multiple data views. In *ECML/PKDD*, 2009.

[10] A. Kumar and H. D. III. A co-training approach for multi-view spectral clustering. In *ICML*, 2011.

[11] A. Kumar, P. Rai, and H. D. III. Co-regularized multi-view spectral clustering. In *NIPS*, 2011.

[12] J. Liu, C. Wang, J. Gao, and J. Han. Multi-view clustering via joint nonnegative matrix factorization. In *SDM*, 2013.

[13] C.-Y. Lu, J. Feng, Z. Lin, and S. Yan. Correlation adaptive subspace segmentation by trace lasso. In *ICCV*, 2013.

[14] Y. Pang, Z. Ji, P. Jing, and X. Li. Ranking graph embedding for learning to rerank. *IEEE Trans. Neural Netw. Learning Syst*, 24(8):1292–1303, 2013.

[15] H. Wang, F. Nie, H. Huang, and C. Ding. Heterogeneous visual feature fusion via sparse multimodal machine. In *CVPR*, 2013.

[16] Y. Wang, X. Lin, and Q. Zhang. Towards metric fusion on multi-view data: a cross-view based graph random walk approach. In *CIKM*, 2013.