Maximum Unfolded Embedding: Formulation, Solution, and Application for Image Clustering

Huan Wang[1]
[1] IE, Chinese University of Hong Kong, Hong Kong
hwang5@ie.cuhk.edu.hk

Shuicheng Yan[2], Thomas Huang[2]
[2] ECE, University of Illinois at Urbana Champaign, USA
{scyan,huang}@ifp.uiuc.edu

Xiaoou Tang [1,3]
[3] Microsoft Research Asia, Beijing, China
xitang@microsoft.com

ABSTRACT

In this paper, we present a novel spectral analysis algorithm for image clustering. First, the image manifold is embedded onto a low-dimensional feature space with dual objectives, i.e., maximizing the distances of faraway sample pairs meanwhile preserving the local manifold structure, which essentially results in a *Trace Ratio* optimization problem. Then an efficient iterative procedure is proposed to directly optimize the trace ratio and finally the clustering process is implemented on the derived low-dimensional embedding. Moreover, the linear approximation is also presented for handling the out-of-sample data. Experimental results show that our algorithm, referred to as Maximum Unfolded Embedding, brings an encouraging improvement in clustering accuracy over the state-of-the-art algorithms, such as K-Means, PCA-Kmeans, normalized cut [8], and Locality Preserving Clustering [13].

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval - Clustering; I.4.m [**Image Processing and Computer Vision**]: Miscellaneous - Image Clustering

General Terms

Algorithms, Performance, Theory

Keywords

Maximum Unfolded Embedding, Spectral Analysis, Image clustering

1. INTRODUCTION

With the rapid development of internet and storage hardware, larger and larger image repositories are being built. As a result, image search and retrieval are becoming more and more popular. Image clustering is a powerful high-level technique that helps analyze the image content of a large database [4].

The aim of image clustering is to find a mapping from the archive images to a set of label numbers. Traditional clustering methods, such as K-means [6] and Gaussian Mixture Model [6], directly operate on the original feature space. They often cannot produce satisfactory results due to the underlying complex data distribution, and many researches show that the image data may lie on a nonlinearly distributed manifold [13].

The spectral clustering algorithms [3][5] utilize a graph to characterize the relationship of the input data. By relaxing the cluster labels to the real value domain, they apply the efficient eigen-decomposition method for data embedding. The most popular spectral clustering algorithm is normalized cut [8], which has been widely used in various applications [9][12]. An extension of the normalized cut was proposed by Ng et al. [2], which deals with the clustering problem in a more effective way.

The spectral embedding algorithms Laplacian Eigenmap (LE) [3] and Locality Preserving Projections (LPP) [7] have shown their effectiveness for image clustering [13]. The disadvantage of LE and LPP is that their objective functions for embedding cannot ensure the separability of those faraway point pairs in the derived low-dimensional feature space, which may greatly degrade the performance of the consequent clustering process.

To overcome the above disadvantage, we propose a novel spectral embedding algorithm called Maximum Unfolded Embedding (MUE), and consequently a new clustering algorithm, Maximum Unfolded Clustering (MUC), is presented by utilizing K-means on the derived embedding. The aim of MUE is two-fold: one is to preserve the local manifold structure, and the other is to maximize the distance between the faraway sample pairs. These dual objectives are integrated to form a Trace Ratio optimization problem. We then propose a novel iterative procedure to directly optimize the trace ratio, instead of transforming the Trace Ratio optimization problem into a Ratio Trace optimization problem which has the advantage of a closed form solution yet sacrifices the potential clustering capability. Moreover, the linear extension of MUE is also presented so as to map the out-of-sample data into the low-dimensional feature space.

2. MAXIMUM UNFOLDED EMBEDDING

An image set is often considered lying on a compact nonlinear manifold embedded in a high-dimensional Euclidean

MM'06, October 23–27, 2006, Santa Barbara, California, USA.
ACM 1-59593-447-2/06/0010.

space. Image clustering directly performed on the original image feature space may suffer from the curse of dimensionality, and it is desirable to conduct dimensionality reduction before the clustering process. In this work, we propose a novel dimensionality reduction algorithm called Maximum Unfolded Embedding (MUE), for image clustering by taking into account the underlying manifold structure of the image set.

For a given data set, we construct a weighted adjacency graph $G^a = (V, W^a)$ by connecting each point with its k_n nearest neighboring points, and a weighted separation Graph $G^s = (V, W^s)$ that indicates the faraway point pairs measured in the original feature space and connects each point with its k_f farthest contrary points. Without loss of generality, we assume these two graphs are both connected. The element of the weight matrix equals to 1 when there is an edge connecting the corresponding two vertices; otherwise 0.

The aim of MUE is to find a low-dimensional representation such that the connected vertices are close to one another and the faraway vertex pairs are faraway to each other. Let an $N \times m$ matrix $Y = [y_1, y_2, ..., y_m]$ denote such a low-dimensional representation (N denotes the sample number and m is the desired reduced feature dimension). The i-th row corresponds to the i-th dimension of the mapped data. To embrace the above two objectives, the following optimization problem is posed in MUE:

$$Y = \underset{Y^TY=I}{\text{argmax}} \frac{\sum_{ij} \|y^{(i)} - y^{(j)}\|^2 W_{ij}^s}{\sum_{ij} \|y^{(i)} - y^{(j)}\|^2 W_{ij}^a}. \tag{1}$$

Here $y^{(i)} = [y_1^{(i)}, ..., y_m^{(i)}]$ is the m dimensional representation of the i-th vertex. It is easy to have the following theorem:

Theorem 1. The problem defined in (1) is equal to a Trace Ratio optimization problem:

$$Y = \underset{Y^TY=I}{\text{argmax}} \frac{Tr(Y^T L^s Y)}{Tr(Y^T L^a Y)}, \tag{2}$$

where L^a and L^s are the Laplacian matrices of the adjacency graph and separation graph respectively. That is, $L^a = D^a - W^a$ and $L^s = D^s - W^s$, where D^a and D^s are both diagonal and their diagonal elements are the row sums of the matrices W^a and W^s respectively.

3. OPTIMIZE TRACE RATIO

In this section we propose an iterative algorithm to optimize the Trace Ratio problem defined in (2), and the detailed algorithm is listed in Algorithm 1. The algorithmic convergence is proved in the theorem 2.

Theorem-2. For the Algorithm 1, we have:

$$\lambda^{n+1} \geq \lambda^n.$$

Proof. Denote $g_n(V) = Tr(V^T(L_p^s - \lambda^n L_p^a)V)$, then

$$g_n(V^{n-1}) = 0.$$

Moreover, from the constraint of Trace Ratio problem that $V^TV = I_m$, we have

$$\sup_{V^TV=I_m} g_n(V) = \sum_{k=1}^{m} \tau_k^n = g_n(V^n).$$

Then, $Tr[{V^n}^T(L_p^s - \lambda_n L_p^a)V^n] \geq 0.$

Algorithm 1 : Iterative Solution to Trace Ratio

```
1: Remove the Null Space. Conduct singular value decomposition $L^a = P\Lambda P^T$, where the diagonal elements of $\Lambda$ are all positive and $P \in \mathbb{R}^{N \times m'}$. Then set matrix $L_p^s = P^T L^s P$ and $L_p^a = P^T L^a P$.
2: Initialize $V^0$ as arbitrary columnly orthogonal m' × m matrix;
3: For n=1,2,..., $N_{max}$, Do
    1. Compute the trace ratio value $\lambda^n$ from the projection matrix $V^{n-1}$:
       $$\lambda^n = \frac{Tr[{V^{n-1}}^T L_p^s V^{n-1}]}{Tr[{V^{n-1}}^T L_p^a V^{n-1}]}. \tag{3}$$
    2. Construct the trace difference problem as:
       $$V^n = \arg\max_V Tr[V^T(L_p^s - \lambda^n L_p^a)V]. \tag{4}$$
    3. Solve the trace difference problem using the eigenvalue decomposition method
       $$(L_p^s - \lambda^n L_p^a)V_k^n = \tau_k^n V_k^n, \tag{5}$$
       where $\tau_k^n$ is the $k$-th largest eigenvalue of $(L_p^s - \lambda^n L_p^a)$ with the corresponding eigenvector $V_k^n$.
    4. Set $V^n = [V_1^n, V_2^n, \ldots, V_m^n]$.
    5. If $\|V^n - V^{n-1}\| < \sqrt{m'm}\,\varepsilon$ ($\varepsilon$ is set as $10^{-4}$ in this work), then break.
4: Output $Y = PV^n$.
```

As the matrix L_p^a is positive definite, we have

$$\frac{Tr({V^n}^T L_p^s V^n)}{Tr({V^n}^T L_p^a V^n))} \geq \lambda^n,$$

that is, $\lambda^{n+1} \geq \lambda^n$.

Clustering: Based on the derived low-dimensional representation, the traditional K-means algorithm can be used to conduct further clustering process in a much faster manner.

4. ANOTHER JUSTIFICATION: MAXIMUM VARIANCE EMBEDDING

To preserve local information, LE computes a nonlinear embedding with spectral decomposition method, which has shown to be an approximation of the Laplace Beltrami operator on the manifold [3]. Merely minimizing the objective function $y^T L^a y$ in LE may encourage consistent output for the neighboring samples in the input space.

Compared with LE, the objective function of MUE exerts additional power on the faraway point pairs and thus prevents the mapping from collapsing into one point. A special case is that the separation matrix W^s can be constructed by utilizing k_f=N-1-k_n farthest points for each sample respectively, and then the objective function becomes

$$Y = \underset{Y}{\text{argmax}} \frac{\sum_{ij} \|y^{(i)} - y^{(j)}\|^2 \bar{W}_{ij}^a}{\sum_{ij} \|y^{(i)} - y^{(j)}\|^2 W_{ij}^a}, \tag{6}$$

with $\bar{W}_{ij}^a = 1 - W_{ij}^a$.

Summing up the numerator and the denominator, we can rewrite the objective function (1) as:

$$\frac{\sum_{ij}\|y^{(i)}-y^{(j)}\|^2}{\sum_{ij}\|y^{(i)}-y^{(j)}\|^2 W^a_{ij}} = \frac{Tr(Y^T L^{cp} Y)}{Tr(Y^T L^a Y)}, \tag{7}$$

where L^{cp} is the Laplacian matrix of the Binary Complete Graph which connects all the samples with equal weights. It means that the objective function is to maximize the global variance (numerator) and preserve the local characteristic (denominator) simultaneously. In this case, the proposed algorithm has the property of maximum variance, and hence may be justified as a Maximum Variance Embedding.

5. LINEAR EXTENSION: MAXIMUM UNFOLDED PROJECTION

To provide the low-dimensional representation for the out-of-sample data, a simple but effective way is to assume that there exists a linear relationship between the original data and the embedding, i.e., $Y^T = U^T X$, where the i-th column vector of X is x_i. With the constraint of column orthogonality on U, the optimal embedding for the objective function (1) is then transformed into the following trace ratio optimization problem:

$$U = \max_{U^T U = I} \frac{Tr(U^T X L^s X^T U)}{Tr(U^T X L^a X^T U)}. \tag{8}$$

It can also be solved with the iterative procedure discussed above.

6. EXPERIMENTS

In this section we present a set of experiments to verify the effectiveness of our proposed algorithms for image embedding and clustering. For comparison, we also report the experimental results from K-means, PCA Kmeans, normalized cut (NCut) and Locality Preserving Clustering (LPC).

6.1 Data set

The algorithmic performance is evaluated on the general-purpose image database, a subset of COREL [1], which contains 10,000 photo images from 79 categories. The image number for each cluster varies from 100 to 300.

We make a combination of the Color Histogram and Color Texture Moment (CTM) [11] for image description. As an extension to color moments that characterize the color texture distribution of the image, CTM provides a 48 dimensional feature vector. For Color Histogram, we set the bin number in HSI (Hue, Saturation and Intensity) space as $4 \times 4 \times 4$. Thus for each image we obtain a 112 dimensional feature vector, which is normalized in the preprocessing step so that the norm of each vector is 1.

6.2 Evaluation Metric

We utilize the normalized mutual information as the evaluation metric for clustering accuracy. Mutual information, as a measure for the mutual dependence of two variables in information theory, is defined as:

$$I(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log \frac{p(x,y)}{p(x)p(y)}.$$

An advantage of the mutual information is that it stays invariant when there exist permutations between the clustering results and the ground truth labels. The normalized mutual information ($\bar{I}$) is defined as

$$\bar{I}(X,Y) = \frac{I(X,Y)}{max(H(X),H(Y))}.$$

Where $H(X)$ and $H(Y)$ are the entropy of the set X and Y respectively. It is clear that the range of $\bar{I}$ is [0,1].

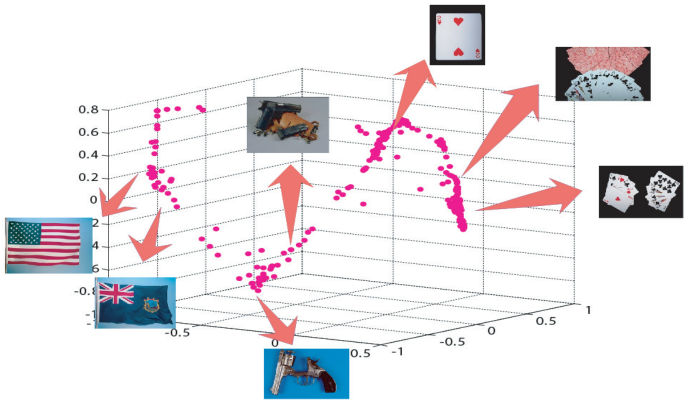

Figure 1: Embedding Visualization of 3 Clusters

Table 1: Clustering accuracies of MUC, NCut, PCA-Kmeans and K-Means in the cases with different cluster numbers.

	MUC	NCut	LPC	PCA-Km.	K-Means
k=2	0.9022	0.8337	0.7943	0.7367	0.7367
k=3	0.5511	0.5407	0.4859	0.3815	0.3755
k=4	0.5535	0.4744	0.5026	0.3561	0.3481
k=5	0.4450	0.3960	0.4132	0.3315	0.3447
k=6	0.3730	0.3462	0.3447	0.3103	0.3075
k=7	0.3382	0.3016	0.2798	0.2760	0.2748
k=8	0.3286	0.3137	0.2676	0.2976	0.2739
k=9	0.3432	0.3310	0.2646	0.2784	0.2818
k=10	0.3381	0.3125	0.2851	0.2766	0.2620

6.3 Performance Comparison

We vary the cluster number k from 2 to 10 and altogether five algorithms are compared and evaluated: MUC, NCut, LPC, PCA Kmeans and K-means. All these algorithms utilize the K-means method for final clustering, either on the feature domain or on the spectral domain. To overcome the influence of the local convergence, we perform the K-means process 20 times and report the best performance. For PCA-Kmeans, Principal Component Analysis (PCA) [10] is designed to retain 95% of the total energy and the algorithm proposed by Ng et al.[2] is adopted for NCut. We make an exhaustive search over the reduced feature dimensions and report the best performance. The comparison performances are demonstrated in Table 1. From these results, we can see that the proposed MUC algorithm consistently obtains the highest clustering accuracy, followed by NCut and LPC. The accuracy of PCA Kmeans is lower; yet is still higher than that of direct K-Means on original features in most cases, since certain amount of noise has been removed by PCA. We can also observe a clear gap between the spectral clustering algorithms and the PCA Kmeans. As the cluster number increases, the performance of LPC approaches that of the K-Means, mainly because the linear separability decreases with the increase of the cluster number; while MUC does not have this limitation. For a better understanding of MUE, the embedding visualization is displayed in Figure 1 and some clustering results are shown in Figure 2.

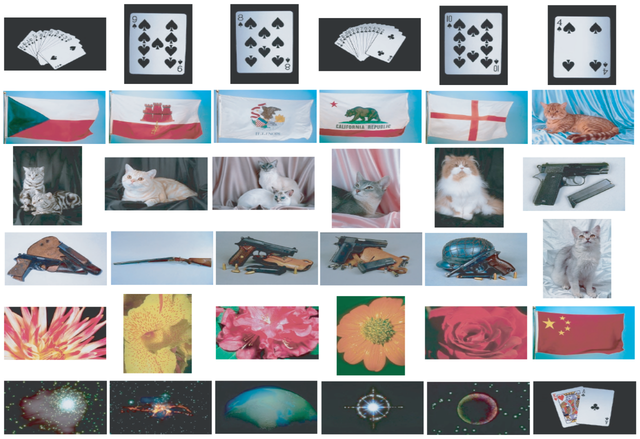

Figure 2: Some clustering results from 6 clusters

6.4 Generalization Capability

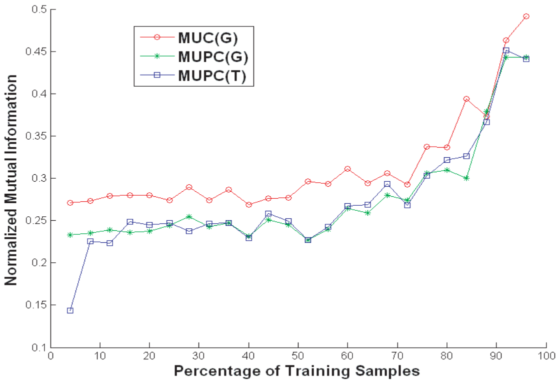

Figure 3: Generalization capability in the cases with different percentages of training samples. The cluster number is 10.

As described in Section 5, MUE can be linearized to construct the Maximum Unfolded Projections (MUP), and correspondingly we call the new clustering algorithm as MUP Clustering (MUPC). One merit of MUP is that the linearization process facilitates us with the generalization capability from training set to testing set.

Under the assumption that the training set and the testing set share similar data distributions, the projection derived from the training set can be extended to the testing set, thus we need not implement the Maximum Unfolded Embedding on the entire data set.

The second experiment is designed to demonstrate the generalization capability of the proposed algorithm. The data set contains 1000 images. We vary the percentage of training samples and evaluate the performance of MUPC on different training sets. For a comparison, the clustering result from MUC is also presented. Two configurations are adopted in this experiment. One is that we implement the algorithm on the entire data set and test the performance on the testing data set, which is denoted by MUPC (G) and MUC (G). The other configuration, MUPC (T), is as follows: first the projection is learned from the training set; then the embedding projection is performed on the testing set and K-means clustering algorithm is carried out; finally the normalized mutual information is calculated and compared among different configurations on the testing set. Figure 3 displayed the generalization results.

It is observed that as the percentage of training samples increases, the performance on the testing set is improved. We also notice that 30% of the total samples will be enough for the MUPC to produce a satisfying clustering accuracy and MUPC (G) can achieve comparable results as the MUC (G) with a much lower computational cost.

7. CONCLUSION

To maximize the distance of faraway sample pairs and simultaneously preserve the locality of the image manifold, a novel spectral embedding algorithm, Maximum Unfolded Embedding, was formulated. Then, an iterative procedure was proposed to directly optimize the Trace Ratio form of objective function. Consequently, a new spectral clustering algorithm was finalized by integrating the derived embedding and the K-Means approach. Extensive experiments verified the effectiveness of the proposed algorithm and its linearized version.

8. ACKNOWLEDGEMENT

The work described in this paper was fully supported by grants from the Research Grants Council of the Hong Kong Special Administrative Region.

9. REFERENCES

[1] http://www.corel.com, 2005.

[2] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, 2001.

[3] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 2003.

[4] Y. Chen, J. Z. Wang, and R. Krovetz. Content-based image retrieval by clustering. In *Multimedia Information Retrieval*, 2003.

[5] I. S. Dhillon, Y. Guan, and B. Kulis. Kernel k-means: spectral clustering and normalized cuts. In *KDD*, 2004.

[6] G.J.McLachlan and K.E. Basford. Mixture models: Inference and applications to clustering. 1988.

[7] X. He and P. Niyogi. Locality preserving projections. In *NIPS*, 2003.

[8] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE TPAMI*, 2000.

[9] J. Odobez, D. G. Perez, and M. Guillemot. Spectral Structuring of Home Videos. In *CIVR*, 2003.

[10] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd Ed).*

[11] H. Yu, M. Li, H. Zhang, and J. Feng. Color texture moment for content-based image retrieval. In *ICIP*, 2002.

[12] Z. Rasheed and M. Shah. A graph theoretic approach for scene detection in produced videos. In *SIGIR*, 2003.

[13] X. Zheng, D. Cai, X. He, W. Ma, and X. Lin. Locality preserving clustering for image database. In *ACM Multimedia*, 2004.