Image Analysis with Nonlinear Adaptive Dimension Reduction

Xinlei Chen Xinquan Qu Zijian Li
State Key Lab of CAD&CG, College of Computer Science
Zhejiang University, Hangzhou, 310027, China
{endernewton,04dzyqqxq,lizijian9630}@gmail.com

ABSTRACT

In multimedia applications, dimension reduction is essential to the effectiveness and efficiency of an algorithm due to the *curse of dimensionality*. Recently, its adaptive variants have received considerable attention in unsupervised learning since a single pass without label information often fails to guarantee an optimal representation, especially when the parameters are not set properly. However, most such methods are basically linear, therefore unable to consider the geometrical structure of the data space. In this paper, we propose a novel algorithm called *Nonlinear Adaptive Dimension Reduction* (NADR), which adaptively learns the optimal low-dimensional coordinates that preserve the intrinsic geometric structure of the original data. Moreover, the incorporation of K-means enables NADR to be a powerful alternative for cluster analysis. Experiments on benchmark image data sets illustrate that NADR outperforms the state-of-the-art adaptive dimension reduction methods.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Algorithms, Experimentation, Performance

Keywords

Dimension Reduction, Adaptiveness, Unsupervised Learning, Graph Laplacian

1. INTRODUCTION

In many application domains, the dimension of data to be processed is usually very high. For instance, in multimedia applications, each image or video is typically linearized into a vector, whose dimension easily scales up to thousands or even millions. However, dealing with high-dimensional data is challenging for the redundancy or irrelevancy within features and the prohibitive computational cost, which is often termed the *curse of dimensionality* [9].

To handle this problem, many unsupervised dimension reduction algorithms have been proposed to serve as a preprocessing step. Classical ones such as PCA [9] have been widely applied nowadays. Another category of algorithms is based on matrix factorization, such as NMF [10]. More recently, researchers have realized that in many situations the data points lie on a low-dimensional manifold embedded in the feature space, and empirical study have shown that Nonlinear dimension reduction is more powerful to preserve local geometric information in these situations [1]. In fact, one of the most popular clustering algorithms, *Normalized Cut* (NCut) [12], is essentially a combination of the *Laplacian Eigenmaps* (LE) [1] and K-means clustering. However, LE and other Nonlinear approach are entirely based on the graph, which is artificially defined in advance and does not necessarily benefit the follow-up steps, especially when the parameters are not set properly.

On the other hand, a new category of dimension reduction [5, 6, 7, 8, 11, 13, 14] that emphasizes the adaptiveness has received considerable attention in the last decades. Essentially, it integrates supervised dimension reduction and clustering together into a joint framework, and the low-dimensional features are adaptively learned. As a result, it often yields a better representation for the original data. A more detailed discussion of them will be found in the next section. However, most of the methods are basically linear and thus fail to consider the geometrical structure of the data space, which is critical in clustering and classification problems.

To overcome this limitation, we propose the *Nonlinear Adaptive Dimension Reduction* (NADR). NADR integrates nonlinear dimension reduction and the canonical K-means clustering into a joint process, in which the geometrical information is initially encoded in a nearest neighbor graph and is updated in every iteration. The algorithm benefits from both LE, which emphasizes the natural clusters in the data [1], and the inherited "adaptiveness" which serves as an automatic information filter for a better graph. These merits lead to a more optimized data representation, together with more accurate cluster assignments.

The rest of this paper is organized as follows. In Section 2, we give a brief review of the related works. Section 3 introduces our algorithm. The experimental results are shown in Section 4. Finally, we give concluding remarks in Section 5.

ICIMCS '11 August 5-7, 2011, Chengdu, Sichuan, China

2. RELATED WORKS

In this section, we review some of the previous works closely related to ours. First, several state-of-the-art adaptive dimension reduction methods are introduced, followed by a brief profile of Laplacian eigenmaps.

2.1 Adaptive Dimension Reduction

The central idea of the "adaptiveness" is to integrate supervised dimension reduction and clustering into a joint framework. Specifically, clustering generates class labels for supervised dimension reduction, while dimension reduction finds the discriminative low-dimensional representation for clustering. The major advantage of such algorithms lies in its ability to dynamically re-adjust the space and label assignment for global optimality, while still effectively avoid the curse of dimensionality.

Early works on this idea includes *adaptive dimension reduction* (ADM) [6] and *adaptive subspace iteration* (ASI) [11]. Given label information, ADM performs a singular value decomposition on the centroids in the original space to obtain a new subspace direction. On the other hand, given data points $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \in \mathbb{R}^{p\times n}$, ASI aims to optimize the objective function:

$$\min_{\hat{C},H,U} \frac{1}{2}\|X^T U - H\hat{C}^T\|_F^2, \tag{1}$$

where $\hat{C} = \{\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_k\} \in \mathbb{R}^{d\times k}$ are centroids in the subspace, $H \in \mathbb{R}^{n\times k}$ is the indicator matrix ($H_{ij} = 1$ if and only if $\mathbf{x}_i \in \mathcal{C}_j$), and $U \in \mathbb{R}^{p\times d}$ is the projection matrix. Formulated as a matrix factorization problem, the iterative algorithm essentially co-clusters samples and features [11]. In [5], *discriminative cluster analysis* (DCA) was proposed to solve another optimization problem:

$$\max_{H,U} \mathrm{Tr}(U^T X X^T U)^{-1}(U^T X H (H^T H)^{-1} H^T X^T U). \tag{2}$$

Afterwards, [7] proposed LDA-Km. It combines LDA [9] and K-means to adaptively select the most discriminative subspace for clustering. They also showed the intimate relationship between LDA-Km, ASI, ADM and DCA, which all seeks to maximize the between-cluster scatter S_b or minimize the within-cluster scatter S_w with respect to total scatter S_t [7] when performing dimension reduction. Then in [14], the authors gave a unified formulation of discriminative clustering methods [13, 7, 5] with the regularization technique on S_t, i.e., replacing $S_t = XX^T$ with $\hat{S}_t = S_t + \alpha I_p$ in Eq.(2). The optimal U can be obtained through an alternative trace optimization problem (Dis-Km),

$$\max_{V} \mathrm{Tr}(V^T(I_n - (I_n + \frac{1}{\alpha}X^T X)^{-1})V). \tag{3}$$

Here $V = H(H^T H)^{-1/2} \in \mathbb{R}^{n\times k}$ is defined as the weighted indicator matrix. Note that in Dis-Km, the adaptiveness is implicit: there's no explicit computation or updates on the projection matrix U.

Until now, our discussion of adaptive dimension reduction is restricted to form the low-dimensional features as linear combinations of original ones. Most non-linear versions, e.g., [7, 14, 4] are kernel extensions of the above methods.

2.2 Laplacian Eigenmaps

The *Laplacian Eigenmaps* [1] finds the low-dimensional representation by minimizing the weighted sum of the squared distances between neighboring data points:

$$\mathcal{O}_{LE} = \sum_{i=1}^{n}\sum_{i'=1}^{n} W_{ii'}\|\mathbf{y}_i - \mathbf{y}_{i'}\|^2 = \mathrm{Tr}(YLY^T). \tag{4}$$

Here $Y = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\} \in \mathbb{R}^{d\times n}$ are the low-dimensional features mapped from the original data points. $L = D - W$ is the graph Laplacian [1], $W_{ii'}$ is the similarity between $\mathbf{x}_i$ and $\mathbf{x}_{i'}$, and $D_{ii} = \sum_{i'=1}^{n} W_{ii'}$ is the degree of $\mathbf{x}_i$. The construction of similarity matrix W is critical in LE: numerous alternatives are available, which all model the local geometric structure of the data through K-NN or ϵ-NN scheme [2]. Typical similarity includes $0-1$ weighting, heat kernel weighting, dot-product weighting [2], etc.

To recover the low-dimensional coordinates, the authors of [1] proposed the following optimization problem:

$$\begin{aligned} \min_{Y} \quad & \mathrm{Tr}(YLY^T) \\ s.t. \quad & YDY^T = I,\ YD\mathbf{1} = \mathbf{0}, \end{aligned} \tag{5}$$

which essentially maps the sample points onto a hyper-ball in the d-dimensional space. Applying Lagrange multiplier, Eq.(5) can be solved by the generalized eigen-decomposition

$$L\mathbf{y} = \lambda D\mathbf{y}, \tag{6}$$

where the d eigenvectors corresponding to the smallest non-zero eigenvalue are selected as the low-dimensional coordinates.

Algorithm 1 Nonlinear Adaptive Dimension Reduction

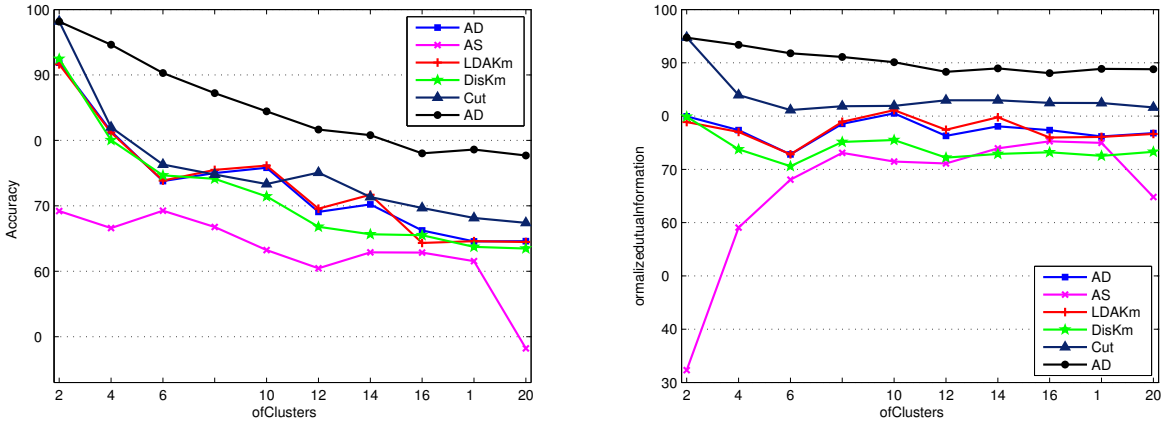
Input:
n data points $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \in \mathbb{R}^{p\times n}$;
Cluster number k;
Regularization parameter λ;
Output:
n coordinates $Y = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\} \in \mathbb{R}^{d\times n}$;
k clusters;
`Initialization step`: compute the affinity matrix $W \in \mathbb{R}^{n\times n}$ and degree matrix $D \in \mathbb{R}^{n\times n}$, set $\tilde{W} = I_n$;
repeat
`Updating coordinates`: solve the generalized eigenvalue problem according to Eq.(8), and compute the first d eigenvectors as data matrix $Y \in \mathbb{R}^{d\times n}$;
`Updating graph`: apply the standard K-means algorithm to Y to produce the new indicator matrix $H \in \mathbb{R}^{n\times k}$, and compute $\tilde{W}$ accordingly;
until H converges

3. OUR ALGORITHM

In this section, we introduce our *Nonlinear Adaptive Dimension Reduction* (NADR) algorithm. It incorporates the adaptiveness property into LE, and updates the coordinates based on the clustering assignments in each iteration. Then the initial embedding can drift to the optimal one.

3.1 Updating the Graph

Our major goal is to find the most discriminative coordinates in a unsupervised manner. From Section 2, we see that the low-dimensional representation of LE is totally dependent on the similarity graph W, therefore in order to achieve our objective, W should be able to adaptively modified. Assume the label of each sample is previously known,



(a) AC (b) NMI

Figure 1: Clustering performance on COIL20 with varying # of clusters

we introduce *supervised graph* [3], defined as

$$\tilde{W}_{ii'} = \begin{cases} 1/c_j, & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_{i'} \text{ belong to } \mathcal{C}_j \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

where $\mathcal{C}_j$ denotes the j-th cluster. It puts an edge between a pair of points only if they are belong to the same class, resulting in a blocked similarity matrix $\tilde{W}$. Since it's easy to verify that $X\tilde{W}X^T$ is precisely the between-class scatter in LDA [9], which aims to find the most discriminative linear projection of the data, it may well be the best choice for embedding. Besides, $\tilde{W}$ has two more impressive properties:

1. $\tilde{W} = H(H^TH)^{-1}H^T$ is closely related to the indicator matrix H.
2. The corresponding degree matrix $\tilde{D} = I_n$.

We build up the framework of adaptive dimension reduction based on this graph. We formulate it as:

$$\min_{Y,\tilde{L}} \quad \mathcal{O}_{NADR} = \mathrm{Tr}(Y(L + \lambda\tilde{L})Y^T) \quad (8)$$

$$s.t. \quad Y(D + \lambda\tilde{D})Y^T = I,\ Y(D + \lambda\tilde{D})\mathbf{1} = \mathbf{0}$$

with $\tilde{L} = I_n - \tilde{W}$ and a regularization parameter λ.

3.2 Optimization

The objective function in Eq.(8) can be minimized (local minima) by alternatively optimizing one of Y or $\tilde{L}$ while fixing the other. The two iterative steps are listed as:

NADR-1. Update coordinates while fixing the graph. We can perceive the graph as a normal similarity graph, and solve the generalized eigenvalue problem via standard Laplacian Eigenmaps.

NADR-2. Update graph while fixing the coordinates. In this case, we notice that neither the original graph Laplacian L in the objective function, nor the imposed constraint is reverent to $\tilde{L}$. Therefore, Eq.(8) can be simplified:

$$\begin{aligned} \mathrm{Tr}(\lambda Y\tilde{L}Y^T) &= \lambda\mathrm{Tr}(Y(I - H(H^TH)^{-1}H^T)Y^T) \\ &= \lambda\mathrm{Tr}(S_t - S_w). \end{aligned} \quad (9)$$

Eq.(9) is precisely the K-means clustering objective of Y [7], which indicates the K-means can be directly applied for minimization. Moreover, in this step, only the value of $\mathcal{O}_{NADR}$ is decreased, the constraint on Y is still preserved in that $\tilde{D} = I_n$ is left unchanged.

These two steps are executed iteratively in NADR. Since they monotonically decreases the objective function Eq.(8), the algorithm is bounded to converge. The whole procedure is described in Algorithm 1.

4. EXPERIMENTS

Previous studies show that adaptive dimension reduction methods are very powerful on clustering [7, 13]. And actually many of them are specially designed for a more discriminative unsupervised learning method. Therefore, we also evaluate our *Nonlinear Adaptive Dimension Reduction* (NADR) algorithm on clustering problems. And we used both the accuracy (AC) and normalized mutual information (NMI) metric [2] to measure the clustering result.

4.1 Data Sets and Comparing Methods

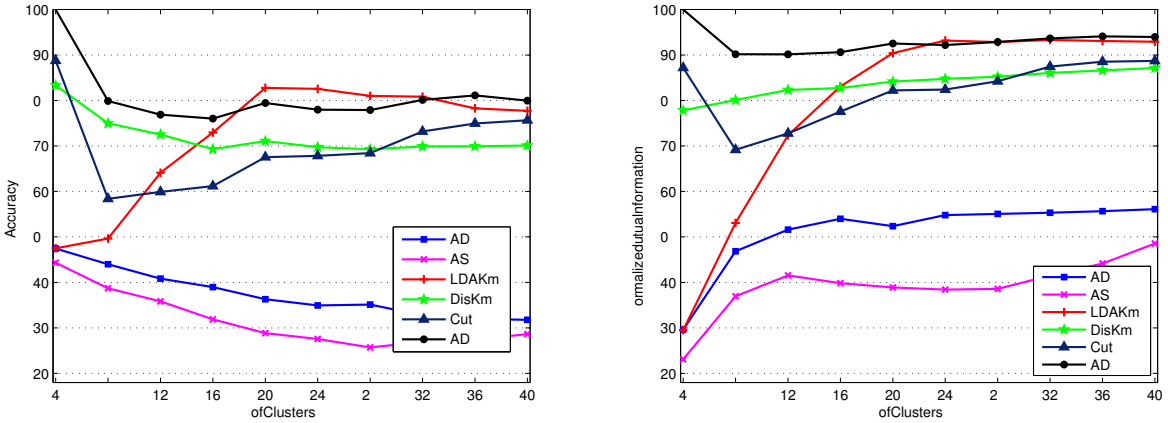
To sum up, we compared it with five algorithms on two real-world data sets. For adaptive methods, we chose four, namely ADM, ASI, LDA-Km and Dis-Km described in Section 2. DCA was left out since it was shown to be equivalent to ADM [7]. To further demonstrate the impact of adaptiveness in NADR, we also involved the non-adaptive version of NADR - normalized cut (NCut) [12].

There are two parameters in our NADR approach: the regularization parameter λ and the number of nearest neighbors p. Throughout our experiments, we empirically set $\lambda = 0.1$ and $p = 5$ (same in NCut). The reduced dimension is fixed to $k - 1$ for all methods where k is the cluster number.

Two image data sets are used. The first one is COIL20 image library[1]. It contains images of 20 objects viewed from varying angles. The second is CMU PIE face database[2]. It contains face images of 68 persons. We used the C27 near frontal pose and each person has 21 facial images under various illuminations. The images are scaled to 32×32 and no other pre-process was performed.

[1] `http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php`

[2] `http://www.ri.cmu.edu/projects/project_418.html`



(a) AC (b) NMI

Figure 2: Clustering performance on PIE with varying # of clusters

4.2 Results

Figure 1 and 2 show the evaluation results on the COIL20 and the PIE data set. The evaluations were conducted with the cluster numbers ranging from 2 to 10 for COIL20 and 4 to 40 for PIE data set. For each given cluster number k, 20 runs were conducted on different randomly chosen clusters. For fairness considerations, the initial assignment of data points was fixed for all competing methods. The results reveal a number of interesting points:

- LDA-Km and Dis-Km outperform other adaptive algorithms on PIE database while fail to get good performance on COIL20 database. Our NADR approach gets significantly better performance than them, particularly when k is small. This shows that by considering the intrinsic geometrical structure of the data, NADR can truly learn a better representation.
- Both NCut and NADR consider the geometrical structure of the data and performs well, and by imposing adaptiveness, NADR can significantly improve the clustering result.

5. CONCLUSION

In this paper, we have presented a novel algorithm for adaptive dimension reduction, called *Nonlinear Adaptive Dimension Reduction* (NADR). NADR adaptively learns a low-dimensional representation, and produces a cluster assignment simultaneously based on graph Laplacian. Experiments on image data sets illustrate that NADR outperforms state-of-the-art adaptive dimension reduction methods from the perspective of clustering performance.

6. REFERENCES

[1] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[2] D. Cai, X. He, and J. Han. Document clustering using locality preserving indexing. *IEEE Transactions on Knowledge and Data Engineering*, 17:1624–1637, 2005.

[3] D. Cai, X. He, and J. Han. Using graph model for face analysis. *University of Illinois Urbana-Champaign, Urbana, IL, Department of Computer Science and Technology*, 2005.

[4] J. Chen, Z. Zhao, J. Ye, and H. Liu. Nonlinear adaptive distance metric learning for clustering. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 123–132. ACM, 2007.

[5] F. De la Torre and T. Kanade. Discriminative cluster analysis. In *Proceedings of the 23rd International Conference on Machine learning (ICML'06)*, pages 241–248, 2006.

[6] C. Ding, X. He, H. Zha, and H. Simon. Adaptive dimension reduction for clustering high dimensional data. In *Proceedings of the 2nd IEEE International Conference on Data Mining (ICDM'02)*, pages 147–154, 2002.

[7] C. Ding and T. Li. Adaptive dimension reduction using discriminant analysis and k-means clustering. In *Proceedings of the 24th International Conference on Machine learning (ICML'07)*, pages 521–528, 2007.

[8] C. Domeniconi, D. Papadopoulos, D. Gunopulos, and S. Ma. Subspace clustering of high dimensional data. In *Proceedings of the 4th SIAM International Conference on Data Mining (SDM'04)*, 2004.

[9] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2001.

[10] D. Lee and H. Seung. Algorithms for non-negative matrix factorization. In *NIPS'01*, volume 13, 2001.

[11] T. Li, S. Ma, and M. Ogihara. Document clustering via adaptive subspace iteration. In *SIGIR'04*.

[12] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[13] J. Ye, Z. Zhao, and H. Liu. Adaptive distance metric learning for clustering. In *Proceedings of the 12th IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07)*, pages 1–7, 2007.

[14] J. Ye, Z. Zhao, and M. Wu. Discriminative k-means for clustering. In *NIPS'07*, volume 20, pages 1649–1656, 2007.