Adaptive Distance Metric Learning for Clustering

Jieping Ye, Zheng Zhao, Huan Liu
Computer Science and Engineering Department
Arizona State University
{jieping.ye, zhaozheng, hliu}@asu.edu

Abstract

A good distance metric is crucial for unsupervised learning from high-dimensional data. To learn a metric without any constraint or class label information, most unsupervised metric learning algorithms appeal to projecting observed data onto a low-dimensional manifold, where geometric relationships such as local or global pairwise distances are preserved. However, the projection may not necessarily improve the separability of the data, which is the desirable outcome of clustering. In this paper, we propose a novel unsupervised **A***daptive* **M***etric* **L***earning algorithm, called* AML*, which performs clustering and distance metric learning simultaneously.* AML *projects the data onto a low-dimensional manifold, where the separability of the data is maximized. We show that the joint clustering and distance metric learning can be formulated as a trace maximization problem, which can be solved via an iterative procedure in the EM framework. Experimental results on a collection of benchmark data sets demonstrated the effectiveness of the proposed algorithm.*

1. Introduction

A good distance metric is crucial to many real-world applications involving high-dimensional data, such as image classification and clustering, microarray data analysis, text mining, and content-based image retrieval. In distance metric learning, the goal is to learn a metric, under which the relationships of the observed data are better represented in comparison with the usual distance metrics, such as the Euclidian distance. With a good distance metric, the construction of the learning models becomes easier and the accuracy of the learning models usually improves [24]. Based on the availability of the constraint information (or class label information), metric learning algorithms fall into two categories: supervised distance metric learning [19, 22, 23, 25] and unsupervised distance metric learning [2, 15, 16, 18, 20]. In this paper, we focus on the case of unsupervised distance metric learning, which is more challenging due to the lack of any prior knowledge. Without any constraint or class label information, most unsupervised metric learning algorithms appeal to projecting observed data onto a low-dimensional manifold, where geometric relationships, such as the pairwise distances are preserved. Most approaches in this category, such as the Principle Component Analysis (PCA) [15], Locally Linear Embedding (LLE) [18], Laplacian Eigenmap [2] and ISOMAP [20], are also dimension reduction approaches and are related to manifold learning. Unsupervised learning algorithms, such as K-means can then be applied in the dimension-reduced space, thus avoiding the *curse of dimensionality*.

In unsupervised clustering, the goal is to find a set of clusters so that the separability between different clusters is maximized. Applying unsupervised dimensionality reduction algorithms such as PCA and other methods as a separate data pre-processing step before clustering may not help improve the separability of the data, which is the desirable outcome of clustering. In this paper, we propose AML, which stands for **A**daptive **M**etric **L**earning for simultaneous distance metric learning (via dimensionality reduction) and clustering. The key idea in AML is to integrate dimensionality reduction and clustering in a joint framework so that the separability of the data is maximized in the low-dimensional space. It has the same flavor as supervised metric learning approaches, which try to adjust the distance among instances to improve the separability of the data. For example, in [19, 23], the distance metric adjusts the geometry of data, so that the distance between data points from the same class under the metric is small. The metric improves the separability of the data and enhances the performance of classifiers, such as K-Nearest-Neighbor (KNN). The novel aspect of the proposed approach in comparison with supervised metric learning approaches is that AML does not use any class label information.

We show that the joint dimensionality reduction and clustering can be formulated as a trace maximization problem, which can be solved by an iterative algorithm based on the EM framework. We evaluate the proposed algorithm us-

1-4244-1180-7/07/$25.00 ©2007 IEEE

ing six benchmark data sets, including Soybean Large (Soybean), Segment, and Letter from UCI Machine Learning Repository [3], GCM microarry data set from [17], and two image data sets: USPS handwritten data [14] and Yale Face B (YaleFaceB) data [10]. We use the K-means algorithm as the baseline for comparison. We also compare the proposed algorithm with two representative unsupervised algorithms: Principle Component Analysis (PCA) and Locally Linear Embedding (LLE). Experimental results show that AML outperforms K-means, PCA, and LLE in most cases, which demonstrates the effectiveness of the proposed algorithm in learning a good distance metric for clustering. We also conducted a preliminary study on incorporating partial label information into the proposed framework, known as semi-supervised learning [4, 27]. Our preliminary results showed that using both labeled and unlabeled data does helps to learn a better distance metric.

The remainder of the paper is organized as follows. In Section 2, we present the idea behind our approach and formulate the adaptive metric learning problem. We propose an EM based algorithm, AML, to solve the problem in Section 3. We also investigate the properties of the proposed algorithm in this section. In Section 4, we present an empirical study to evaluate the effectiveness of the proposed algorithm in comparison with the representative algorithms and conduct a sensitivity study to evaluate various components of the algorithm. We conclude in Section 5 with discussions and future work.

2. Adaptive Metric Learning: Problem Formulation

Let X denote a data set with n instances, $\{x_j\}_{j=1}^n \in \mathbb{R}^m$. Let $G \in \mathbb{R}^{m\times l}$ be a linear transformation that maps each x_i in the m-dimensional space to a vector $\hat{x}_i$ in the l-dimensional space:

$$G : x_i \in \mathbb{R}^m \rightarrow \hat{x}_i = G^T x_i \in \mathbb{R}^l (l < m). \tag{1}$$

We focus on orthogonal transformations in this paper, that is, $G^T G = I_l$, where I_l is the identity matrix of size l. It has been shown [7, 11] that for most high-dimensional data sets, almost all low dimensional projections are nearly normal. That is, for large m, we expect the projected data $\{\hat{x}_i\}_{i=1}^n$ to be nearly normal. In this case, a good distance measure is the well-known *Mahalanobis* distance measure defined as follows:

$$d_M(\hat{x}_i, \hat{x}_j) = \sqrt{(\hat{x}_i - \hat{x}_j)^T \hat{\Sigma}^{-1} (\hat{x}_i - \hat{x}_j)}, \tag{2}$$

where $\hat{\Sigma}$ is the covariance matrix defined as follows:

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n (\hat{x}_i - \hat{\mu})(\hat{x}_i - \hat{\mu})^T, \tag{3}$$

and $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n \hat{x}_i$ is the mean of $\{\hat{x}_i\}_{i=1}^n$. It follows that

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n G^T (x_i - \mu)(x_i - \mu)^T G = G^T \Sigma G, \tag{4}$$

where $\mu = \frac{1}{n}\sum_{i=1}^n x_i$ is the mean of $\{x_i\}_{i=1}^n$, and

$$\Sigma = \frac{1}{n}\sum_{i=1}^n (x_i - \mu)(x_i - \mu)^T \tag{5}$$

is the class covariance matrix of the original data in X. For high-dimensional data, the estimation of the covariance matrix in Equation (5) is often not reliable. We apply the regularization technique [9] to improve the estimation as follows:

$$\Sigma = \frac{1}{n}\sum_{i=1}^n (x_i - \mu)(x_i - \mu)^T + \lambda I_m, \tag{6}$$

where I_m is the identity matrix of size m and $\lambda > 0$ is a regularization parameter.

Under this new distance measure, K-means clustering can be applied to assign $\{\hat{x}_i\}_{i=1}^n$ into K disjoint clusters, $\{C_j\}_{j=1}^K$, which minimize the following *Sum of Squared Error* (SSE):

$$\text{SSE}(\{C_j\}_{j=1}^K) = \sum_{j=1}^K \sum_{\hat{x}_i \in C_j} d_M(\hat{x}_i, \mu_j)^2, \tag{7}$$

where the Manalanobis distance $d_M(\cdot,\cdot)$ is defined as in Equation (2), and μ_j is the mean of the j-th cluster C_j.

As the summation of all pair-wise distances is a constant for a fixed G. The minimization of the SSE is equivalent to the maximization of *Sum of Squared Intra-cluster Error* (SSIE) defined as follows:

$$\text{SSIE}\left(\{C_j\}_{j=1}^K\right) = \sum_{j=1}^K n_j\, d_M(\mu_j, \hat{\mu})^2, \tag{8}$$

where n_j is the sample size of the j-th cluster C_j, μ_j is the mean of the j-th cluster C_j, and $\hat{\mu}$ is the global mean as defined above. SSIE can be expressed in a compact matrix form as follows. Let $F \in \mathbb{R}^{n\times K}$ be the cluster indicator matrix defined as follows:

$$F = \{f_{i,j}\}_{n\times K}, \text{ where } f_{i,j} = 1, \text{ iff } x_i \in C_j. \tag{9}$$

We can define the weighted cluster indicator matrix

$$L = [L_1, L_2, \cdots, L_K]$$

as follows [8]:

$$L = F(F^T F)^{-\frac{1}{2}}. \tag{10}$$

It follows that the i-th column of L is given by

$$L_i = (0, \ldots, 0, \overbrace{1, \ldots, 1}^{n_i}, 0, \ldots, 0)^T / n_i^{\frac{1}{2}}. \quad (11)$$

With the weighted cluster indicator matrix L, the Sum of Squared Intra-cluster Error (SSIE) can be expressed as [8]:

$$\begin{aligned} & \frac{1}{n} trace\left(L^T X^T G \left(\hat{\Sigma}\right)^{-1} G^T X L\right) \\ = & \frac{1}{n} trace\left(L^T X^T G (G^T \Sigma G)^{-1} G^T X L\right). \quad (12) \end{aligned}$$

The adaptive metric learning problem can be formulated as follows:

$$\max_{G:\, G^T G = I_l, L} \frac{1}{n} trace\left(L^T X^T G (G^T \Sigma G)^{-1} G^T X L\right). \quad (13)$$

The optimization problem in Equation (13) maximizes the squared intra-cluster distance under the Mahalanobis distance measure determined by the transformation G. Thus, it simultaneously computes the distance metric and the clustering. However, the optimization problem is highly nonlinear and difficult to solve. In the following, we propose to develop an iterative algorithm to compute the transformation matrix G and the clustering captured in L.

3. Adaptive Metric Learning: The Main Algorithm

The objective function in Equation (13) is not convex. However, if one of the two components (G and L) is fixed, the objective function is convex in terms of the other component and the optimization problem is easy to solve. This enables us to solve the problem in the **EM** framework, in which we update G and L iteratively to find a (local) optimal solution for the adaptive metric learning problem.

3.1. Calculation of L for a given G

For a given transformation matrix G, the maximization problem specified in Equation (13) reduces to the maximization of

$$trace\left(L^T \tilde{K} L\right),$$

where $\tilde{K}$ is defined as

$$\tilde{K} = \frac{1}{n} X^T G \left(G^T \Sigma G\right)^{-1} G^T X.$$

This is equivalent to a kernel K-means problem with $\tilde{K}$ as the kernel, as summarized below [6]:

THEOREM 1 *Give the transformation matrix G and the cluster number k, the computation of L that solves the optimization problem specified in Equation (13) is equivalent to a kernel K-means problem, where the kernel matrix $\tilde{K}$ is given by*

$$\tilde{K} = \frac{1}{n} X^T G \left(G^T \Sigma G\right)^{-1} G^T X. \quad (14)$$

3.2. Calculation of G for a given L

Since trace(AB) = trace(BA), for any two matrices A and B, we have

$$\begin{aligned} \text{SSIE}\left(\{C_j\}_{j=1}^K\right) &= \frac{1}{n} \text{trace}\left(L^T X^T G (G^T \Sigma G)^{-1} G^T X L\right) \\ &= \frac{1}{n} \text{trace}\left((G^T \Sigma G)^{-1} G^T X L L^T X^T G\right). \end{aligned}$$

For a given L, the optimal G^* can be computed by solving an eigenvalue problem associated with Σ and XLL^TX^T as summarized below:

THEOREM 2 *Let $A = XLL^TX^T$ and $B = \Sigma$. Let $M = [v_1, \cdots, v_l]$ be the matrix consisting of the first l eigenvectors corresponding to the largest l nonzero eigenvalues of $B^{-1}A$. Let $M = QR$ be the QR decomposition of M, where Q has orthonormal columns and R is upper triangular. Then $G^* = Q$.*

PROOF: For a given L, the optimal G is given by solving the following optimization problem:

$$\max_{G:\, G^T G = I} f(G), \quad (15)$$

where $f(G) = \text{trace}\left((G^T B G)^{-1} G^T A G\right)$.

Since $f(G) = f(GM)$, for any nonsingular matrix M, the optimal G^* is given by first maximizing $f(G)$ without the orthogonality constraint and then computing the QR decomposition. The result follows, since the optimal G maximizing $f(G)$ is given by the top eigenvectors of $B^{-1}A$. □

Note that the computation of G is closely related to the well-known linear discriminant analysis [12]. The optimization problem in Equation (13) is similar to the one in [16], however the algorithm in [16] is based on gradient descent and is computationally more expensive. Since the rank of the matrix XLL^TX^T is no larger than $K-1$, the number of columns of G is no larger than $K-1$, i.e., $l \leq K-1$. We set $l = K-1$ in the following discussion.

3.3. The Main Algorithm

Based on the analysis above, we propose to develop an iterative algorithm for the adaptive metric learning problem defined in Equation (13). The corresponding algorithm, called AML is presented in Algorithm 1.

The convergence of the AML algorithm is guaranteed, as summarized in the following theorem below:

THEOREM 3 *Algorithm AML always converges.*

PROOF: It is easy to see that in steps 4 and 5 of AML, the objective value defined in Equation (13) always increases. As the objective value is bounded from above by a finite number, the algorithm always converges. □

Algorithm 1: AML

Input: X, k, ϵ
Output: G, L

1 Using K-means to obtain the initial weighted cluster indication matrix L; Compute Σ as in Equation (6);
2 Compute the optimal G, by computing the QR decomposition of the matrix, which consists of the first $K-1$ eigenvectors of the matrix $\Sigma^{-1}XLL^TX^T$;
3 **while** *trace incensement* $\geq \epsilon$ **do**
4 For a given G, update L by running kernel K-means as in Section 3.1 with the initial set of centroids given by the current L;
5 For a given L, update G by computing the QR decomposition of the matrix, which consists of the first $K-1$ eigenvectors of $\Sigma^{-1}XLL^TX^T$;
6 **end**
7 return G and L;

4. Experimental Result

In this section, we present an empirical study to evaluate the AML algorithm in comparison with several other representative algorithms and conduct a sensitivity study to evaluate various components of the algorithm.

4.1. Experimental Setup

We use the K-means algorithm as the baseline for comparison. We also compare the proposed algorithm with two representative unsupervised learning algorithms including Principal Component Analysis (PCA) and Locally Linear Embedding (LLE).We implemented AML in the MATLAB environment. All experiments were conducted on a PENTIUM IV 2.4G PC with 1.5GB RAM. We compared all algorithms on six benchmark data sets, including Soybean Large (Soybean), Segment, and Letter (Letter (a-d)) from UCI Machine Learning Repository, GCM microarray data set, and two image data sets: USPS handwritten data and Yale Face B (YaleFaceB)[1]. The statistics of all data sets are summarized in Table 1. All data have been normalized so that each feature has mean 0 and standard deviation 1.

For each data set, we ran different algorithms for 20 times and the comparison was based on the average performance. We used the existing label information for all six benchmark data sets to evaluate the performance of different algorithms. Two standard measurements are used: the accuracy (ACC) and the normalized mutual information metric (MI) [13]. In all experiments, the dimensionality, l, of the subspace produced by PCA is selected so that at least 95% information of the original data is kept. The dimensionality of AML is set to $K-1$.

[1]For YaleFaceB data, image size is reduced from 648*480 to 40*30. For Soybean Large data, instances with unknown value are removed, which results in a data set with 562 instances and 15 classes. For Letter data, the first 4 letters "a, b, c, d" were selected.

Data set	Dimension	Instance	Class
GCM	11485	190	14
Soybean	35	562	15
Segment	19	2309	7
Letter (a-d)	16	3096	4
USPS	256	9298	10
YaleFaceB	1200	5850	10

Table 1. Summary of the test data sets used in our experiment.

4.2. Experimental Results

Table 2 presents the accuracy (ACC) and normalized mutual information (MI) results of various algorithms on all six data sets. We include the results of AML using 3 different λ values: 0, 1, and 100. We can observe from the table that in terms of accuracy, AML with $\lambda = 100$ performs the best on 5 data sets. Considering all six benchmark data sets, AML with $\lambda = 100$ performs the best with an average accuracy of 0.685, which is followed by AML with $\lambda = 1$ with an average accuracy of 0.681, and then AML with $\lambda = 0$ with an average accuracy of 0.677. We can also observe that AML improves K-means on all six data sets. On Segment data, AML with $\lambda = 1$ improves its accuracy from 0.552 to 0.756, an 37% improvement. Similar trends can also be observed for the MI measure.

4.2.1 Sensitivity Study

The Effect of Regularization. As mentioned in Section 2, the regularization parameter λ is introduced to improve the estimation of the covariance matrix. From Table 2, we can see that in general the regularization helps to improve the performance of AML. For example, on most data sets, the performance of AML with $\lambda = 100$ is significantly better than that with no regularization (i.e., $\lambda = 0$). To obtain a better understanding of the effect of the regularization parameter, we tried a series of different λ values from 0 to 10^5. The results are plotted in Figure 1. We can observe that a λ value of about hundreds is usually helpful. On segment data, however, with $\lambda \leq 1$, the performance of AML is the best. Once the λ value becomes higher, the performance degrades sharply. This may be related to characteristics of the Segment data, which has a small dimensionality, but a large number of instances. In this

Table 2. Accuracy (ACC) and Mutual information (MI) comparison on six benchmark data sets. AML with different λ values are included. For each data set, the first row and the second row stand for ACC (or MI) and p-value respectively. (The p-value is obtained by pairwise student t-test.) The p-value of each algorithm is generated by comparing its ACC (or MI) value with the highest one. ACCs (or MIs) with bold face are the highest ones, or the second highest if it has no significant difference with the highest one. AML$_{EU}$ is for sensitivity study (see Section 4.2.1 for detail).

Meas	Data set	AML(0)	AML(1)	AML(100)	K-means	AML$_{EU}$	PCA	LLE
ACC	GCM	0.568	0.58	**0.583**	0.568	0.568	0.569	0.571
		0	0.007	-	0	0	0	0
	Soybean	0.674	0.678	**0.725**	0.671	0.671	0.668	0.705
		0	0	-	0	0	0	0.002
	Segment	**0.751**	**0.756**	0.644	0.552	0.552	0.551	0.533
		0.324	-	0	0	0	0	0
	Letter (a-d)	0.635	0.631	**0.662**	0.606	0.606	0.606	0.647
		0	0	-	0	0	0	0.003
	USPS	0.703	0.704	**0.726**	0.708	0.708	0.709	0.655
		0	0	-	0.001	0	0.001	0
	YaleFaceB	0.733	0.735	**0.771**	0.733	0.733	0.733	0.746
		0	0	-	0	0	0	0.002
	Average	0.677	0.681	0.685	0.64	0.64	0.639	0.643
MI	GCM	0.57	**0.585**	**0.587**	0.57	0.57	0.571	0.573
		0	0.133	-	0	0	0	0.001
	Soybean	0.675	0.683	**0.716**	0.651	0.651	0.648	0.678
		0	0.004	-	0	0	0	0
	Segment	**0.683**	**0.683**	0.586	0.502	0.502	0.501	0.421
		-	0.929	0	0	0	0	0
	Letter (a-d)	0.493	0.484	**0.508**	0.423	0.423	0.422	**0.509**
		0.398	0.238	0.934	0	0.012	0	-
	USPS	0.607	0.614	**0.627**	0.603	0.603	0.603	0.539
		0	0	-	0	0	0	0
	YaleFaceB	0.764	0.766	**0.809**	0.764	0.764	0.764	0.782
		0	0	-	0	0	0	0.001
	Average	0.632	0.636	0.639	0.586	0.586	0.585	0.584

case, the covariance matrix may well be estimated from the given data, even without any regularization.

The Effect of Mahalanobis Distance. In AML, we employ the Mahalanobis distance in the K-means clustering as defined in Equation (2). If instead we use the traditional Euclidian distance, AML is reduced to the following iterative process: for a given G, solve the kernel K-means, in which the kernel is defined as: $W = X^T GG^T X$; and for a given L, solve the eigenvalue problem on the matrix $A = XLL^T X^T$ to obtain the transformation matrix G. The algorithm is called AML$_{EU}$, which stands for **AML** with **EU**clidian distance. From Table 2, we can observe that AML$_{EU}$ has the same performance as K-means. We further observe from the experiment that when the Euclidian distance is used, the cluster affiliations are exactly kept after the projection, therefore in the reduced space, the initial centroids used in the algorithm will not change. This validates the use of the Mahalanobis distance in AML.

Incorporating Partial Label Information. One nice feature of the proposed AML algorithm is that it is convenient to incorporate the partial label information into the proposed framework. This is known as the semi-supervised learning [4, 27]. Recall that one of the key steps in AML is the kernel K-means, which can be readily extended to include the labeled data by applying semi-supervised K-means clustering, such as the constrained K-means in [21]. The algorithm is called *semiAML*. We expect that using both labeled and unlabeled data would help to learn a better distance metric. Figure 2 shows the accuracy results on USPS and YaleFaceB data. (Similar trends have been observed from other data sets and results are omitted here.) In the experiment, we varied the number of labeled data points per class and reported the corresponding accuracy. We can observe from Figure 2 that the use of partial label information does help to improve the performance of both constrained K-means and AML.

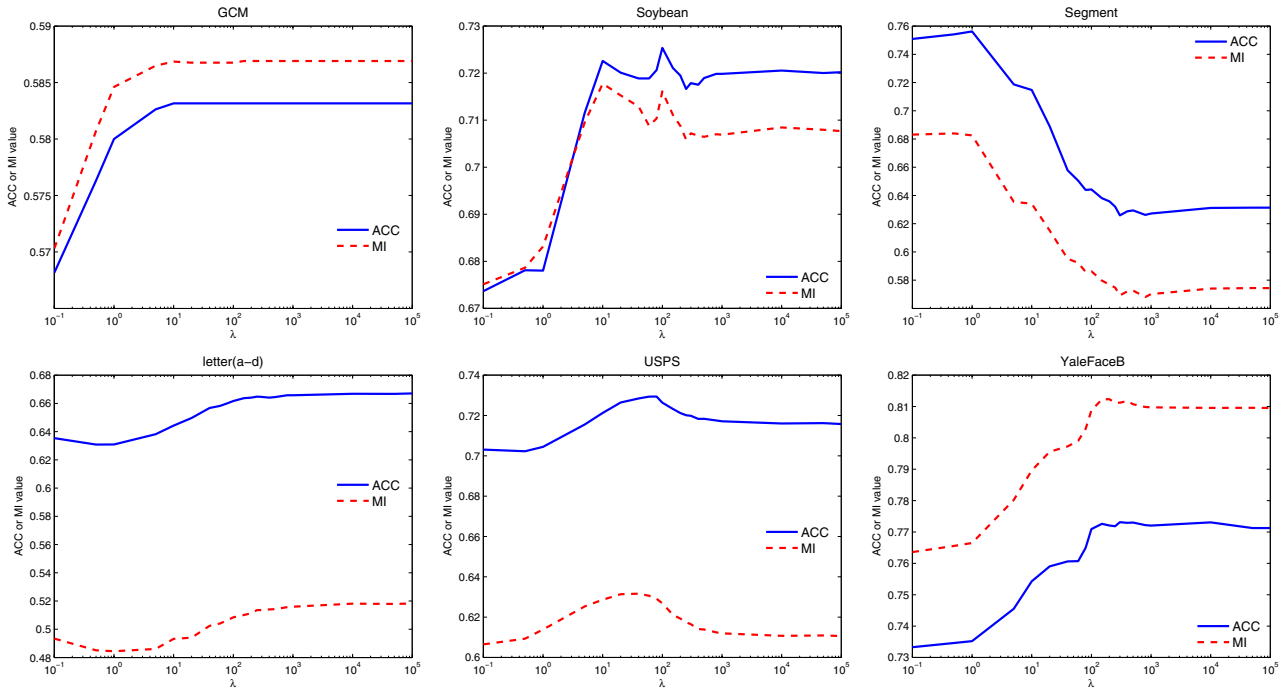

Figure 1. Comparison of ACC and MI with different λ values. In each figure, the x-axis corresponds to different λ values and the y-axis corresponds to ACC or MI value. The logarithmic (base 10) scale is used for the x-axis.

5. Conclusion

In this paper, we propose a novel unsupervised **A**daptive **M**etric **L**earning algorithm, called AML, which performs distance metric learning and clustering simultaneously. AML projects the data onto a low-dimensional manifold, where the separability of the data is maximized. We show that the joint clustering and distance metric learning can be formulated as a trace maximization problem, which can be solved via an iterative procedure in the EM framework. Experimental results on six benchmark data sets demonstrated the effectiveness of the proposed algorithm.

AML is a global unsupervised distance metric learning algorithm. Compared with global approaches, local approaches are usually more flexible and have been shown to be effective in many applications. We plan to develop a local unsupervised distance metric learning algorithm in the framework of AML. Our preliminary results have shown the promise of semiAML for semi-supervised distance metric learning and dimensionality reduction. We plan to explore various existing techniques [1, 5, 26, 28] from semi-supervised learning in the framework of AML.

Acknowledgments

This research is sponsored by the Center for Evolutionary Functional Genomics of the Biodesign Institute at ASU and by the National Science Foundation Grant IIS-0612069.

References

[1] M. Belkin, I. Matveeva, and P. Niyogi. Regularization and semi-supervised learning on large graphs. In *COLT*, pages 624–638, 2004. 6

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. In *NIPS*, 2003. 1

[3] C. Blake and C. Merz. UCI repository of machine learning databases, 1998. 2

[4] O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning*. The MIT Press, 2006. 2, 5

[5] O. Delalleau, Y. Bengio, and N. L. Roux. Efficient non-parametric function induction in semi-supervised learning. In *AISTATS*, 2005. 6

[6] I. S. Dhillon, Y. Guan, and B. Kulis. A unified view of kernel k-means, spectral clustering and graph partitioning. Technical report, Department of Computer Sciences, University of Texas at Austin, 2005. 3

[7] P. Diaconis and D. Freedman. Asymptotics of graphical projection pursuit. *Annuals of Statistics*, 12:793815, 1984. 2

[8] C. Ding, X. He, and H. D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *SDM*, pages 606–610, 2006. 2, 3

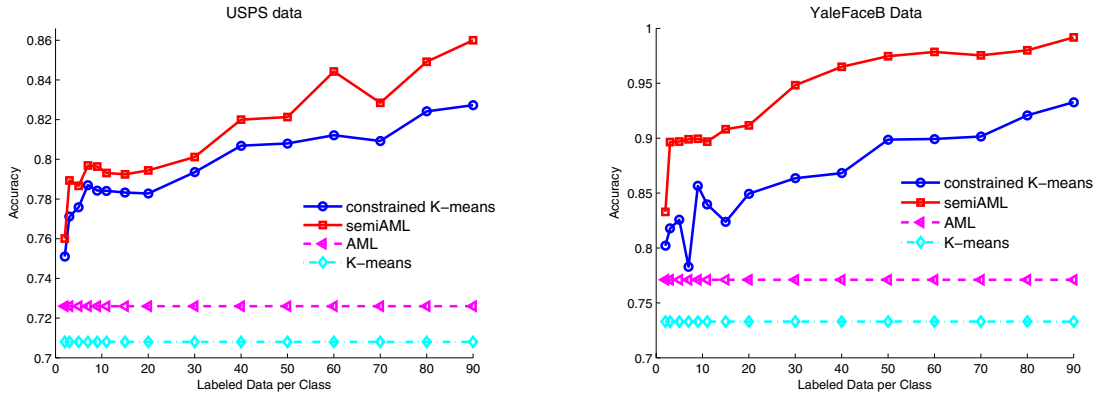

Figure 2. Effect of incorporating partial label information in AML. The x-axis denotes the number of labeled data points per class and the y-axis denotes accuracy. semiAML denotes semi-supervised AML, which employs constrained K-means in AML to incorporating partial label information. It is clear that labeled data help to improve the performance of both constrained K-means and AML.

[9] J. H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175, 1989. 2

[10] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 23:643–660, 2001. 2

[11] P. Hall and K. Li. On almost linearity of low dimensional projections from high dimensional data. *Annuals of Statistics*, 21:867–889, 1993. 2

[12] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001. 3

[13] X. He, D. Cai, and P. Niyogi. Laplacian score for feature selection. In *NIPS*. 4

[14] J. J. Hull. A database for handwritten text recognition research. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 16(5):550–554, 1994. 2

[15] I. Jolliffe. *Principal Component Analysis*. Springer; 2nd edition, 2002. 1

[16] F. D. la Torre Frade and T. Kanade. Discriminative cluster analysis. In *ICML*, pages 241–248, 2006. 1, 3

[17] S. Ramaswamy, P. Tamayo, R. Rifkin, S. Mukherjee, C.-H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. Mesirov, T. Poggio, W. Gerald, M. Loda, E. Lander, and T. Golub. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences*, 98:15149–15154, 2001. 2

[18] L. K. Saul and S. T. Roweis. Think globally, fit locally : Unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4:119–155, 2003. 1

[19] N. Shental, T. Hertz, D. Weinshall, and M. Pavel. Adjustment learning and relevant component analysis. In *ECCV*, pages 776–792, 2002. 1

[20] J. Tenenbaum, V. de Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000. 1

[21] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl. Constrained k-means clustering with background knowledge. In *ICML*, pages 577–584, 2001. 5

[22] K. Weinberger, J. Blitzer, and L. Saul. Distance metric learning for large margin nearest neighbor classification. In *NIPS*, 2005. 1

[23] E. Xing, A. Ng, M. Jordan, and S. Russell. Distance metric learning with application to clustering with side-information. In *NIPS*, 2002. 1

[24] L. Yang and R. Jin. Distance metric learning: A comprehensive survey. Technical report, Department of Computer Science and Engineering, Michigan State University, 2006. 1

[25] L. Yang, R. Jin, R. Sukthankar, and Y. Liu. An efficient algorithm for local distance metric learning. In *AAAI*, 2006. 1

[26] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *NIPS*, 2004. 6

[27] X. J. Zhu. Semi-supervised learning literature survey. Technical report, Computer Sciences, University of Wisconsin-Madison, 2005. 2, 5

[28] X. J. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, 2003. 6