SUBSPACE LEARNING USING CONSENSUS ON THE GRASSMANNIAN MANIFOLD

Jayaraman J. Thiagarajan† *and Karthikeyan Natesan Ramamurthy*‡

†Lawrence Livermore National Laboratory, ‡IBM Thomas J. Watson Research Center

ABSTRACT

High-dimensional structure of data can be explored and task-specific representations can be obtained using manifold learning and low-dimensional embedding approaches. However, the uncertainties in data and the sensitivity of the algorithms to parameter settings, reduce the reliability of such representations, and make visualization and interpretation of data very challenging. A natural approach to combat challenges pertinent to data visualization is to use linearized embedding approaches. In this paper, we explore approaches to improve the reliability of linearized, subspace embedding frameworks by learning a plurality of subspaces and computing a geometric mean on the Grassmannian manifold. Using the proposed algorithm, we build variants of popular unsupervised and supervised graph embedding algorithms, and show that we can infer high-quality embeddings, thereby significantly improving their usability in visualization and classification.

***Index Terms*—** graph embedding, subspace learning, Grassmannian manifold, visualization.

1. INTRODUCTION

Low-dimensional representations of high dimensional data are desired for several reasons. If the data generating process has low degrees of freedom, it may be useful to obtain representations that disentangle the various degrees of freedom and reject the noise. This is the goal of several manifold learning algorithms, many of which can be posed using a Graph embedding formulation [1]. The representations created by these methods retain either the geometric characteristics or the topology (local neighborhood) or both, depending on the type of the embedding algorithm used. Furthermore, supervised embedding approaches create representations, that improve the discrimination across classes apart from preserving the similarity between within-class samples.

Some examples of unsupervised graph embedding approaches include Principal Components Analysis (PCA) [2], Multi-Dimensional Scaling (MDS) [3], ISOMAP [4], Laplacian Eigenmaps (LE) [5], Locality Preserving Projections (LPP) [6], and Neighborhood Preserving Embedding (NPE) [7]. Some well-known supervised embedding approaches include Linear Discriminant Analysis (LDA) [8], Marginal Fisher Analysis (MFA) [1], and Local Discriminant Embedding (LDE) [9]. Though direct graph embedding approaches such as ISOMAP, LE, and MDS have been successfully used to learn underlying non-linear manifolds, linearized embedding algorithms have been of particular interest to practitioners. Explicitly inferring the linear subspace (e.g. PCA) greatly simplifies the out-of-sample extension procedure. In addition, when compared to direct embeddings where the interpretation of the embedded space is entirely opaque, linearized embeddings can lead to more meaningful data visualization.

The success of graph embedding depends not only on the parameters such as the embedding dimension or the neighborhood graph that needs to be carefully designed, but also the suitability of the embedding approach to the data under consideration. Furthermore, most algorithms only have global parameters, whereas the topological characteristics of data could vary locally. An example of this is when the data is perfectly sampled from a smooth manifold of constant intrinsic dimensionality, the embedding obtained may not be meaningful if its dimensionality does not match the intrinsic dimension of the manifold. This is common in visualization applications, where the embedding dimension is constrained to be 2 or 3. An alternative approach proposed in [10] attempts to alleviate this challenge by posing embedding inference as an information retrieval problem, and constructs a direct $2-D$ embedding that optimizes a function of neighborhood precision and recall. Similar issues exist with supervised embedding approaches as well.

In this paper, we propose approaches that improve the performance of linearized, unsupervised and supervised graph embedding methods, by computing an ensemble of incoherent linear subspaces for the data. Since the subspaces obtained lie on a Grassmannian manifold, a consensus subspace can be inferred on the Grassmannian. The final embedding will be obtained by projecting the data onto the consensus subspace. Using the proposed technique, we build variants of some of the popular unsupervised and supervised embedding algorithms and evaluate their performance on several datasets. Experiment results show that the proposed algorithm can significantly improve the quality of linearized subspace learning methods, and can also impact their performance in conventional tasks such as classification, and visualization for data exploration.

2. THEORY AND BACKGROUND

2.1. Graph Embedding

Let us consider a set of samples of T samples denoted by $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T]$, where $\mathbf{x}_i \in \mathbb{R}^M$. For supervised embedding, we will also assume that there are C classes with labels $\{1, \ldots, C\}$. The class c has T_c samples and $\mathcal{I}_c$ contains the indices of samples in that class. The goal of embedding approaches is to find a mapping function $v : \mathbf{x} \rightarrow \mathbf{y}$, where $\mathbf{y} \in \mathbb{R}^N$ and $N \ll M$. Direct methods obtain the embedding $\mathbf{y}$ directly using an implicitly defined v, whereas linearized methods obtain v as the linear subspace whose orthonormal basis is defined as $\mathbf{V} \in \mathbb{R}^{M \times N}$, and hence $\mathbf{y} = \mathbf{V}^T\mathbf{x}$.

We will define a similarity graph G that comprises the vertex set $\mathbf{X}$ and the similarity between each of the vertices encoded in the symmetric adjacency matrix $\mathbf{W} \in \mathbb{R}^{T \times T}$. With the exception of NPE, for all the linearized methods we consider, the Laplacian matrix is defined as $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ is the diagonal degree matrix of the graph with $D_{ii} = \sum_{i \neq j} W_{ij}$. In NPE, $\mathbf{L}$ is directly defined as $(\mathbf{I} - \mathbf{W})^T(\mathbf{I} - \mathbf{W})$. Furthermore, we create a penalty matrix $\mathbf{H}$, which imposes a constraint on $\mathbf{Y}^T\mathbf{H}\mathbf{Y} = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. For example, with LPP, $\mathbf{H}$ is chosen to be the degree matrix $\mathbf{D}$, whereas with NPE it is chosen to be $\mathbf{I}$. In supervised dimension-

ality reduction, $\mathbf{H}$ is chosen to the penalty Laplacian, $\mathbf{L} = \mathbf{D}' - \mathbf{W}'$, where $\mathbf{D}'$ and $\mathbf{W}'$ are respectively the degree and adjacency matrices of the penalty graphs. The forms of $\mathbf{L}$ and $\mathbf{H}$ for various graph embeddings can be found in [1].

The embedding directions $\mathbf{V}$ with linearized methods can be optimized as

$$\mathbf{V} = \underset{\mathbf{V}^T\mathbf{X}\mathbf{H}\mathbf{X}^T\mathbf{V}=\mathbf{I}}{\arg\min} \text{Tr}(\mathbf{V}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{V}). \quad (1)$$

This optimization can be carried out by choosing the N eigen vectors corresponding to the minimum eigen values of the generalized eigen decomposition (GED),

$$\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{v} = \lambda\mathbf{X}\mathbf{H}\mathbf{X}^T\mathbf{v}, \quad (2)$$

and the low-dimensional embedding can be computed as $\mathbf{Y} = \mathbf{V}^T\mathbf{X}$. Note that $\mathbf{V}$ obtained with the GED is not guaranteed to contain only orthonormal columns.

Note that the eigen decomposition method is typically suited for unsupervised methods since $\mathbf{H}$ will truly represent a constraint. In supervised approaches, ideally we would like to compute embeddings that simultaneously maximize $\text{Tr}(\mathbf{V}^T\mathbf{X}\mathbf{H}\mathbf{X}^T\mathbf{V})$, while minimizing $\text{Tr}(\mathbf{V}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{V})$. In these cases, a provable globally optimal solution can be obtained for graph embedding by including an additional constraint $\mathbf{V}^T\mathbf{V} = \mathbf{I}$. The optimal embedding can be now obtained by maximizing the trace ratio

$$\mathbf{V} = \underset{\mathbf{V}^T\mathbf{V}=\mathbf{I}}{\arg\max} \frac{\text{Tr}(\mathbf{V}^T\mathbf{X}\mathbf{H}\mathbf{X}^T\mathbf{V})}{\text{Tr}(\mathbf{V}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{V})}. \quad (3)$$

Approaches such as iterative trace ratio (ITR) maximization or decomposed Newton's method (DNM) can be used for this purpose and it has been shown that trace ratio maximization performs better than generalized eigen decomposition in supervised dimensionality reduction [11, 12].

2.2. Grassmannian Manifolds

A Grassmannian manifold, $\text{Gr}(N, M)$ is a set of $N-$dimensional subspaces in $\mathbb{R}^M$, where each subspace maps to a unique point on the manifold [13]. Each point in $\text{Gr}(N, M)$ can be conveniently represented using an $M \times N$ orthonormal matrix which forms a basis for that subspace. Given two points in a Grassmannian, represented by their orthonormal bases, $\mathbf{A}$ and $\mathbf{B}$ of size $M \times N$, there are several distance measures that can be computed between them. The distance measured along the geodesic is the Grassmann distance and can be computed by decomposing $\mathbf{A}^T\mathbf{B}$ using its SVD and obtaining $\sum_{i=1}^{N} \left(\theta_i^2\right)^{\frac{1}{2}}$. Here, θ_i denotes a principal angle and is obtained as $\cos^{-1}\sigma_i$, where σ_i is the corresponding singular value [14].

In general, Grassmann distances are difficult to optimize with, and hence the chordal distance [14], given as $\sqrt{N - \|\mathbf{A}^T\mathbf{B}\|_F^2}$ is widely adopted in lieu of the true geodesic distance [15, 16], among several others. The chordal distance is also referred to as the symmetric directional distance. The idea of distance between the subspaces can also be extended to the case when we have two subspaces of different dimensions [14]. For example, suppose $\mathbf{A} \in \mathbb{R}^{M\times N_1}$ and $\mathbf{B} \in \mathbb{R}^{M\times N_2}$, the chordal distance is given by $\sqrt{\max(N_1, N_2) - \|\mathbf{A}^T\mathbf{B}\|_F^2}$. If we use chordal distances, incoherence between subspaces corresponds to large distances and vice-versa. For a given distance metric, the mean of two subspaces is defined as the subspace that has the minimum sum of squared distance with the given subspaces. The mean of two subspaces of different dimensions is known for several metrics including the Grassmann and the chordal distances [14]. However, with more than two subspaces, it is much easier to use chordal distances to compute the mean, and also it is flexible to be coupled with other constraints, and hence we will focus on this distance measure in our algorithm.

3. PROPOSED ALGORITHM

Our proposed approach aims to improve the reliability of linearized unsupervised and supervised embedding algorithms by learning a plurality of subspaces with incoherence constraints between them and finding the mean subspace on the Grassmannian. The final embedding will be obtained by projecting the data onto the mean subspace.

We will begin by describing the algorithm for the unsupervised case. As described in Section 2.1, we will begin with the data matrix $\mathbf{X}$, the Laplacian $\mathbf{L}$ and the constraint matrix $\mathbf{H}$. We aim to construct K low-dimensional projections $\{\mathbf{V}_i\}_{i=1}^K$, each with dimensions $M \times N_i$ where $N_i < M$, that satisfy the objective in (1), and are far from each other in the Grassmannian. Here, $\mathbf{V}_i$ are assumed to contain orthonormal columns.

3.1. Obtaining Multiple Incoherent Subspaces

Let us assume that we have already computed j subspaces and we wish to compute the $j+1$ subspace that is far from all the previous subspaces on the Grassmannian. The squared chordal distance between $\mathbf{V}_{j+1}$ and $\mathbf{V}_i$, where $i \leq j$ is given by

$$\begin{aligned} d^2(\mathbf{V}_{j+1}, \mathbf{V}_i) &= \max(N_{j+1}, N_i) - \|\mathbf{V}_i^T\mathbf{V}_{j+1}\|_F^2, \\ &= \max(N_{j+1}, N_i) - \text{Tr}\left(\mathbf{V}_{j+1}^T\mathbf{V}_i\mathbf{V}_i^T\mathbf{V}_{j+1}\right), \end{aligned}$$

and hence the sum of squared distances $\sum_{i=1}^{j} d^2(\mathbf{V}_{j+1}, \mathbf{V}_i)$ is given as

$$\sum_{i=1}^{j} \max(N_{j+1}, N_i) - \text{Tr}\left(\mathbf{V}_{j+1}^T \sum_{i=1}^{j}\left(\mathbf{V}_i\mathbf{V}_i^T\right)\mathbf{V}_{j+1}\right). \quad (4)$$

Maximizing (4) w.r.t. $\mathbf{V}_{j+1}$ is equivalent to minimizing the negative of the second term. Combining this with (1), the overall optimization to compute $\mathbf{V}_{j+1}$ is given as

$$\mathbf{V}_{j+1} = \underset{\mathbf{V}^T\mathbf{X}\mathbf{H}\mathbf{X}^T\mathbf{V}=\mathbf{I}}{\arg\min} \text{Tr}\left(\mathbf{V}^T\left(\mathbf{X}\mathbf{L}\mathbf{X}^T + \alpha\sum_{i=1}^{j}\left(\mathbf{V}_i\mathbf{V}_i^T\right)\right)\mathbf{V}\right), \quad (5)$$

where $\alpha > 0$ is the tradeoff parameter between the embedding cost and the incoherence of the subspace. This can be solved by choosing N_{j+1} eigen vectors corresponding to the smallest eigen values from the GED

$$\left(\mathbf{X}\mathbf{L}\mathbf{X}^T + \alpha\sum_{i=1}^{j}\left(\mathbf{V}_i\mathbf{V}_i^T\right)\right)\mathbf{v} = \lambda\mathbf{X}\mathbf{H}\mathbf{X}^T\mathbf{v}. \quad (6)$$

The K incoherent subspaces are obtained by repeating this process K times, by including an additional subspace in each round. Note that we orthonormalize the columns of subspace basis computed in each round before proceeding to the next round.

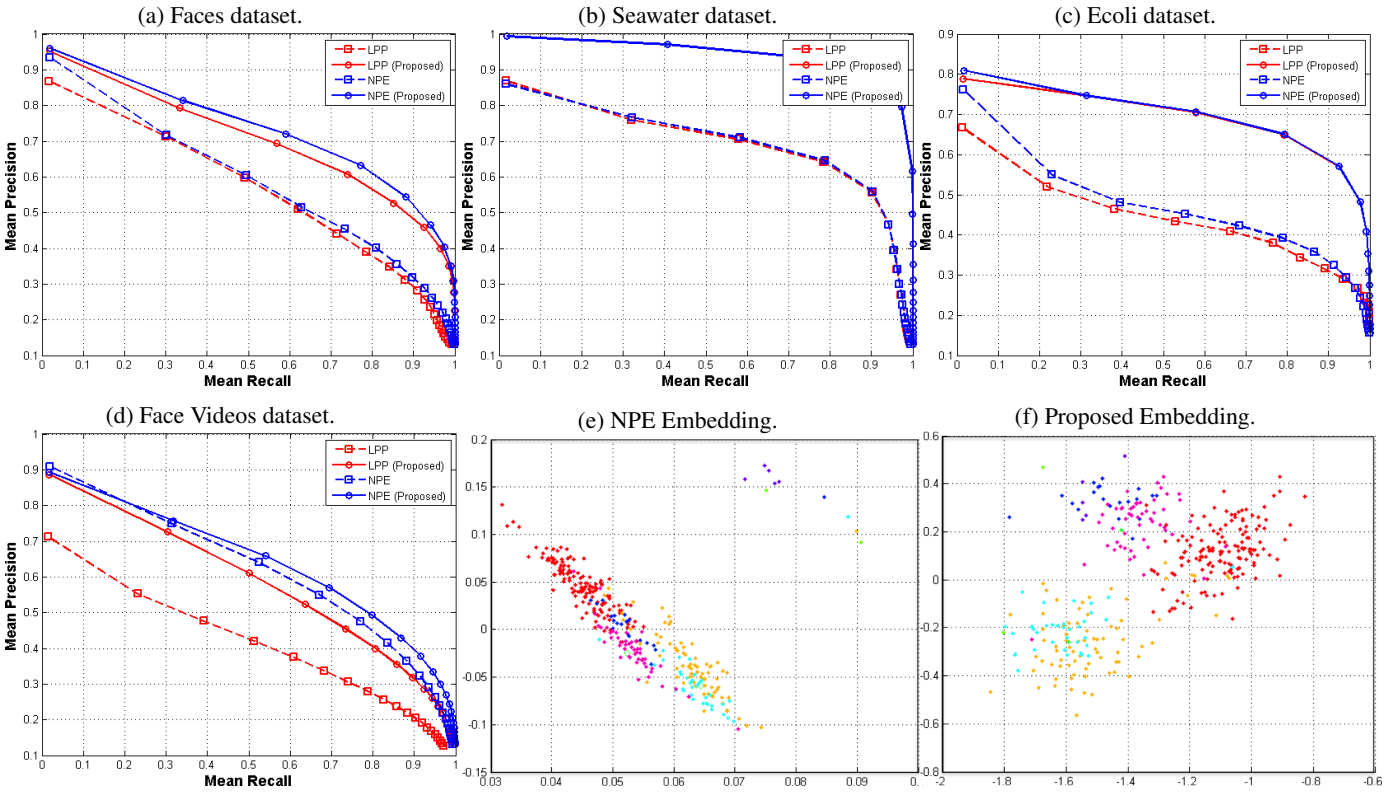

Fig. 1. Unsupervised Graph Embedding: (a)-(d) Mean precision - Mean Recall curves for different datasets; (e)-(f) 2-D embeddings for the UCI ecoli datasets obtained using NPE and the proposed variant.

3.2. Grassmannian Averaging

After obtaining the K subspaces, $\{V_i\}_{i=1}^K$, we will compute the consensus (mean) subspace that will have minimum sum of squared chordal distance with these subspaces. The dimension of the mean subspace spanned by the columns of $\mathbf{V}$, will be chosen to be the maximum of $\{N_i\}_{i=1}^K$ and we will denote it as N_m. The optimization problem can be now posed as

$$\begin{aligned}\mathbf{V} &= \underset{\mathbf{V}^T\mathbf{V}=\mathbf{I}}{\arg\min} \sum_{i=1}^{K} w_i \left(N_m - \|\mathbf{V}^T\mathbf{V}_i\|_F^2\right), \\ &= \underset{\mathbf{V}^T\mathbf{V}=\mathbf{I}}{\arg\min} \operatorname{Tr}\left(\mathbf{V}^T \sum_{i=1}^{K} \left(w_i\mathbf{I} - w_i\mathbf{V}_i\mathbf{V}_i^T\right)\mathbf{V}\right) \end{aligned} \tag{7}$$

where w_k are the optional positive weights. $\mathbf{V}$ can be evaluated as the N_m eigen vectors corresponding to the smallest eigen values of $\sum_{i=1}^{K}\left(w_i\mathbf{I} - w_i\mathbf{V}_i\mathbf{V}_i^T\right)$.

3.3. Supervised Embedding with Trace Ratio

As discussed in Section 2.1, globally optimal projection directions can be obtained using trace ratio optimization in supervised embedding. This can be exploited in our proposed approach as well for supervised embedding. The only modification will be in obtaining the incoherent subspaces, $\{V_i\}_{i=1}^K$. We will incorporate the incoherence constraint of (4) into (3) and obtain the following optimization,

$$\mathbf{V}_{j+1} = \underset{\mathbf{V}^T\mathbf{V}=\mathbf{I}}{\arg\max} \frac{\operatorname{Tr}\left(\mathbf{V}^T\mathbf{X}\mathbf{H}\mathbf{X}^T\mathbf{V}\right)}{\operatorname{Tr}\left(\mathbf{V}^T\left(\mathbf{X}\mathbf{L}\mathbf{X}^T + \alpha\sum_{i=1}^{j}\left(\mathbf{V}_i\mathbf{V}_i^T\right)\right)\mathbf{V}\right)}. \tag{8}$$

This can be solved using the existing trace ratio optimization approaches such as DNM or ITR. After computing the incoherent subspaces, they can be averaged using the approach described in Section 3.2 to obtain the final subspace.

4. RESULTS AND DISCUSSION

We evaluate the proposed algorithms using unsupervised and supervised, linearized graph embedding algorithms on different standard datasets. A natural way to evaluate the usefulness of dimensionality reduction algorithms is to measure the quality of the preserved neighborhood with respect to the original high-dimensional data. From an information retrieval perspective, this is equivalent to measuring the precision/recall of the relevant neighbors. A high-quality embedding can significantly impact the interpretability and visualization of high-dimensional data. In addition, for the case of supervised embeddings, we measure the classification performance using a simple k-Nearest Neighbor (k-NN) classifier.

4.1. Unsupervised Graph Embedding

For unsupervised learning experiments, we consider Locality Preserving Projections (LPP), Neighborhood Preserving Embedding (NPE), and the proposed variants of these approaches. To infer the LPP projections, we construct a k-NN graph and compute similarities using a heat kernel, $W_{ij} = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|_2)$. For NPE, the similarity graph is constructed using a least-squares fit in the neighborhood of each sample. In both approaches, we fixed the neighborhood size, $k = 15$. For the proposed algorithm, we choose

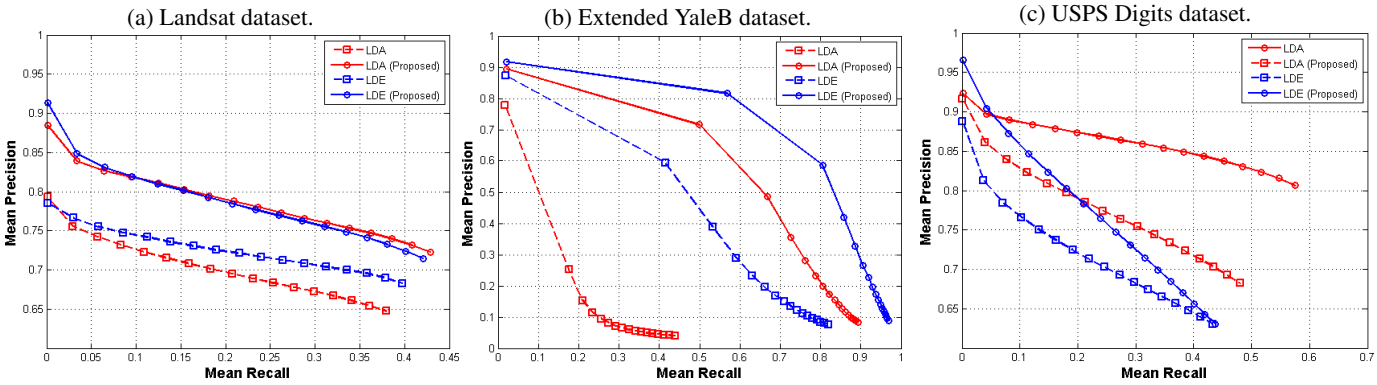

Fig. 2. Supervised Graph Embedding: Mean precision - Mean Recall curves for novel test data with respect to the training samples.

Table 1. Classification Performance obtained using different supervised embedding techniques. In each case, the performance was evaluated using a k-NN classifier and the highest performance for each dataset is shown in bold. GM denotes our proposed approach.

Dataset	LDA	LDE	LDA-GM	LDE-GM
Landsat	83.99	84.2	89.9	**91.45**
Letter	88.95	91.43	89.33	**92.25**
USPS	92.79	93.51	89.17	**96.7**
HeartDisease	69.7	82.5	77.5	**84.9**
YaleB	78.55	85.64	89.67	**91.57**
ORL	93.33	92.9	95.1	**95.7**

the size of the ensemble $K = 20$ and learn multiple incoherent subspaces with LPP and NPE respectively.

We measure the quality of the embedding by choosing $N_h = 50$ neighbors for each sample in the high-dimensional space and varying the neighborhood size N_l between 1 and 200 in the low-dimensional space. For each sample index i, we measure the precision and recall in the embedding as

$$precision(i) = \frac{|\Omega_h(i) \cap \Omega_l(i)|}{N_l}, \tag{9}$$

and

$$recall(i) = \frac{|\Omega_h(i) \cap \Omega_l(i)|}{N_h}, \tag{10}$$

where Ω_h and Ω_l denote the set of neighborhood samples for original and embedded data respectively. Finally, the set of precision-curves are generated by averaging across all data samples.

We used the following datasets for our experiments: (a) faces dataset [17] that contains 10 different face images of 40 different subjects; (b) sea-water temperature time-series dataset [18], where each data sample is a time window of 52 weeks; (c) UCI ecoli dataset of protein localization sites [19]; (d) face videos dataset [20]. In each case, we repeated experiments with varying number of embedding dimensions (N) and report the results for the case which provided the highest F-measure ($2(P.R)/(P+R)$).

Figure 1 shows the mean precision-mean recall curves for the four datasets obtained using different linear graph embedding strategies. As it can be clearly observed, the proposed algorithm significantly improves the performance of both LPP and NPE approaches. It is particularly interesting to note that it does not compromise the recall performance in order to better preserve the closest neighbors.

In order to illustrate the utility of the proposed algorithm in data visualization, we consider the UCI ecoli dataset and show the $2-$D embeddings obtained using NPE and the proposed algorithm in Figure 1(e) and (f) respectively. Though the embedding is constructed using unsupervised strategies, the relation between the different classes (color) is strongly evident in the proposed embedding and hence contains much richer information from a visualization perspective.

4.2. Supervised Graph Embedding

In the next set of experiments, two linearized graph embedding algorithms (Linear Discriminant Analysis (LDA) and Local Discriminant Embedding (LDE)) are used with the proposed technique to infer a consensus subspace for supervised classification. In both cases, the projection directions are computed using the iterative trace ratio (ITR) method. The inter-class and intra-class neighborhood sizes for learning the LDE projections were fixed at 10. Similar to the previous experiment, we used $K = 20$ subspaces to infer the consensus. In order to evaluate the performance of the embeddings, we split each dataset into train and test sets (70% for training), learned the subspace using the train set, and projected the test samples onto the inferred subspace. Similar to an information retrieval setup, we measure the precision and recall for each test sample with respect to the training samples, in the embedded space (using ground truth labels). In addition, we measure the classification accuracy using a k-NN classifier ($k = 5$).

Table 1 lists the classification performance of the different supervised embedding strategies with the following datasets: (a) UCI landsat dataset [19]; (b) UCI letter recognition dataset [19]; (c) USPS handwritten digits [19]; (d) UCI heart disease dataset [19]; (e) Extended YaleB dataset [21]; (f) ORL face database [17]. Figure 2 illustrates the mean precision-mean recall curves for the test data with different embedding strategies. Significant performance gains were obtained using the proposed algorithm in all cases.

5. CONCLUSIONS

In this paper, we proposed to build consensus of multiple incoherent subspace projections on the Grassmannian, to obtain high-quality graph embeddings. Evaluation with popular unsupervised and supervised approaches show lot of promise, and we believe it is crucial to study the theoretical characteristics of the proposed approach. In addition, kernelization of the proposed algorithm can further improve its performance on more complex datasets.

6. REFERENCES

[1] Shuicheng Yan, Dong Xu, Benyu Zhang, Hong-Jiang Zhang, Qiang Yang, and Stephen Lin, "Graph embedding and extensions: a general framework for dimensionality reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40–51, 2007.

[2] George H Dunteman, *Principal components analysis*, Sage, 1989.

[3] Joseph B Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, 1964.

[4] Joshua B Tenenbaum, Vin De Silva, and John C Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[5] Mikhail Belkin and Partha Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *NIPS*, 2001, vol. 14, pp. 585–591.

[6] Xiaofei He and Partha Niyogi, "Locality preserving projections," in *Neural information processing systems*, 2004, vol. 16, p. 153.

[7] Xiaofei He, Deng Cai, Shuicheng Yan, and Hong-Jiang Zhang, "Neighborhood preserving embedding," in *IEEE International Conference on Computer Vision*. IEEE, 2005, vol. 2, pp. 1208–1213.

[8] Peter A Lachenbruch, *Discriminant analysis*, Wiley Online Library, 1975.

[9] Hwann-Tzong Chen, Huang-Wei Chang, and Tyng-Luh Liu, "Local discriminant embedding and its variants," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2005, vol. 2, pp. 846–853.

[10] Jarkko Venna, Jaakko Peltonen, Kristian Nybo, Helena Aidos, and Samuel Kaski, "Information retrieval perspective to nonlinear dimensionality reduction for data visualization," *The Journal of Machine Learning Research*, vol. 11, pp. 451–490, 2010.

[11] Huan Wang, Shuicheng Yan, Dong Xu, Xiaoou Tang, and Thomas Huang, "Trace ratio vs. ratio trace for dimensionality reduction," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[12] Yangqing Jia, Feiping Nie, and Changshui Zhang, "Trace ratio problem revisited," *IEEE Transactions on Neural Networks*, vol. 20, no. 4, pp. 729–735, 2009.

[13] Joe Harris, *Algebraic geometry: a first course*, vol. 133, Springer, 1992.

[14] Ke Ye and Lek-Heng Lim, "Distance between subspaces of different dimensions," *arXiv preprint arXiv:1407.0900*, 2014.

[15] Xichen Sun, Liwei Wang, and Jufu Feng, "Further results on the subspace distance," *Pattern recognition*, vol. 40, no. 1, pp. 328–329, 2007.

[16] Liwei Wang, Xiao Wang, and Jufu Feng, "Subspace distance analysis with application to adaptive bayesian algorithm for face recognition," *Pattern Recognition*, vol. 39, no. 3, pp. 456–464, 2006.

[17] "Machine Learning datasets," available online at: http://cs.nyu.edu/∼roweis/data.html.

[18] A Lendasse and E. Liitiainen, "Variable scaling for time series prediction: Application to the ESTSP'07 and the NN3 forecasting competitions," in *IJCN*, Aug 2007.

[19] K. Bache and M. Lichman, "UCI machine learning repository," available online at: https://archive.ics.uci.edu/ml/index.html, 2013.

[20] "Face Video," available online at: http://www.cs.toronto.edu/∼roweis/data.html.

[21] AS. Georghiades, P.N. Belhumeur, and D. Kriegman, "From few to many: illumination cone models for face recognition under variable lighting and pose," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, Jun 2001.