2015 IEEE Symposium Series on Computational Intelligence

Graph Embedding Exploiting Subclasses

Anastasios Maronidis, Anastasios Tefas and Ioannis Pitas
Department of Informatics,
Aristotle University of Thessaloniki,
P.O.Box 451, 54124
Thessaloniki, Greece
Email: amaronidis@iti.gr, tefas@aiia.csd.auth.gr, pitas@aiia.csd.auth.gr

***Abstract*—Recently, subspace learning methods for Dimensionality Reduction (DR), like Subclass Discriminant Analysis (SDA) and Clustering-based Discriminant Analysis (CDA), which use subclass information for the discrimination between the data classes, have attracted much attention. In parallel, important work has been accomplished on Graph Embedding (GE), which is a general framework unifying several subspace learning techniques. In this paper, GE has been extended in order to integrate subclass discriminant information resulting to the novel Subclass Graph Embedding (SGE) framework. The kernelization of SGE is also presented. It is shown that SGE comprises a generalization of the typical GE including subclass DR methods. In this context, the theoretical link of SDA and CDA with SGE is established. The efficacy and power of SGE has been substantiated by comparing subclass DR methods versus a diversity of unimodal methods all pertaining to the SGE framework via a series of experiments on various real-world data.**

I. Introduction

In recent years, various subspace learning algorithms for dimensionality reduction (DR) have been developed. Locality Preserving Projections (LPP) [1] and Principal Component Analysis (PCA) [2] are two of the most popular unsupervised linear DR algorithms with multiple applications. Besides, supervised methods like Linear Discriminant Analysis (LDA) [3] have shown superior performance in many classification problems, since through the DR process they aim at achieving data class discrimination.

Usually in practice, there is the case where many data clusters appear inside the same class imposing the need to integrate this information in the DR approach. Along these lines, techniques such as Clustering Discriminant Analysis (CDA) [4] and Subclass Discriminant Analysis (SDA) [5] have been proposed. Both of them utilize a specific objective criterion that incorporates the data subclass information in an attempt to discriminate subclasses that belong to different classes, while they put no constraints to subclasses within the same class.

In parallel to the development of subspace learning techniques, a lot of work has been carried out in DR from a graph theoretic perspective. Towards this direction, Graph Embedding (GE) has been introduced as a generalized framework, which unifies several existing DR methods and furthermore offers as a platform for developing novel algorithms [6]. In [1], [6] the connection of LPP, PCA and LDA with the GE framework has been illustrated and in [6], employing GE, the authors propose Marginal Fisher Analysis (MFA), while a Subclass Marginal Fisher Analysis (SMFA) method has also been proposed in [7]. In addition, the ISOMAP [8], Locally Linear Embedding (LLE) [9] and Laplacian Eigenmaps (LE) [10] algorithms have been interpreted within the GE framework [6].

From the perspective of GE, the data are considered as vertices of a graph, which is accompanied by two matrices, the intrinsic and the penalty matrix, weighing the edges among vertices. The intrinsic matrix encodes the similarity relationships, while the penalty matrix encodes the undesirable connections among the data. In this context, the DR task is translated to the problem of transforming the initial graph into a new one in a way that the weights of the intrinsic matrix are reinforced, while the weights of the penalty matrix are suppressed.

Apart from the core idea on GE presented in [6], some other interesting works have also been published recently in the literature. A graph-based supervised DR method has been proposed in [11] for circumventing the problem of non-Gaussian distributed data. The importance degrees of the same-class and not-same-class vertices are encoded by the intrinsic and extrinsic graphs, based on a monotonically decreasing function. Moreover, the kernel extension of the proposed approach is also presented. In [12], the selection of the neighbor parameters of the intrinsic and extrinsic graph matrices is adaptively performed based on the different local manifold structure of different samples, enhancing in this way the intra-class similarity and inter-class separability.

978-1-4799-7560-0/15 $31.00 © 2015 IEEE
DOI 10.1109/SSCI.2015.206

Methodologies that convert a set of graphs into a vector space have also been presented. For instance, a novel prototype selection method from a class-labeled set of graphs has been proposed in [13]. A dissimilarity metric between a pair of graphs is established and the dissimilarities of a graph from a set of prototypes are calculated providing an n-dimensional feature vector. Several deterministic algorithms are used to select the prototypes with the most discriminative power [13]. The flexibility of GE has also been combined with the generalization ability of the support vector machine classifier resulting to improved classification performance. In [14], the authors propose the substitution of the support vector machine kernel with sub-space or sub-manifold kernels, that are constructed based on the GE framework.

Despite the intense activity around GE, no extension has been proposed to integrate subclass information. In this paper, such an extension is proposed, leading to the novel Subclass Graph Embedding (SGE) framework, which is the main contribution of our work. Using subclass block form in both the intrinsic and penalty graph matrices, SGE optimizes a criterion which preserves the subclass structure and simultaneously the local geometry of the data. One big advantage of SGE and generally of graph-based methods is that their functionality is based merely on the existence of the two above graph matrices regardless of the way they have been constructed. This allows for employing any similarity measure for modelling the local geometry and any clustering approach for extracting the subclasses of the data.

Choosing the appropriate parameters, SGE becomes one of the well-known aforementioned algorithms. Along these lines, in this paper it is shown that a variety of unimodal DR algorithms are encapsulated within SGE. Furthermore, the theoretical link between SGE and CDA, SDA methods is also established, which is another novelty of our work. Finally, the kernelization of SGE (K-SGE) is also presented. The efficacy of SGE and K-SGE is demonstrated through a comparison between subclass DR methods and a diversity of unimodal ones – all pertaining to the SGE framework – via a series of experiments on various datasets.

The remainder of this paper is organized as follows. The subspace learning algorithms CDA and SDA are presented in Section II in order to pave the way for their connection with SGE. The novel SGE framework along with its kernelization is presented in Section III. The connection between the SGE framework and the several subspace learning techniques is given in Section IV. A comparison of the aforementioned methods on real-world datasets is presented in Section V. Finally, conclusions are drawn in Section VI.

II. Subspace Learning Techniques

In this section, we provide the mathematical formulation of the subspace learning techniques CDA and SDA in order to allow their connection with the SGE framework. The other methods mentioned in the Introduction are encapsulated in the proposed SGE framework as well. However, their detailed description is omitted, as they have already been described in [6].

In the following analysis, we consider that each data sample denoted by $\mathbf{x}$ is an m-dimensional real vector, i.e., $\mathbf{x} \in \mathbb{R}^m$. We also denote by $\mathbf{y} \in \mathbb{R}^{m'}$ its projection $\mathbf{y} = \mathbf{V}^T\mathbf{x}$ to a new m'-dimensional space using a projection matrix $\mathbf{V} \in \mathbb{R}^{m\times m'}$. CDA and SDA attempt to minimize:

$$J(\mathbf{v}) = \frac{\mathbf{v}^T\mathbf{S}_W\mathbf{v}}{\mathbf{v}^T\mathbf{S}_B\mathbf{v}}\,, \tag{1}$$

where $\mathbf{S}_W$ is called the *within* and $\mathbf{S}_B$ the *between* scatter matrix [15]. These matrices are symmetric and positive semi-definite. The minimization of the ratio (1) leads to the following generalized eigenvalue decomposition problem to find the optimal discriminant projection eigenvectors:

$$\mathbf{S}_W\mathbf{v} = \lambda\mathbf{S}_B\mathbf{v}\,. \tag{2}$$

The eigenvalues λ_i of the above eigenproblem are by definition positive or zero:

$$0 \le \lambda_1 \le \lambda_2 \le \cdots \le \lambda_m\,. \tag{3}$$

Let $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_m$ be the corresponding eigenvectors. Then the projection $\mathbf{y} = \mathbf{V}^T\mathbf{x}$, from the initial space to the new space of reduced dimensionality employs the projection matrix $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_{m'}]$ whose columns are the eigenvectors $\mathbf{v}_i$, $i = 1, \ldots, m'$ and $m' \ll m$.

Looking for a linear transform that effectively separates the projected data of each class, CDA makes use of potential subclass structure. Let us denote the total number of subclasses inside the i-th class by d_i and, for the j-th subclass of the i-th class, the number of its samples by n_{ij}, its q-th sample by $\mathbf{x}_q^{ij}$ and its mean vector by $\boldsymbol{\mu}^{ij}$. CDA attempts to minimize (1), where $\mathbf{S}_W^{(CDA)}$ is the *within-subclass* and $\mathbf{S}_B^{(CDA)}$ the *between-subclass* scatter matrix, defined in [4]:

$$\mathbf{S}_W^{(CDA)} = \sum_{i=1}^{c}\sum_{j=1}^{d_i}\sum_{q=1}^{n_{ij}} \left(\mathbf{x}_q^{ij} - \boldsymbol{\mu}^{ij}\right)\left(\mathbf{x}_q^{ij} - \boldsymbol{\mu}^{ij}\right)^T, \tag{4}$$

$$\mathbf{S}_B^{(CDA)} = \sum_{i=1}^{c-1} \sum_{l=i+1}^{c} \sum_{j=1}^{d_i} \sum_{h=1}^{d_l} \left(\boldsymbol{\mu}^{ij} - \boldsymbol{\mu}^{lh}\right)\left(\boldsymbol{\mu}^{ij} - \boldsymbol{\mu}^{lh}\right)^T. \tag{5}$$

The difference between SDA and CDA mainly lies on the definition of the within scatter matrix, while the between scatter matrix of SDA is a modified version of that of CDA. The exact definitions of the two matrices are:

$$\mathbf{S}_W^{(SDA)} = \sum_{q=1}^{n} \left(\mathbf{x}_q - \boldsymbol{\mu}\right)\left(\mathbf{x}_q - \boldsymbol{\mu}\right)^T , \tag{6}$$

$$\mathbf{S}_B^{(SDA)} = \sum_{i=1}^{c-1} \sum_{l=i+1}^{c} \sum_{j=1}^{d_i} \sum_{h=1}^{d_l} p_{ij} p_{lh} \left(\boldsymbol{\mu}^{ij} - \boldsymbol{\mu}^{lh}\right)\left(\boldsymbol{\mu}^{ij} - \boldsymbol{\mu}^{lh}\right)^T, \tag{7}$$

where $p_{ij} = \frac{n_{ij}}{n}$ is the relative frequency of the j-th cluster of the i-th class [5]. It is worth mentioning that $\mathbf{S}_W^{(SDA)}$ is actually the total covariance matrix of the data.

The previously described DR methods along with LPP, PCA and LDA can be seen under a common prism, since their basic calculation element towards the construction of the corresponding optimization criteria is the similarity among the samples. Thus we can unify them in a common framework if we consider that the samples form a graph and we set criteria on the similarities between the nodes of this graph. In the following section we describe in detail this approach.

III. Subclass Graph Embedding

In this section, the problem of dimensionality reduction is described from a graph theoretic perspective. Before we present the novel SGE, let us first briefly provide the main ideas of the core GE framework.

A. Graph Embedding

In the GE framework, the set of the data samples to be projected in a low dimensionality space is represented by two graphs, namely, the *intrinsic* $G_{int} = \{\mathcal{X}, \mathbf{W}_{int}\}$ and the *penalty* $G_{pen} = \{\mathcal{X}, \mathbf{W}_{pen}\}$ graph, where $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ is the set of the data samples in both graphs. The intrinsic graph models the similarity connections between every pair of data samples that have to be reinforced after the projection. The penalty graph contains the connections between the data samples that must be suppressed after the projection. For both of the above matrices these connections might have negative values imposing the opposite effects. Choosing the values of both the intrinsic and the penalty graph matrices, may lead to either supervised, unsupervised or semi-supervised DR algorithms.

Along these lines, it is desirable to project the initial data to the new low dimensional space, such that the geometrical structure of the data is preserved. The corresponding objective function for optimization is:

$$\underset{tr\{\mathbf{YBY}^T\}=d}{\operatorname{argmin}} \; J(\mathbf{Y}) , \tag{8}$$

$$J(\mathbf{Y}) = \frac{1}{2} tr\{\sum_q \sum_p (\mathbf{y}_q - \mathbf{y}_p)\mathbf{W}_{int}(q,p)(\mathbf{y}_q - \mathbf{y}_p)^T\} , \tag{9}$$

where $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n]$ are the projected vectors, d is a constant, $\mathbf{B}$ is a constraint matrix, defined to remove an arbitrary scaling factor in the embedding and $\mathbf{W}_{int}(q,p)$ is the value of $\mathbf{W}_{int}$ at position (q,p) [6]. The structure of the objective function (9) postulates that, the larger the value $\mathbf{W}_{int}(q,p)$ is, the smaller the distance between the projections of the data samples $\mathbf{x}_q$ and $\mathbf{x}_p$ has to be. By using some simple algebraic manipulations, equation (9) becomes:

$$J(\mathbf{Y}) = tr\{\mathbf{Y}\mathbf{L}_{int}\mathbf{Y}^T\} , \tag{10}$$

where $\mathbf{L}_{int} = \mathbf{D}_{int} - \mathbf{W}_{int}$ is the intrinsic Laplacian matrix and $\mathbf{D}_{int}$ is the degree matrix defined as the diagonal matrix, which has at position (q,q) the value $\mathbf{D}_{int}(q,q) = \sum_p \mathbf{W}_{int}(q,p)$.

The Laplacian matrix $\mathbf{L}_{pen} = \mathbf{D}_{pen} - \mathbf{W}_{pen}$ of the penalty graph is often used as the constraint matrix $\mathbf{B}$. Thus (8) becomes:

$$\operatorname{argmin} \frac{tr\{\mathbf{Y}\mathbf{L}_{int}\mathbf{Y}^T\}}{tr\{\mathbf{Y}\mathbf{L}_{pen}\mathbf{Y}^T\}} . \tag{11}$$

The optimization of the above objective function is achieved by solving the generalized eigenproblem:

$$\mathbf{L}_{int}\mathbf{v} = \lambda \mathbf{L}_{pen}\mathbf{v} , \tag{12}$$

keeping the eigenvectors, which correspond to the smallest eigenvalues.

This approach leads to the optimal projection of the given data samples. In order to achieve the out of sample projection, the linearization [6] of the above approach should be used. If we employ $\mathbf{y} = \mathbf{V}^T\mathbf{x}$, the objective function (9) becomes:

$$\underset{tr\{\mathbf{V}^T\mathbf{X}\mathbf{L}_{pen}\mathbf{X}^T\mathbf{V}\}=d}{\operatorname{argmin}} \; J(\mathbf{V}) , \tag{13}$$

where $J(\mathbf{V})$ is defined as:

$$\frac{1}{2} tr\{\mathbf{V}^T \left(\sum_q \sum_p (\mathbf{x}_q - \mathbf{x}_p)\mathbf{W}_{int}(q,p)(\mathbf{x}_q - \mathbf{x}_p)^T \right) \mathbf{V}\} , \tag{14}$$

where $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$. By using simple algebraic manipulations, we have:

$$J(\mathbf{V}) = tr\{\mathbf{V}^T \mathbf{X} \mathbf{L}_{int} \mathbf{X}^T \mathbf{V}\}. \quad (15)$$

Similarly to the straight approach, the optimal eigenvectors are given by solving the generalized eigenproblem:

$$\mathbf{X}\mathbf{L}_{int}\mathbf{X}^T\mathbf{v} = \lambda \mathbf{X}\mathbf{L}_{pen}\mathbf{X}^T\mathbf{v}. \quad (16)$$

B. Linear Subclass Graph Embedding

In this section, we propose a GE framework that allows the exploitation of subclass information. In the following analysis, it is assumed that the subclass labels are known. We attempt to minimize the scatter of the data samples within the same subclass, while separating data samples from subclasses that belong to different classes. Finally, we are not concerned about samples that belong to different subclasses of the same class.

Usually, in real-world problems, local geometry of the data is related to the global supervised structure. Samples that belong to the same class or subclass, should be "sufficiently close" to each other. SGE actually exploits this fact. It simultaneously handles supervised and unsupervised information. As a consequence, it combines the global labeling information with the local geometrical characteristics of the data samples. This is achieved by weighing the above connections with the similarities of the data samples. The *Gaussian similarity function* (see eq. 17), has been used in this paper for this purpose.

$$S_{qp} = S(\mathbf{x}_q, \mathbf{x}_p) = \exp\left(-\frac{d^2(\mathbf{x}_q, \mathbf{x}_p)}{\sigma^2}\right), \quad (17)$$

where $d(\mathbf{x}_q, \mathbf{x}_p)$ is a distance metric (e.g., Euclidean) and σ^2 is a parameter (variance) that determines the distance scale.

Let us denote as $\mathbf{P}$ an affinity matrix. Without limiting the generality, we assume that this matrix has block form, depending on the subclass and the class of the data samples. Using the linearized approach, we attempt to optimize a more general discrimination criterion. We consider again that $\mathbf{y} = \mathbf{V}^T\mathbf{x}$ is the projection of $\mathbf{x}$ to the new subspace. Let $\mathbf{P}^{ij}(q,p)$ be the value of $\mathbf{P}$ at position (q,p) of the submatrix that contains the j-th subclass of the i-th class. Then, the proposed criterion is:

$$\operatorname{argmin} J(\mathbf{Y}), \quad (18)$$

$$J(\mathbf{Y}) = \frac{1}{2} tr\{\sum_{i=1}^{c}\sum_{j=1}^{d_i}\sum_{q=1}^{n_{ij}}\sum_{p=1}^{n_{ij}} \left(\mathbf{y}_q^{ij} - \mathbf{y}_p^{ij}\right) \mathbf{P}^{ij}(q,p) \left(\mathbf{y}_q^{ij} - \mathbf{y}_p^{ij}\right)^T\} \quad (19)$$

$$= \frac{1}{2} tr\{\mathbf{V}^T \left(\sum_{i=1}^{c}\sum_{j=1}^{d_i}\sum_{q=1}^{n_{ij}}\sum_{p=1}^{n_{ij}} \left(\mathbf{x}_q^{ij} - \mathbf{x}_p^{ij}\right) \mathbf{P}^{ij}(q,p) \left(\mathbf{x}_q^{ij} - \mathbf{x}_p^{ij}\right)^T \right) \mathbf{V}\} \quad (20)$$

$$= tr\{\mathbf{V}^T \mathbf{X} \left(\mathbf{D}_{int} - \mathbf{W}_{int}\right) \mathbf{X}^T \mathbf{V}\} \quad (21)$$

$$= tr\{\mathbf{V}^T \mathbf{X} \mathbf{L}_{int} \mathbf{X}^T \mathbf{V}\}. \quad (22)$$

The derivation of (22) is omitted due to lack of space. The matrix $\mathbf{W}_{int}$ is block diagonal with blocks that correspond to each class and is given by:

$$\mathbf{W}_{int} = \begin{pmatrix} \mathbf{W}_{int}^1 & & & 0 \\ & \mathbf{W}_{int}^2 & & \\ & & \ddots & \\ 0 & & & \mathbf{W}_{int}^c \end{pmatrix}. \quad (23)$$

$\mathbf{W}_{int}^i$ are block diagonal submatrices, with blocks that correspond to the subclasses and are given by:

$$\mathbf{W}_{int}^i = \begin{pmatrix} \mathbf{P}^{i1} & & & 0 \\ & \mathbf{P}^{i2} & & \\ & & \ddots & \\ 0 & & & \mathbf{P}^{id_i} \end{pmatrix}. \quad (24)$$

$\mathbf{P}^{ij}$ is the submatrix of $\mathbf{P}$ that corresponds to the data of the j-th cluster of the i-th class. By looking carefully at the form of $\mathbf{W}_{int}$, it is clear that the degree intrinsic matrix $\mathbf{D}_{int}$ has values

$$\mathbf{D}_{int}(\sum_{s=0}^{i-1}\sum_{t=0}^{j-1} n_{st}+q, \sum_{s=0}^{i-1}\sum_{t=0}^{j-1} n_{st}+q) = \sum_{p} \mathbf{P}^{ij}(q,p), \quad (25)$$

where p runs over the indices of the j-th cluster of i-th class.

In parallel, we demand to maximize a criterion, which encodes the similarities among the centroid vectors of the subclasses. Let the value Q_{ij}^{lh} express the similarity between the centroid vectors $\boldsymbol{\mu}^{ij}$ and $\boldsymbol{\mu}^{lh}$. The more similar two centroids that belong to different classes are, the further apart their projections $\mathbf{m}^{ij} = \mathbf{V}^T\boldsymbol{\mu}^{ij}$ have to be from each other:

$$\operatorname{argmax} G(\mathbf{m}^{ij}), \quad (26)$$

$$G(\mathbf{m}^{ij}) = tr\{\sum_{i=1}^{c-1}\sum_{l=i+1}^{c}\sum_{j=1}^{d_i}\sum_{h=1}^{d_l} \left(\mathbf{m}^{ij} - \mathbf{m}^{lh}\right) Q_{ij}^{lh} \left(\mathbf{m}^{ij} - \mathbf{m}^{lh}\right)^T\} \quad (27)$$

$$= tr\{\mathbf{V}^T \left(\sum_{i=1}^{c-1} \sum_{l=i+1}^{c} \sum_{j=1}^{d_i} \sum_{h=1}^{d_l} \left(\boldsymbol{\mu}^{ij} - \boldsymbol{\mu}^{lh} \right) Q_{ij}^{lh} \left(\boldsymbol{\mu}^{ij} - \boldsymbol{\mu}^{lh} \right)^T \right) \mathbf{V} \} \quad (28)$$

$$= tr\{\mathbf{V}^T \mathbf{X} \left(\mathbf{D}_{pen} - \mathbf{W}_{pen} \right) \mathbf{X}^T \mathbf{V} \} \quad (29)$$

$$= tr\{\mathbf{V}^T \mathbf{X} \mathbf{L}_{pen} \mathbf{X}^T \mathbf{V} \} . \quad (30)$$

The block matrix $\mathbf{W}_{pen}$ in (29) consists of block submatrices:

$$\mathbf{W}_{pen} = \begin{pmatrix} \mathbf{W}_{pen}^{1,1} & \mathbf{W}_{pen}^{1,2} & \cdots & \mathbf{W}_{pen}^{1,c} \\ \mathbf{W}_{pen}^{2,1} & \mathbf{W}_{pen}^{2,2} & \cdots & \mathbf{W}_{pen}^{2,c} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{W}_{pen}^{c,1} & \mathbf{W}_{pen}^{c,2} & \cdots & \mathbf{W}_{pen}^{c,c} \end{pmatrix} . \quad (31)$$

The submatrices $\mathbf{W}_{pen}^{i,i}$ lying on the main block diagonal are given by:

$$\mathbf{W}_{pen}^{i,i} = \begin{pmatrix} \mathbf{W}^{i1} & & & \\ & \mathbf{W}^{i2} & & 0 \\ & & \ddots & \\ 0 & & & \mathbf{W}^{id_i} \end{pmatrix} , \quad (32)$$

where $\mathbf{W}^{ij}$ corresponds to the j-th subclass of the i-th class and is given by:

$$\mathbf{W}^{ij} = -\frac{\left(\sum_{\omega \neq i} \left(\sum_{t=1}^{d_\omega} Q_{ij}^{\omega t} \right) \right)}{\left(n_{ij} \right)^2} \mathbf{e}^{n_{ij}} \left(\mathbf{e}^{n_{ij}} \right)^T , \quad (33)$$

where $\mathbf{e}^{n_{ij}} = [\overbrace{11\cdots1}^{n_{ij}\text{-times}}]^T$. Respectively, the off-diagonal submatrices of $\mathbf{W}_{pen}$ are given by:

$$\mathbf{W}_{pen}^{i,l} = \begin{pmatrix} \mathbf{W}_{i1}^{l1} & \mathbf{W}_{i1}^{l2} & \cdots & \mathbf{W}_{i1}^{ld_l} \\ \mathbf{W}_{i2}^{l1} & \mathbf{W}_{i2}^{l2} & \cdots & \mathbf{W}_{i2}^{ld_l} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{W}_{id_i}^{l1} & \mathbf{W}_{id_i}^{l2} & \cdots & \mathbf{W}_{id_i}^{ld_l} \end{pmatrix} , i \neq l , \quad (34)$$

where:

$$\mathbf{W}_{ij}^{lh} = \frac{Q_{ij}^{lh}}{n_{ij} n_{lh}} \mathbf{e}^{n_{ij}} \left(\mathbf{e}^{n_{lh}} \right)^T . \quad (35)$$

It can be easily shown that $\mathbf{D} = \mathbf{0}$, so that $\mathbf{L}_{pen} = -\mathbf{W}_{pen}$.

C. Kernel Subclass Graph Embedding

In this section, the kernelization of SGE is presented. Let us denote by $\mathcal{X}$ the initial data space, by $\mathcal{F}$ a Hilbert space and by f the non-linear mapping function from $\mathcal{X}$ to $\mathcal{F}$. The main idea is to firstly map the original data from the initial space into another high-dimensional Hilbert space and then perform linear subspace analysis in that space. If we denote by $m_{\mathcal{F}}$ the dimensionality of the Hilbert space, then the above procedure is described as:

$$\mathcal{X} \ni \mathbf{x}_q \rightarrow \mathbf{y}_q = f(\mathbf{x}_q) = \begin{pmatrix} \sum_{p=1}^{n} a_{1p} k(\mathbf{x}_q, \mathbf{x}_p) \\ \vdots \\ \sum_{p=1}^{n} a_{m_{\mathcal{F}} p} k(\mathbf{x}_q, \mathbf{x}_p) \end{pmatrix} \in \mathcal{F} , \quad (36)$$

where k is the kernel function. From the above equation it is obvious that

$$\mathbf{Y} = \mathbf{A}^T \mathbf{K} , \quad (37)$$

where $\mathbf{K}$ is the Gram matrix, which has at position (q, p) the value $K_{qp} = k(\mathbf{x}_q, \mathbf{x}_p)$ and

$$\mathbf{A} = [\mathbf{a}_1 \cdots \mathbf{a}_{m_{\mathcal{F}}}] = \begin{pmatrix} a_{11} & \cdots & a_{m_{\mathcal{F}}1} \\ \vdots & \ddots & \vdots \\ a_{1n} & \cdots & a_{m_{\mathcal{F}}n} \end{pmatrix} \quad (38)$$

is the map coefficient matrix. Consequently, the final SGE optimization becomes:

$$\operatorname{argmin} \frac{tr\{\mathbf{A}^T \mathbf{K} \mathbf{L}_{int} \mathbf{K} \mathbf{A}\}}{tr\{\mathbf{A}^T \mathbf{K} \mathbf{L}_{pen} \mathbf{K} \mathbf{A}\}} . \quad (39)$$

Similarly to the linear case, in order to find the optimal projections, we resolve the generalized eigenproblem:

$$\mathbf{K} \mathbf{L}_{int} \mathbf{K} \mathbf{a} = \lambda \mathbf{K} \mathbf{L}_{pen} \mathbf{K} \mathbf{a} , \quad (40)$$

keeping the eigenvectors that correspond to the smallest eigenvalues.

IV. SGE as a General Dimensionality Reduction Framework

In this section, it is shown that SGE is a generalized framework that can be used for subspace learning, since all the standard approaches are specific cases of SGE. Let us use the *Gaussian similarity function* (17), in order to construct the affinity matrix.

In the following analysis, we initially let the variance of Gaussian σ^2 tend to infinity. Hence,

$$S(\mathbf{x}_q, \mathbf{x}_p) = 1, \forall (q, p) \in \{1, 2, \cdots, n\}^2 .$$

Let the intrinsic matrix elements be:

$$\mathbf{P}^{ij}(q, p) = \begin{cases} \frac{S(\mathbf{x}_q, \mathbf{x}_p)}{n_{ij}} = \frac{1}{n_{ij}} & , \text{if } \mathbf{x}_q, \mathbf{x}_p \in \mathcal{C}_{ij} \\ 0 & , \text{otherwise} \end{cases} , \quad (41)$$

where $\mathcal{C}_{ij}$ is the set of the samples that belong to the j-th subclass of the i-th class. Obviously, (20) becomes the within-subclass criterion of CDA (also

see eq. 4). Thus, in this case, $\mathbf{W}_{int}$ is the intrinsic graph matrix of CDA. Let also:

$$Q_{ij}^{lh} = S(\boldsymbol{\mu}^{ij}, \boldsymbol{\mu}^{lh}) = 1, \forall\, i, j, h, l \tag{42}$$

the penalty matrix elements. Then, (28) becomes the between-subclass criterion of CDA (also see eq. 5). Thus, $\mathbf{W}_{pen}$ is the penalty graph matrix of CDA and the connection between CDA and GE has been established.

Let us consider that each data sample constitutes its own class, i.e., $c = n$, $d_i = 1$ and $n_i = 1$, $\forall i \in \{1, 2, \cdots, c\}$. Thus, each class-block of the penalty graph matrix reduces to a single element of the matrix. Obviously, each data sample coincides with the mean of its class. By setting:

$$Q_{i1}^{l1} = \frac{S(\boldsymbol{\mu}^{i}, \boldsymbol{\mu}^{l})}{n} = \frac{1}{n}, \forall\, (i, l) \in \{1, 2, \cdots, c\}^2\,, \tag{43}$$

then:

$$-\frac{\left(\sum_{\omega \neq i}\left(\sum_{t=1}^{d_\omega} Q_{i1}^{\omega t}\right)\right)}{(n_i)^2} = -\sum_{\omega \neq i}\left(\frac{1}{n}\right) = \frac{1}{n} - 1\,. \tag{44}$$

These values lie on the main diagonal of the penalty graph matrix. Regarding the off diagonal elements we have:

$$\frac{Q_{i1}^{l1}}{n_i n_l} = \frac{1}{n}\,. \tag{45}$$

It can be easily shown that the degree penalty matrix is $\mathbf{D} = \mathbf{0}$, so that $\mathbf{L}_{pen} = -\mathbf{W}_{pen}$. Obviously, $\mathbf{L}_{pen} = \mathbf{I} - \frac{1}{n}\mathbf{e}^n \left(\mathbf{e}^n\right)^T$ and $\mathbf{X}\mathbf{L}_{pen}\mathbf{X}^T$ becomes the covariance matrix $\mathbf{C}$ of the data. By using as intrinsic graph matrix the identity matrix, SGE becomes identical to PCA:

$$\operatorname{argmin} \frac{tr\{\mathbf{V}^T\mathbf{X}\mathbf{L}_{int}\mathbf{X}^T\mathbf{V}\}}{tr\{\mathbf{V}^T\mathbf{X}\mathbf{L}_{pen}\mathbf{X}^T\mathbf{V}\}} = \operatorname{argmin} \frac{tr\{\mathbf{V}^T\mathbf{I}\mathbf{V}\}}{tr\{\mathbf{V}^T\mathbf{C}\mathbf{V}\}} \tag{46}$$

leading to the following generalized eigenproblem:

$$\mathbf{I}\mathbf{v} = \lambda\mathbf{C}\mathbf{v}\,, \tag{47}$$

solved by keeping the smallest eigenvalues, or by setting $\mu = \frac{1}{\lambda}$, since $\lambda \neq 0$, this leads to:

$$\mathbf{C}\mathbf{v} = \mu\mathbf{I}\mathbf{v}\,, \tag{48}$$

solved by keeping the greatest eigenvalues, which is obviously the PCA solution.

Now, consider that every class consists of a unique subclass, thus $d_i = 1, \forall i \in \{1, 2, \ldots, c\}$. If we set:

$$\mathbf{P}(q, p) = \begin{cases} \frac{S(\mathbf{x}_q, \mathbf{x}_p)}{n_i} = \frac{1}{n_i} & \text{, if } \mathbf{x}_q, \mathbf{x}_p \in \mathcal{C}_i \\ 0 & \text{, otherwise} \end{cases}, \tag{49}$$

then the intrinsic graph matrix becomes that of LDA. Furthermore, if we set:

$$Q_{i1}^{l1} = \frac{n_i n_l}{n}\,, \forall\, (i, l) \in \{1, \ldots, c\}^2 \tag{50}$$

then

$$-\frac{\left(\sum_{\omega \neq i}\left(\sum_{t=1}^{d_\omega} Q_{i1}^{\omega t}\right)\right)}{(n_i)^2} = \frac{n_i - n}{n n_i} \tag{51}$$

and

$$\frac{Q_{i1}^{l1}}{n_i n_l} = \frac{1}{n}\,. \tag{52}$$

These are the values of the penalty graph matrix of LDA. So, by taking the Laplacians of the above matrices, we end up to the LDA algorithm.

Let us now reject the assumption that the variance of Gaussian tends to infinity. Consider that there is only one class which contains the whole set of the data, i.e., $c = 1$. Also consider that there are no subclasses within this unique class, i.e., $d_1 = 1$. In this case the intrinsic graph matrix becomes equal to $\mathbf{P}$. Thus, by setting $\mathbf{P}$ equal to the affinity matrix $\mathbf{S}$, the intrinsic Laplacian matrix becomes that of LPP.

We observe that by utilizing the identity matrix as the penalty Laplacian matrix, obviously we get the LPP algorithm. Since we consider a unique class, which contains a unique subclass, from (31) and (32) we have that $\mathbf{W}_{pen} = \mathbf{W}^{11}$. The values of $\mathbf{W}^{11}$ are given from (33), which in this case reduces to:

$$\mathbf{W}^{11} = -\frac{Q_{11}^{11}}{n^2}\mathbf{e}^n \left(\mathbf{e}^n\right)^T\,. \tag{53}$$

If we set:

$$Q_{11}^{11} = \frac{n^2}{1 - n}\,, \tag{54}$$

then $\mathbf{W}_{pen} = \mathbf{W}^{11} = \frac{1}{n-1}\mathbf{e}^n \left(\mathbf{e}^n\right)^T$. Consequently,

$$\mathbf{L}_{pen} = \begin{pmatrix} 1 & \frac{1}{1-n} & \cdots & \frac{1}{1-n} \\ \frac{1}{1-n} & 1 & \cdots & \frac{1}{1-n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{1-n} & \frac{1}{1-n} & \cdots & 1 \end{pmatrix}. \tag{55}$$

Thus, if we make the assumption that the number of the data-samples becomes very large, then asymptotically we have $\mathbf{L}_{pen} = \mathbf{I}$.

Finally, to complete the analysis, if we consider as the intrinsic Laplacian matrix, the matrix

$$\mathbf{L}_{int} = \mathbf{I} - \frac{1}{n}\mathbf{e}^n \left(\mathbf{e}^n\right)^T \tag{56}$$

and if we set:

$$Q_{ij}^{lh} = \frac{n_{ij} n_{lh}}{n}\,, \tag{57}$$

TABLE I: Dimensionality Reduction Using SGE Framework.

	$\mathbf{P}(\mathbf{L}_{int})$	$Q(\mathbf{L}_{pen})$	σ^2	c	d_i	d
LPP	$\mathbf{P}^{11}(q,p) = \exp\left(-\frac{d^2(\mathbf{x}_q,\mathbf{x}_p)}{\sigma^2}\right), \forall \mathbf{x}_q, \mathbf{x}_p$	$Q_{11}^{11} = \frac{n^2}{1-n}\,(\mathbf{L}_{pen} = \mathbf{I})$	σ^2	1	1	1
PCA	$\mathbf{L}_{int} = \mathbf{I}$	$Q_{i1}^{l1} = \frac{1}{n}$	∞	n	1	n
LDA	$\mathbf{P}^{i1}(q,p) = \frac{1}{n_i}\,, \mathbf{x}_q, \mathbf{x}_p \in c_i$	$Q_{i1}^{l1} = \frac{n_i n_l}{n}$	∞	c	1	c
CDA	$\mathbf{P}^{ij}(q,p) = \frac{1}{n_{ij}}\,, \mathbf{x}_q, \mathbf{x}_p \in c_{ij}$	$Q_{ij}^{lh} = 1$	∞	c	d_i	d
SDA	$\mathbf{L}_{int} = \mathbf{I} - \frac{1}{n}\mathbf{e}^n\,(\mathbf{e}^n)^T$	$Q_{ij}^{lh} = \frac{n_{ij} n_{lh}}{n}$	∞	c	d_i	d

in (33) and (35), SGE becomes identical to SDA. The parameters that determine the connection of the several methods with SGE are pooled in Table I.

V. Experimental Results

We conducted *5-fold cross-validation* classification experiments on several real-world datasets using the proposed linear and kernel (RBF) SGE framework. For extracting automatically the subclass structure, we have utilized the multiple Spectral Clustering technique [16], keeping the most plausible partition for each dataset. For classifying the data, the Nearest Centroid (NC) classifier has been used with LPP, PCA and LDA algorithms, while the Nearest Cluster Centroid (NCC) [17] has been used with CDA and SDA algorithms. In NCC, the cluster centroids are calculated and the test sample is assigned to the class of the nearest cluster centroid. NC and NCC were selected because they provide the optimal classification solutions in Bayesian terms, thus proving whether the DR methods have reached the goal described by their specific criterion.

A. Classification experiments

For the classification experiments, we have used diverse publicly available datasets offered for various classification problems. More specifically, FERAIIA, BU, JAFFE and KANADE were used for facial expression recognition, XM2VTS for face frontal view recognition, while MNIST and SEMEION for optical digit recognition. Finally, IONOSPHERE, MONK and PIMA were used in order to further extend our experimental study to diverse data classification problems.

The classification accuracy rates for the several subspace learning methods over the utilized datasets are summarized in Table II. The optimal dimensionality of the projected space that returned the above results is depicted in parenthesis. For each dataset, the best performance rate among linear and kernel methods separately is highlighted with bold, while the best overall performance rate among all methods, both linear and kernel, is surrounded by a rectangle. The classification performance rank of each method is also referred in the last two rows of Table II. Specific Rank denotes the method rank for the linear and the kernel methods, independently. Overall Rank refers to the rank of each method among both the linear and the kernel methods. The ranking has been achieved through a post-hoc Bonferroni test [18].

An immediate remark from Table II is that in both linear and kernel case, multimodal methods exhibit better classification performance than the unimodal ones. In particular, the top overall performance is shown by SDA followed by CDA, while the worst performance is shown by KLPP and KPCA. This result undoubtedly shows that the inclusion of subclass information in the DR process offers a strong potential to improve the performance of the state-of-the-art in many classification domains.

In comparing linear with kernel methods, a simple calculation yields mean overall rank equal to 5.08 for the linear methods and 5.90 for the kernel ones. Although the average performance of linear methods is clearly better than that of kernel ones, we must admit that there is ample space for improving the kernel results by varying the RBF parameter, as the selection of this parameter is not trivial and may easily lead to over-fitting. Actually, the top performance rates presented in this paper have been obtained by testing indicative values of the above parameter. As a matter of fact, it is interesting to observe that the use of kernels proves to be beneficial for some methods in certain datasets, while deteriorates the performance of others.

VI. Conclusions

In this paper, data subclass information has been incorporated within Graph Embedding (GE) leading to a novel Subclass Graph Embedding (SGE) framework, which constitutes the main contribution of our work. In particular, it has been shown that SGE comprises a generalization of GE, encapsulating a number of state-of-the-art unimodal subspace learning techniques already integrated within GE. Besides, the connection of SGE with subspace learning algorithms that use subclass information in the embedding process has been analytically proven. The kernelization of SGE has also been presented.

TABLE II: Classification accuracy (%) of linear and kernel methods on several real-world datasets.

DATASET	LPP	PCA	LDA	CDA	SDA	KLPP	KPCA	KDA	KCDA	KSDA
FER-AIIA	40.9(3)	31.0(120)	64.6(6)	73.2	**75.5(11)**	50.2(252)	41.5(29)	54.9(6)	**56.1(12)**	53.5(12)
BU	39.4(298)	38.1(49)	51.6(6)	49.1(16)	**52.3(15)**	**52.7(317)**	35.9(290)	46.6(6)	41.0(13)	48.0(14)
JAFFE	46.8(18)	37.6(39)	53.2(6)	40.0(15)	**54.1(6)**	28.8(98)	25.9(58)	42.4(6)	36.1(18)	**46.3(5)**
KANADE	34.2(92)	43.3(46)	**67.1(6)**	59.7(7)	**67.1(5)**	32.7(99)	33.2(88)	**44.3(6)**	40.0(6)	38.5(6)
MNIST	71.1(259)	79.9(135)	84.6(9)	84.8(15)	**85.1(14)**	81.4(299)	64.5(155)	**86.0(9)**	83.4(19)	85.2(15)
SEMEION	53.6(99)	83.2(55)	88.2(9)	89.2(19)	**89.4(19)**	83.8(99)	77.4(77)	95.3(9)	94.1(19)	**95.9(19)**
XM2VTS	95.7(54)	92.0(86)	70.5(1)	**98.1(3)**	97.4(2)	71.3(297)	**74.7(56)**	61.3(1)	71.5(3)	57.3(4)
IONOSPHERE	**84.6(23)**	72.3(15)	78.9(1)	80.6(2)	83.4(2)	83.7(23)	70.3(2)	92.9(1)	**93.1(1)**	92.9(1)
MONK 1	66.7(3)	68.3(5)	50.8(1)	70.0(4)	**74.2(3)**	63.3(2)	**72.5(1)**	55.8(1)	58.3(4)	61.7(3)
MONK 2	**56.0(1)**	53.3(4)	52.0(1)	54.2(1)	54.0(2)	54.8(1)	59.8(3)	69.7(1)	**78.7(1)**	54.5(1)
MONK 3	77.2(5)	**80.9(4)**	49.4(1)	74.6(2)	66.3(2)	62.5(2)	**79.2(5)**	51.7(1)	67.5(2)	58.3(1)
PIMA	61.8(1)	63.5(6)	56.5(1)	60.5(3)	**73.5(3)**	50.7(3)	**67.5(4)**	48.9(1)	52.5(3)	52.9(1)
SPECIFIC RANK	3.3	3.8	3.6	2.5	1.6	3.5	3.4	2.9	2.4	2.7
OVERALL RANK	5.8	6.4	6.0	4.2	3.0	6.7	6.7	5.4	5.2	5.5

Through an extensive experimental study, it has been shown that subclass learning techniques outperform a number of state-of-the-art unimodal learning methods in many real-world datasets pertaining to various classification domains. In addition, although the superiority of linear methods over kernel ones is evident, there is ample space for improving kernel methods by optimizing the involved parameters.

In the near future, we intend to employ SGE as a template to design novel DR methods. For instance, as current subclass methods are strongly dependent on the underlying distribution of the data, we anticipate that novel methods, which use neighbourhood information among the data of the several subclasses, will succeed in alleviating this sort of limitations.

REFERENCES

[1] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, "Face recognition using laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 27, no. 3, pp. 328–340, 2005.

[2] I. Jolliffe, *Principal Component Analysis*. Springer Verlag, 1986.

[3] D. J. Kriegman, J. P. Hespanha, and P. N. Belhumeur, "Eigenfaces vs. fisherfaces: Recognition using class-specific linear projection," in *ECCV*, 1996, pp. I:43–58.

[4] X. W. Chen and T. S. Huang, "Facial expression recognition: A clustering-based approach," *Pattern Recognition Letters*, vol. 24, no. 9-10, pp. 1295–1302, Jun. 2003.

[5] M. L. Zhu and A. M. Martinez, "Subclass discriminant analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 8, pp. 1274–1286, Aug. 2006.

[6] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 1, pp. 40–51, 2007.

[7] A. Maronidis, A. Tefas, and I. Pitas, "Subclass graph embedding and a marginal fisher analysis paradigm," *Pattern Recognition*, vol. 48, no. 12, pp. 4024–4035, 2015.

[8] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction." *Science*, vol. 290, no. 5500, pp. 2319–2323, dec 2000.

[9] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding." *Science*, vol. 290, no. 5500, pp. 2323–2326, dec 2000.

[10] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering." *Advances in Neural Information Processing Systems (NIPS)*, vol. 14, pp. 585–591, 2001.

[11] Y. Cui and L. Fan, "A novel supervised dimensionality reduction algorithm: Graph-based fisher analysis," *Pattern Recognition*, vol. 45, no. 4, pp. 1471–1481, 2012.

[12] J. Shi, Z. Jiang, and H. Feng, "Adaptive graph embedding discriminant projections," *Neural Processing Letters*, pp. 1–16, 2013.

[13] E. Zare Borzeshi, M. Piccardi, K. Riesen, and H. Bunke, "Discriminative prototype selection methods for graph embedding," *Pattern Recognition*, 2012.

[14] G. Arvanitidis and A. Tefas, "Exploiting graph embedding in support vector machines," in *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*. IEEE, 2012, pp. 1–6.

[15] R. A. Fisher, "The statistical utilization of multiple measurements." *Annals of Eugenics*, vol. 8, pp. 376–386, 1938.

[16] A. Azran and Z. Ghahramani, "Spectral methods for automatic multiscale data clustering." in *IEEE Computer Vision and Pattern Recognition (CVPR) (1)*. IEEE Computer Society, 2006, pp. 190–197.

[17] A. Maronidis, A. Tefas, and I. Pitas, "Frontal view recognition using spectral clustering and subspace learning methods." in *ICANN (1)*, ser. Lecture Notes in Computer Science, W. D. K. I. Diamantaras and L. S. Iliadis, Eds., vol. 6352. Springer, 2010, pp. 460–469.

[18] O. J. Dunn, "Multiple comparisons among means," *Journal of American Statistical Association*, vol. 56, no. 293, pp. 52–64, 1961.