2015 IEEE Symposium Series on Computational Intelligence

Subclass Marginal Fisher Analysis

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*—Subspace learning techniques have been extensively used for dimensionality reduction (DR) in many pattern classification problem domains. Recently, Discriminant Analysis (DA) methods, which use subclass information for the discrimination between the data classes, have attracted much attention. As DA methods are strongly dependent on the underlying distribution of the data, techniques whose functionality is based on neighbourhood information among the data samples have emerged. For instance, based on the Graph Embedding (GE) framework, which is a platform for developing novel DR methods, Marginal Fisher Analysis (MFA) has been proposed. Although MFA surpasses the above distribution limitations, it fails to model potential subclass structure that might lie within the several classes of the data. In this paper, motivated by the need to alleviate the above shortcomings, we propose a novel DR technique, called Subclass Marginal Fisher Analysis (SMFA), which combines the strength of subclass DA methods with the versatility of MFA. The new method is built by extending the GE framework so as to include subclass information. Through a series of experiments on various real-world datasets, it is shown that SMFA outperforms in most of the cases the state-of-the-art demonstrating the potential of exploiting subclass neighbourhood information in the DR process.**

I. Introduction

Dimensionality reduction (DR) is an important process for achieving efficient pattern classification. In recent years, a variety of subspace learning algorithms for DR has been developed. Locality Preserving Projections (LPP) [1], [2] and Principal Component Analysis (PCA) [3] are two of the most popular unsupervised linear DR algorithms with a wide range of applications. Besides, supervised methods like Linear Discriminant Analysis (LDA) [4] have shown superior performance in many classification problems, since through the DR process they aim at achieving data class discrimination.

In practice, usually there is the case that many data clusters appear inside the same class imposing the need to integrate this information in the DR process. Along these lines, techniques such as Clustering Discriminant Analysis (CDA) [5] and Subclass Discriminant Analysis (SDA) [6] have been proposed. Both of them utilize a specific objective criterion that incorporates data subclass information aiming to discriminate subclasses that belong to different classes, while putting no constraints to subclasses within the same class.

Although the above methods have proven their potential in various classification problems, their correct performance is highly dependent on specific assumptions with respect to the underlying distribution of the data samples [4]. Since in real-world problems such assumptions are rarely satisfied, it is clear that there is a need to overcome the limitations related to the above methods. Towards this end, in [7], the authors have presented a Graph Embedding (GE) framework, which serves as a platform to develop new DR methods. Using GE, they have proposed Marginal Fisher Analysis (MFA), which uses neighbourhood information among adjacent samples within and between the classes of a dataset. The advantage of MFA is that it models the intra-class compactness and the inter-class separability using vicinity information among the samples ignoring the underlying distribution of the data classes.

Although MFA overcomes the limitations related to class distribution, it totally defies potential structure within the classes in the form of subclasses. Such structure is anticipated to provide DR process with crucial information, which may allow better discrimination of the classes. In this paper, extending the GE framework so as to include subclass information [8], we propose a novel Subclass Marginal Fisher Analysis (SMFA) algorithm for supervised dimensionality reduction. The new method combines the modularity of subclass based methods with the strength of MFA, as it models the margins among classes using neighbourhood information between

978-1-4799-7560-0/15 $31.00 © 2015 IEEE
DOI 10.1109/SSCI.2015.198

the samples belonging to the several subclasses. This combination enables SMFA to overcome the shortcomings stemming from the distribution constraints of the data leading to improved classification performance. As a matter of fact, through an experimental comparison, it is shown that our method outperforms a number of state-of-the-art dimensionality reduction methods in terms of classification accuracy.

The rest of this paper is organized as follows. A literature review of related work is presented in Section II. The GE framework, which is employed for developing our method is described in Section III, while the novel SMFA method along with its kernelization is presented in Section IV. A comparison of SMFA with all the state-of-the-art subspace methods mentioned in the Introduction is conducted in Section V on a number of real-world datasets. Finally, conclusions are drawn in Section VI.

II. Related Work

Although LDA proves to be an effective method in many classification problems, it encounters some fundamental limitations. For instance, it suffers from the *small sample size* problem, which occurs when the number of the training samples is smaller than the data dimensionality. In this case, LDA fails to optimize its objective criterion, due to the singularity of the involved matrices. A solution to this problem has been provided in [9], where the authors propose the use of the pseudo-inverse of a matrix, in order to overcome matrix singularity. Another approach is the utilization of PCA as a preprocessing step to reduce data dimensionality and then, the application of LDA, resulting to the combined PCA + LDA method [4].

For overcoming the singularity problem, regularization techniques have also been employed [10]. Moreover, in an indirect way to deal with the singularity problem, a 2D-LDA method, where the data are represented as matrices has been proposed in [11]. As has been clearly stated in [12], an additional problem appears when some of the smallest eigenvalues of the within matrix correspond to noisy features of the data. A factorization that prunes the noisy bases of the within matrix and a correlation-based criterion have been proposed in [12] for solving these problems.

Another strong limitation is that LDA postulates that the data class samples have multivariate Gaussian distribution, common covariance matrix and different means, for achieving the optimal discrimination in Bayesian terms [13]. In real problems though, the class data might not be normally distributed. Many extensions of LDA have been proposed in the literature for circumventing these limitations [14], [15], [16], [17]. Amongst the most effective methods towards this end is Marginal Fisher Analysis [7] designed based on the Graph Embedding framework. MFA uses adjacency information among the data samples and succeeds in overcoming the above-mentioned distribution limitations. However, MFA ignores information stemming from potential subclass structure within the data classes.

As already mentioned in the Introduction, CDA and SDA have been proposed for exploiting subclass structure of the data. Along the same lines, a Mixture Subclass Discriminant Analysis (MSDA) method that modifies the objective function of SDA has been proposed in [18]. Moreover, the link between MSDA and the Gaussian mixture model has been accomplished using the Expectation-Maximization framework. In the same work, MSDA has further been extended in several ways so that the subclass separation problem is solved and nonlinearly separable subclass structure has been tackled using the kernel trick. In [19], a Multiple-Exemplar Discriminant Analysis (MEDA) method is presented. The classes are represented by some exemplar vectors. Using these exemplars, an objective criterion is constructed. In this vein, the subclass means can be used as exemplars, hence exploiting the subclass structure of the data.

Subspace learning and clustering have been treated together into an iterative process in [20]. Intra-cluster similarity and inter-cluster separability are enhanced using initial cluster estimation in the subspace-learning step. Then, affinity propagation is adopted for clustering the reduced data providing an updated clustering estimation. In [21], the authors combine global with local geometric structures using a regularization technique. The singularity problem is tackled by imposing penalty on parameters and the optimal parameter is chosen based on a model selection approach.

Recently, Sparse Representations have attracted great interest in terms of data discrimination. In this context, regularised approaches for solving the generalised eigenvalue problem accompanying the Fisher criterion have been proposed providing sparse discriminant directions. For instance, in [22] a regularised method that can be applied to both matrix-based and tensor-based discriminant techniques has proven its potential. Moreover, the L_1-norm has also been successfully utilised in Fisher discriminant analysis [23]. In a similar vein, in [24] two sparse discriminant analysis methods based on $L_{2,1}$-norm penalty have been used in face recognition with interesting results. Finally, sparse graph-based discriminant analysis for preserving the sparse connection in

a manifold has also been introduced [25].

For conducting nonlinear DR, the application of the kernel trick to the linear approaches has been proposed [26]. The main idea is to firstly map the data from the initial space to a high-dimensional Hilbert space, where they might be linearly separable and then use a linear subspace method. This approach results to the kernelized versions of the linear techniques, that have already been developed, i.e., Kernel Principal Component Analysis (KPCA) [27], Kernel Clustering Discriminant Analysis (KCDA) [28], Kernel Subclass Discriminant Analysis (KSDA) [29], etc.

From the above review, it looks as though the several limitations stemming from the data distributions or the singularity of the involved matrices have been successfully addressed by dedicated methods. However, there is still enough space for improvement as the new methods introduce new limitations. For instance, subclass-based methods postulate that the data subclasses have Gaussian distributions, hence translating the problem from classes to subclasses. Moreover, although some of the above-mentioned techniques manage to deal with such limitations and optimally model the distributions of the training data, the generalization ability to the test data still remains an open challenge. To this end, as we will see in the following sections, our method achieves surpassing any distribution related limitations, while at the same moment offers great generalization chances.

III. Graph Embedding

In the GE framework [7], the set of the data samples to be projected in a low dimensional 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. Moreover, $\mathbf{W}_{int}$ and $\mathbf{W}_{pen}$ is the intrinsic and the penalty weight matrix, respectively. The intrinsic weight matrix models the similarity connections between every pair of data samples that have to be reinforced after the projection. The penalty weight matrix contains the connections between the data samples that must be suppressed after the projection. For both of the above matrices these connections can have negative values. A negative value causes the opposite results, i.e., a negative value in the intrinsic matrix means that the corresponding data samples should diverge and a negative value in the penalty matrix means that the corresponding data samples should converge after the projection.

Now, the problem of DR could be interpreted in an alternative way. 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:

$$\underset{tr\{\mathbf{YBY}^T\}=d}{\operatorname{argmin}} J(\mathbf{Y}), \tag{1}$$

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

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). The structure of the objective function (2) 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 (2) becomes:

$$J(\mathbf{Y}) = tr\{\mathbf{Y}\mathbf{L}_{int}\mathbf{Y}^T\}, \tag{3}$$

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 $\mathbf{D}_{int}(q,q) = \sum_p \mathbf{W}_{int}(q,p)$.

Similarly, 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, the above optimization problem becomes:

$$\operatorname{argmin} \frac{tr\{\mathbf{Y}\mathbf{L}_{int}\mathbf{Y}^T\}}{tr\{\mathbf{Y}\mathbf{L}_{pen}\mathbf{Y}^T\}}. \tag{4}$$

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

$$\mathbf{L}_{int}\mathbf{v} = \lambda \mathbf{L}_{pen}\mathbf{v}, \tag{5}$$

keeping the eigenvectors, which correspond to the smallest eigenvalues.

This approach leads to the optimal projection of the training data samples. For projecting new test samples, the linearization of the above approach could be used [7]. If we employ $\mathbf{y} = \mathbf{V}^T\mathbf{x}$, the objective function (2) becomes:

$$\underset{tr\{\mathbf{V}^T\mathbf{X}\mathbf{L}_{pen}\mathbf{X}^T\mathbf{V}\}=d}{\operatorname{argmin}} J(\mathbf{V}), \tag{6}$$

$$J(\mathbf{V}) = \frac{1}{2} tr\{\mathbf{V}^T \Bigg(\sum_q \sum_p (\mathbf{x}_q - \mathbf{x}_p) \mathbf{W}_{int}(q,p)(\mathbf{x}_q - \mathbf{x}_p)^T \Bigg) \mathbf{V}\}\,, \quad (7)$$

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

$$J(\mathbf{V}) = tr\{\mathbf{V}^T \mathbf{X} \mathbf{L}_{int} \mathbf{X}^T \mathbf{V}\}\,. \quad (8)$$

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

$$\mathbf{X}\mathbf{L}_{int}\mathbf{X}^T\mathbf{v} = \lambda \mathbf{X}\mathbf{L}_{pen}\mathbf{X}^T\mathbf{v}\,. \quad (9)$$

IV. Subclass Marginal Fisher Analysis

In this section, motivated by the well-known Marginal Fisher Analysis (MFA) method presented in [7], we propose a novel algorithm for dimensionality reduction, called Subclass Marginal Fisher Analysis (SMFA) employing the GE framework. The new method combines the power of subclass methods with the agility of the typical MFA to overcome the limitation of the intraclass Gaussian distribution assumption. The intrinsic graph matrix characterizes the intra-subclass compactness, while the penalty graph matrix characterizes the inter-class separability. Both graph matrices are built using neighbouring information of the graph nodes. More specifically, based on the graph embedding formulation presented in Section III, the intrinsic graph matrix is defined as:

$$\mathbf{W}_{int}(p,q) = \begin{cases} 1, \text{ if } p \in \mathcal{N}_{k_{int}}(q) \text{ or } q \in \mathcal{N}_{k_{int}}(p) \\ 0, \text{ otherwise} \end{cases}, \quad (10)$$

where $\mathcal{N}_{k_{int}}(q)$ denotes the index set of the k_{int} nearest neighbours of the q-th sample in the same subclass. The penalty graph matrix is defined as:

$$\mathbf{W}_{pen}(p,q) = \begin{cases} 1, \text{ if } p \in \mathcal{M}_{k_{pen}}(q) \text{ or } q \in \mathcal{M}_{k_{pen}}(p) \\ 0, \text{ otherwise} \end{cases}, \quad (11)$$

where $\mathcal{M}_{k_{pen}}(q)$ denotes the set of samples that belong to the k_{pen} nearest neighbours of q outside the class of q. It is worth noting that in contrast to the intrinsic graph matrix, the values of the penalty graph matrix depend on the class information regardless of the subclass labels. In this way we avoid to put constraints between subclasses belonging to the same class offering better generalization chances.

The proposed SMFA algorithm inherits all the advantages of the typical MFA method. More specifically, there is no assumption on the data distribution, since the intra-subclass compactness is encoded by the nearest neighbours of the data belonging to the same subclass and the inter-class separability is modelled using the margins among the classes. Moreover, the functionality of SMFA is based on two parameters, i.e., k_{int} and k_{pen}, which appropriately adjusted may lead to avoiding potential overfitting, therefore offering huge generalization power to the method. Also, the available projection dimensionality using SMFA is determined by k_{pen}, which almost always is much larger than that of LDA, CDA and SDA. Finally, SMFA is capable of leveraging potential subclass structure of the data, which in many cases may boost its performance. In Section V, the superiority of SMFA over a number of previously presented state-of-the-art DR methods in terms of classification accuracy is demonstrated through a series of experiments.

A. Kernel Subclass Marginal Fisher Analysis

In this section, the kernelization of SMFA (KSMFA) is presented. Kernels are widely used in classification problems, where the data are not linearly separable and in unsupervised learning when the data lie on a nonlinear manifold. 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:

$$\mathcal{X} \ni \mathbf{x}_q \rightarrow \mathbf{y}_q = f(\mathbf{x}_q) = \begin{pmatrix} \sum_{p=1}^n a_{1p} k(\mathbf{x}_q, \mathbf{x}_p) \\ \vdots \\ \sum_{p=1}^n a_{m_{\mathcal{F}}p} k(\mathbf{x}_q, \mathbf{x}_p) \end{pmatrix} \in \mathcal{F}\,, \quad (12)$$

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

$$\mathbf{Y} = \mathbf{A}^T\mathbf{K}\,, \quad (13)$$

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

$$\mathbf{A} = [\mathbf{a}_1 \cdots \mathbf{a}_{m_{\mathcal{F}}}] = \begin{pmatrix} a_{11} & \cdots & a_{m_{\mathcal{F}}1} \\ \vdots & \ddots & \vdots \\ a_{1n} & \cdots & a_{m_{\mathcal{F}}n} \end{pmatrix} \quad (14)$$

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

$$\operatorname{argmin} \frac{tr\{\mathbf{A}^T\mathbf{K}\mathbf{L}_{int}\mathbf{K}\mathbf{A}\}}{tr\{\mathbf{A}^T\mathbf{K}\mathbf{L}_{pen}\mathbf{K}\mathbf{A}\}}\,, \quad (15)$$

where $\mathbf{L}_{int} = \mathbf{D}_{int} - \mathbf{W}_{int}$ and $\mathbf{L}_{pen} = \mathbf{D}_{pen} - \mathbf{W}_{pen}$ and $\mathbf{W}_{int}$, $\mathbf{W}_{pen}$ are those defined in eq. 10 and 11, respectively. Similarly to the linear case, in order to find the optimal projections, we resolve the generalized eigenproblem:

$$\mathbf{KL}_{int}\mathbf{Ka} = \lambda \mathbf{KL}_{pen}\mathbf{Ka}\,, \tag{16}$$

keeping the eigenvectors that correspond to the smallest eigenvalues.

B. Subclass Extraction

From the above discussion, the need for efficient data clustering, is evident. A variety of clustering methods has been proposed in the literature. Techniques such as K-means and Expectation-Maximization (EM) [30] have been used for extracting clusters in a database. It is well-known that there is no method that consistently outperforms the others.

A relatively new technique relying on spectral graph theory [31], called Spectral Clustering (SC), has also been proposed for data clustering. It has been shown that SC often outperforms traditional clustering algorithms such as K-Means [32]. However, the use of this method has certain limitations, described in [33]. SC can be used for the estimation of the correct number of subclasses within each class [32]. Another potential advantage of SC is that it uses the Gram matrix, which is also used by KSMFA. Therefore, when combining SC with KSMFA, the Gram matrix has to be calculated once, hence reducing the computational load. In this paper, a multiscale Spectral Clustering (MSC) approach, proposed in [34] has been used, in order to extract clusters within each class of the data at different scales.

V. Experimental Results

We conducted classification experiments on several real-world datasets using LPP, PCA, LDA, MFA, CDA, SDA and SMFA along with their kernel counterparts. For validating the performance of the algorithms, the *5-fold cross-validation* procedure has been used. For extracting automatically the subclass structure, we have utilized the MSC technique [34], keeping the most plausible clustering for each dataset. For classifying the data, the Nearest Centroid (NC) classifier has been used with LPP, PCA LDA and MFA algorithms, while the Nearest Cluster Centroid (NCC) [35] has been used with CDA, SDA and SMFA 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.

In the following paragraphs, we briefly present the datasets that have been used along with the performance rates of the various subspace learning methods.

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.

In our experiments, for performing DR we have used both the linear and the RBF kernel approach. The maximal dimensionality of the reduced space is determined by the rank of the corresponding matrices utilized by the discriminant analysis methods. Moreover, LPP is a parametric method regarding the variance of *Gaussian similarity function*, when constructing the affinity matrix. Thus, looking for the optimal variance, in order to achieve the best classification results, makes the comparison very complex. In this paper, for the sake of simplicity and relying on some empirical studies of ours, this parameter was allowed to take values in the range $[0.1 \cdot \hat{E}(d_{ij}), 2.0 \cdot \hat{E}(d_{ij})]$, with step $0.1 \cdot \hat{E}(d_{ij})$, where $\hat{E}$ denotes the sample mean and d_{ij} is the Euclidean distance between i, j samples.

The cross-validation classification accuracy rates for the several subspace learning methods over the utilized datasets, are summarized in Tables I and II for the linear and the kernel methods, respectively. 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.

For ranking the methods in terms of classification performance we further conducted a post-hoc Bonferroni test [36] for each pair of methods. The performance of pairwise methods is significantly different, if the corresponding average ranks differ by at least the critical difference $CD = q_\alpha \sqrt{\frac{j(j+1)}{6T}}$ [37], where j is the number of methods compared,

TABLE I: Classification Accuracies (%) of Linear Methods on Several Real-World Datasets

DATASET	LPP	PCA	LDA	MFA	CDA	SDA	SMFA
FER-AIIA	40.9(3)	31.0(120)	64.6(6)	72.6(10)	73.2	**75.5(11)**	72.6(12)
BU	39.4(298)	38.1(49)	51.6(6)	**52.4(6)**	49.1(16)	52.3(15)	49.3(11)
JAFFE	46.8(18)	37.6(39)	53.2(6)	**61.5(14)**	40.0(15)	54.1(6)	44.9(20)
KANADE	34.2(92)	43.3(46)	**67.1(6)**	66.3(19)	59.7(7)	**67.1(5)**	63.8(9)
MNIST	71.1(259)	79.9(135)	84.6(9)	82.8(38)	84.8(15)	85.1(14)	**85.3(40)**
SEMEION	53.6(99)	83.2(55)	88.2(9)	86.9(8)	89.2(19)	**89.4(19)**	87.5(10)
XM2VTS	95.7(54)	92.0(86)	70.5(1)	97.7(4)	98.1(3)	97.4(2)	**98.4(4)**
IONOSPHERE	**84.6(23)**	72.3(15)	78.9(1)	76.0(12)	80.6(2)	83.4(2)	84.3(26)
MONK 1	66.7(3)	68.3(5)	50.8(1)	71.7(2)	70.0(4)	74.2(3)	**78.3(2)**
MONK 2	56.0(1)	53.3(4)	52.0(1)	58.7(2)	54.2(1)	54.0(2)	**60.7(1)**
MONK 3	77.2(5)	80.9(4)	49.4(1)	81.6(1)	74.6(2)	66.3(2)	**86.1(5)**
PIMA	61.8(1)	63.5(6)	56.5(1)	74.4(1)	60.5(3)	73.5(3)	**74.9(1)**
SPECIFIC RANK	5.1	5.8	5.0	3.0	4.0	2.7	2.3
OVERALL RANK	9.0	9.8	8.5	5.0	6.6	5.0	4.0

TABLE II: Classification Accuracies (%) of Kernel Methods on Several Real-World Datasets

DATASET	KLPP	KPCA	KDA	KMFA	KCDA	KSDA	KSMFA
FER-AIIA	50.2(252)	41.5(29)	54.9(6)	**61.3(9)**	56.1(12)	53.5(12)	56.7(39)
BU	**52.7(317)**	35.9(290)	46.6(6)	44.4(29)	41.0(13)	48.0(14)	39.9(18)
JAFFE	28.8(98)	25.9(58)	42.4(6)	**47.8(6)**	36.1(18)	46.3(5)	34.1(13)
KANADE	32.7(99)	33.2(88)	44.3(6)	**46.6(6)**	40.0(6)	38.5(6)	45.8(7)
MNIST	81.4(299)	64.5(155)	86.0(9)	86.4(21)	83.4(19)	85.2(15)	**86.7(34)**
SEMEION	83.8(99)	77.4(77)	95.3(9)	90.0(11)	94.1(19)	**95.9(19)**	94.9(20)
XM2VTS	71.3(297)	74.7(56)	61.3(1)	78.7(31)	71.5(3)	57.3(4)	**81.2(4)**
IONOSPHERE	83.7(23)	70.3(2)	92.9(1)	92.3(1)	**93.1(1)**	92.9(1)	92.6(1)
MONK 1	63.3(2)	**72.5(1)**	55.8(1)	60.0(1)	58.3(4)	61.7(3)	70.8(4)
MONK 2	54.8(1)	59.8(3)	69.7(1)	70.8(2)	78.7(1)	54.5(1)	**79.7(2)**
MONK 3	62.5(2)	**79.2(5)**	51.7(1)	**79.2(2)**	67.5(2)	58.3(1)	73.3(2)
PIMA	50.7(3)	**67.5(4)**	48.9(1)	54.0(3)	52.5(3)	52.9(1)	56.2(3)
SPECIFIC RANK	5.3	5.0	4.3	2.8	3.9	4.1	2.6
OVERALL RANK	10.2	10.0	8.1	6.3	8.2	8.3	6.1

T is the number of data sets and critical values q_α can be found in [38]. In our comparisons we set $\alpha = 0.05$. The ranking has been performed including both linear and kernel methods in the comparison, as well as separately for the linear and kernel methods. The classification performance rank of each method is referred to in the last two rows of Tables I and 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 results are also illustrated in Fig. 1 left and right, for the linear and kernel methods, respectively. The vertical axis in both figures depicts the various methods, while the horizontal axis depicts the performance ranking. The circles indicate the mean rank and the intervals around them indicate the confidence interval as this is determined by the CD value. Overlapping intervals between two methods indicate that there is not a statistically significant difference between the corresponding ranks.

The first remark from Tables I, II and Fig. 1 is that SMFA and KSMFA outperform the rest methods in the linear and kernel case, respectively. Although their superiority is not statistically significant over all remaining methods, undoubtedly these two methods offer a strong potential to improve the performance of the state-of-the-art in many classification domains. In addition, it is interesting to observe the robustness of SMFA and MFA along with their kernel counterparts across the datasets. This observation combined with the fact that both these methods rely on the same motivations shows the advantage gained by encoding the data distributions using neighbouring information between the samples towards overcoming the several limitations previously presented in this paper, offering at the same time great generalization chances.

As a general remark, the superiority of subclass methods against unimodal ones is evident, with MFA and KMFA being vivid exceptions. The top overall performance is shown by SMFA followed by SDA and MFA, while the worst performance is shown by KLPP. More specifically, on the one hand, SDA, MFA and KMFA display on average the best performance in facial expression recognition problems. On the other hand, in optical digit recognition, face

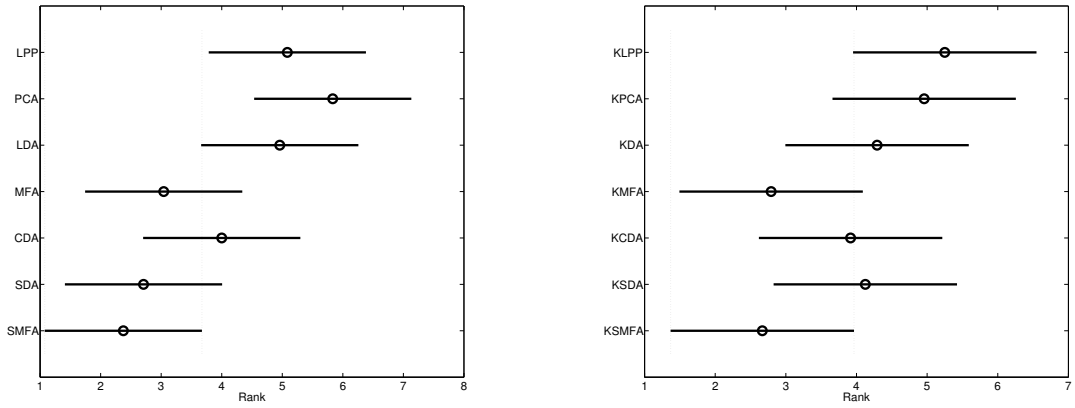

Fig. 1: Ranking of Various Methods After Pairwise Post-Hoc Bonferroni Tests on Real Data. (Left: Linear Methods, Right: Kernel Methods)

frontal view recognition and the remaining classification problems, SMFA and KSMFA clearly have on average the optimal performance.

In comparing linear with kernel methods, a simple calculation yields mean overall rank equal to 6.84 for the linear methods and 8.17 for the kernel ones. Although the difference between the two approaches (i.e., linear and kernel) is significant, 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. For instance, from Tables I and II, the use of kernels boosts the performance of PCA in three out of the four last datasets (i.e., MONK 1, MONK 3 and PIMA), while this is not the case for example in XM2VTS. There are two main reasons for this. Firstly, while some datasets contain linearly separable classes, others may require some kernel to obtain this linearity. The second reason is that in our experiments, for relaxing the computational complexity, we have used the same kernel values per dataset across all methods and there is no fact advocating that the same value constitutes the optimal parameter for each method.

VI. Conclusions

The main contribution of this paper is a novel Subclass Marginal Fisher Analysis (SMFA) dimensionality reduction method. The functionality of SMFA is based on adjacency information of data samples within the same subclass as well as the proximity of "marginal" samples belonging to different classes. In this way, the new method combines the flexibility of neighbourhood modelling methods, like MFA, with the modularity offered by subclass information towards overcoming inherent limitations stemming from the data distributions, offering at the same moment great generalization chances.

Through an extensive experimental study, it has been shown that SMFA outperforms a number of state-of-the-art subspace learning methods in many real-world datasets pertaining to various classification domains. Similar remarks could also be drawn for KSMFA. Moreover, as a general remark, it could be stated that subclass-based methods exhibit superior performance against unimodal ones, in terms of classification accuracy, proving the potential of including subclass information in the dimensionality reduction process.

Although the performance of the proposed method is impressive, there is yet space for exploring new methods employing the Graph Embedding framework, either by designing completely new methods or by modifying SMFA. Experimenting on this direction is encompassed in our future plans. Moreover, in order to reinforce even more the outcomes of this paper and to provide more credibility to SMFA, in the near future we intend to extend our current experimental study to more datasets from additional classification domains.

References

[1] X. He and P. Niyogi, "Locality preserving projections," in *NIPS*, S. Thrun, L. K. Saul, and B. Schölkopf, Eds. MIT

Press, 2003.

[2] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang, "Face recognition using laplacianfaces," *IEEE PAMI*, vol. 27, no. 3, pp. 328–340, 2005.

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

[4] 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.

[5] 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.

[6] M. L. Zhu and A. M. Martinez, "Subclass discriminant analysis," *IEEE PAMI*, vol. 28, no. 8, pp. 1274–1286, Aug. 2006.

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

[8] 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.

[9] J. Ye, R. Janardan, C. H. Park, and H. Park, "An optimization criterion for generalized discriminant analysis on undersampled problems." *IEEE PAMI*, vol. 26, no. 8, pp. 982–994, 2004.

[10] M. Kyperountas, A. Tefas, and I. Pitas, "Weighted piecewise lda for solving the small sample size problem in face verification," *IEEE Transactions on Neural Networks*, vol. 18, no. 2, pp. 506–519, 2007.

[11] W. J. Krzanowski, P. Jonathan, W. V. McCarthy, and M. R. Thomas, "General interest section: Discriminant analysis with singular covariance matrices: Methods and applications to spectroscopic data." *Applied Statistics*, vol. 44, no. 1, pp. 101–115, 1995.

[12] M. Zhu and A. M. Martínez, "Pruning noisy bases in discriminant analysis," *IEEE Transactions on Neural Networks*, vol. 19, no. 1, pp. 148–157, 2008.

[13] O. C. Hamsici and A. M. Martinez, "Bayes optimality in linear discriminant analysis," *IEEE PAMI*, vol. 30, no. 4, pp. 647–657, Apr. 2008.

[14] T. Hastie, A. Buja, and R. Tibshirani, "Penalized discriminant analysis," *Annals of Statistics*, vol. 23, pp. 73–102, 1995.

[15] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Computation*, vol. 12, no. 10, pp. 2385–2404, 2000.

[16] M. Loog, R. P. W. Duin, and R. Haeb-Umbach, "Multiclass linear dimension reduction by weighted pairwise fisher criteria." *IEEE PAMI*, vol. 23, no. 7, pp. 762–766, 2001.

[17] G. Goudelis, S. Zafeiriou, A. Tefas, and I. Pitas, "Class-specific kernel-discriminant analysis for face verification," *IEEE Transactions on Information Forensics and Security*, vol. 2, no. 3-2, pp. 570–587, 2007.

[18] N. Gkalelis, V. Mezaris, and I. Kompatsiaris, "Mixture subclass discriminant analysis," *Signal Processing Letters, IEEE*, vol. 18, no. 5, pp. 319–322, 2011.

[19] S. K. Zhou and R. Chellappa, "Multiple-exemplar discriminant analysis for face recognition." *International Conference on Pattern Recognition (ICPR) (4)*, pp. 191–194, 2004.

[20] X. Wu, X. Chen, X. Li, L. Zhou, and J. Lai, "Adaptive subspace learning: an iterative approach for document clustering," *Neural Computing and Applications*, vol. 25, no. 2, pp. 333–342, 2014.

[21] X. Shu, Y. Gao, and H. Lu, "Efficient linear discriminant analysis with locality preserving for face recognition," *Pattern Recognition*, vol. 45, no. 5, pp. 1892–1898, 2012.

[22] X. Han and L. Clemmensen, "Regularized generalized eigen-decomposition with applications to sparse supervised feature extraction and sparse discriminant analysis," *Pattern Recognition*, 2015.

[23] H. Wang, X. Lu, Z. Hu, and W. Zheng, "Fisher discriminant analysis with l1-norm," *Cybernetics, IEEE Transactions on*, vol. 44, no. 6, pp. 828–842, 2014.

[24] X. Shi, Y. Yang, Z. Guo, and Z. Lai, "Face recognition by sparse discriminant analysis via joint l 2, 1-norm minimization," *Pattern Recognition*, vol. 47, no. 7, pp. 2447–2453, 2014.

[25] N. H. Ly, Q. Du, and J. E. Fowler, "Sparse graph-based discriminant analysis for hyperspectral imagery," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 52, no. 7, pp. 3872–3884, 2014.

[26] K.-R. Müller, S. Mika, G. Rätsch, S. Tsuda, and B. Schölkopf, "An introduction to kernel-based learning algorithms." *IEEE Transactions on Neural Networks*, vol. 12, no. 2, pp. 181–202, 2001.

[27] B. Schölkopf, A. J. Smola, and K.-R. Muller, "Kernel principal component analysis." in *Proceedings of the International Conference on Artificial Neural Networks (ICANN-1997)*, 1997, pp. 583–588.

[28] B. Ma, H. Y. Qu, and H. S. Wong, "Kernel clustering-based discriminant analysis," *Pattern Recognition*, vol. 40, no. 1, pp. 324–327, Jan. 2007.

[29] D. You, O. C. Hamsici, and A. M. Martínez, "Kernel optimization in discriminant analysis," *IEEE PAMI*, vol. 33, no. 3, pp. 631–638, 2011.

[30] G. J. McLachlan and T. Krishnan, *The EM algorithm and extensions.*, 2nd ed., ser. Wiley series in probability and statistics. Hoboken, NJ: Wiley, 2008.

[31] Doob, "Spectral graph theory." in *Handbook of Graph Theory, CRC Press, 2004*, J. L. Gross and J. Yellen, Eds., 2004.

[32] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[33] U. von Luxburg, O. Bousquet, and M. Belkin, "Limits of spectral clustering." in *Advances in Neural Information Processing Systems (NIPS)*, vol. 17. MIT Press, 2005, pp. 857–864.

[34] 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.

[35] 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.

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

[37] H. Chen, P. Tino, and X. Yao, "Probabilistic classification vector machines." *IEEE Transactions on Neural Networks*, vol. 20, no. 6, pp. 901–914, 2009.

[38] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.