Maximum Margin Binary Classifiers using Intrinsic and Penalty Graphs

Berkay Kicanaoglu, Alexandros Iosifidis and Moncef Gabbouj
Department of Signal Processing, Tampere University of Technology, Tampere, Finland
Email: {alexandros.iosifidis,moncef.gabbouj}@tut.fi

***Abstract*—In this paper a variant of the binary Support Vector Machine classifier that exploits intrinsic and penalty graphs in its optimization problem is proposed. We show that the proposed approach is equivalent to a two-step process where the data is firstly mapped to an optimal discriminant space of the input space and, subsequently, the original SVM classifier is applied. Our approach exploits the underlying data distribution in a discriminant space in order to enhance SVMs generalization ability. We also extend this idea to the Least Squares SVM classifier, where the adoption of the intrinsic and penalty graphs acts as a regularizer incorporating discriminant information in the overall solution. Experiments on standard and recently introduced datasets verify our analysis since, in the cases where the classes forming the problem are not well discriminated in the original feature space, the exploitation of both intrinsic and penalty graphs enhances performance.**

I. Introduction

Support Vector Machines (SVM) [1], [2] have been found to be one of the most popular classification methods since its invention. Thanks to their solid theoretical foundation and flexibility, they have been successfully applied in numerous pattern recognition problems, including computer vision tasks, such as isolated handwritten digit, object and activity recognition [3], [4]. The fundamental idea forming the basis of SVMs is the determination of the separating hyperplane that allows maximum margin-based discrimination between classes. Hence, they are often referred to as *maximum-margin classifiers*. One of their most significant properties originates from the fact that SVMs use structural risk minimization (SRM) contrary to empirical/actual risk minimization [5]. This property ensures that the solution is unique under certain conditions.

Given the classification task, one can formulate a quadratic convex optimization problem which can be optimally solved. Although standard SVM is formulated for linear classification tasks, it can be easily transformed to solve nonlinear ones as well. This is achieved by using the "kernel trick" that is used to project the samples from the original space to a feature space of higher (even infinite) dimensions (usually called kernel space), where the classes are likely to become linearly separable and, thus, the classification problem can be solved by applying linear SVM in that space [1].

It is evident that the success of SVM-based classification is closely related to the feature space in which the method is applied. This is why a line of work in SVM-based classification attempts to formulate suitable optimization problems that combine maximum margin-based discrimination with geometric properties of the original feature space [6], [7], [8], [9], [10] or pairwise relationships between the training data [11]. All these methods exploit either global data geometric information (described by using the total scatter matrix) or class geometric information (described by using either the within-class scatter matrix or intrinsic class graph structures). Analyzing the optimization problems proposed till today, it can be seen that they can be interpreted as a two-step process: a) the input data are mapped to a new feature space that is determined by using intrinsic (class) geometric information. This process can be seen as a "whitening" or dimension scaling process that is used to map the data from the input space to a feature space where all classes have similar covariance structures. b) Application of the original SVM optimization problem on the projected data.

Based on this observation, in this paper we are trying to answer the question: "Would the exploitation of discriminant information in the first step of the above-described process be beneficial in terms of performance?". In order to answer this question, we formulate a new optimization problem for maximum margin-based classification that exploits both intrinsic and penalty graphs. We show that by doing so, indeed, the optimization problem to be solved can be interpreted as a two-step process where one determines an optimal discriminant (sub-)space, in terms of Graph Embedding-based discriminant subspace learning [12], and subsequently applies standard SVM-based classification in the transformed space. In addition, we extend this idea in LS-SVM-based classification, where we show that the adopted discriminant term plays the role of a regularizer. This regularizer expresses discrimination criteria, which can vary depending on the intrinsic and penalty graphs exploited. Quantitative and qualitative comparisons between the proposed approach and existing ones are provided and discussed.

The remainder of the paper is structured as follows. Related methods proposed till today for the exploitation of intrinsic graph structures in the SVM optimization problem are described in Section II. The proposed method incorporating discrimination criteria described by exploiting both intrinsic and penalty graphs in the SVM optimization problem is described in Section III. This idea is further extended by using a Least Squares SVM formulation in subsection III-B. Experiments comparing the performance of the proposed methods with related ones are provided in Section IV and conclusions are

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

drawn in Section V.

II. Related Work

Let us denote by $\mathbf{x}_i \in \mathbb{R}^D$, $i = 1, \ldots, N$ the set of training vectors that we would like to employ in order to train a maximum margin classifier. Let us also define the binary labels $y_i \in \{-1, 1\}$ determining whether the vectors $\mathbf{x}_i$ belong to the positive or negative class of the binary classification problem at hand. In SVM, the optimal separating hyperplane is the one that separates the two classes with maximum margin. The SVM optimization problem is defined as:

$$\underset{\mathbf{w},b}{min} \ \frac{1}{2}\mathbf{w}^T\mathbf{S}\mathbf{w} + c\sum_{i=1}^{N}\xi_i, \tag{1}$$

subject to the constraints:

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \qquad i = 1, \ldots, N, \tag{2}$$

where $\mathbf{w} \in \mathbb{R}^D$ is the vector defining the separating hyperplane, b determines the offset of the hyperplane from the origin, ξ_i, $i = 1, \ldots, N$ are the so-called slack variables and $c > 0$ is a regularization parameter denoting the importance of the training error in the optimization problem. The solution of the above-described optimization problem is a quadratic convex optimization problem of the form:

$$\underset{\boldsymbol{\alpha}}{max} \ \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{S}^{-1}\mathbf{x}_j + \sum_{i=1}^{N}\alpha_i, \tag{3}$$

subject to the constraint $0 \leq \alpha_i \leq c, i = 1, \ldots, N$. $\boldsymbol{\alpha} \in \mathbb{R}^N$ is a vector containing the Lagrange multipliers α_i, $i = 1, \ldots, N$.

In (1), $\mathbf{S} \in \mathbb{R}^{N \times N}$ is a matrix that defines properties of the feature space in which classification is applied. We can employ $\mathbf{S}$ in order to define a “whitening” or dimension scaling process that is used to scale the input space $\mathbb{R}^D$ by applying $\tilde{\mathbf{x}}_i = \mathbf{S}^{-\frac{1}{2}}\mathbf{x}_i$. One can also find a mapping of the data to a lower-dimensional feature space by finding the eigenvalue decomposition of $\mathbf{S} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^T$ and applying $\tilde{\mathbf{x}}_i = \tilde{\boldsymbol{\Sigma}}^{-\frac{1}{2}}\tilde{\mathbf{U}}^T\mathbf{x}_i$, where $\tilde{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}(1:d, 1:d)$ is a diagonal matrix containing $d < D$ eigenvalues and $\tilde{\mathbf{U}} = \mathbf{U}(:, 1:d)$ is a matrix formed by the corresponding eigenvectors.

Depending on the choice of the matrix $\mathbf{S}$, the following SVM variants have been proposed:

- Identity matrix: In this case, the original SVM classifier is applied in the original feature space $\mathbb{R}^D$ [2].
- Within-class scatter matrix: In this case, the Minimum Class Variance SVM (MCVSVM) classifier proposed in [6] is applied.
- Subclass within-class scatter matrix: In this case the Minimum Subclass Variance SVM (MSVSVM) classifier proposed in [9] is applied.
- Scatter matrix defined on an intrinsic graph structure by $\mathbf{S} = \mathbf{X}\mathbf{L}\mathbf{X}^T$, where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ and $\mathbf{L}$ is the graph Laplacian matrix defined so as to describe properties of the data that are subject to minimization: In this case the Graph Embedded SVM (GESVM) classifier proposed in [11] is applied. We should note here that the within class and the within subclass scatter matrices used in [6] and [9] can also been expressed by using specific types of Laplacian matrices and, thus, the methods in [6] and [9] can be considered to be special cases of GESVM in [11].

A. Least Squares SVM

Least Squares SVM is a regression model solving the following optimization problem [19]:

$$\underset{\mathbf{w},b}{min} \ \frac{1}{2}\mathbf{w}^T\mathbf{w} + c\sum_{i=1}^{N}\xi_i^2, \tag{4}$$

subject to the constraints:

$$\mathbf{w}^T\mathbf{x}_i + b = y_i - \xi_i, \quad i = 1, \ldots, N. \tag{5}$$

By calculating the saddle point of the optimization criterion with respect to both b and $\mathbf{w}$, we obtain:

$$b = \frac{1}{N + \mathbf{1}^T\mathbf{B}\mathbf{1}}\mathbf{y}^T(\mathbf{I} - \mathbf{B})\mathbf{1} \tag{6}$$

$$\mathbf{w} = \left(\mathbf{X}\mathbf{X}^T + \frac{1}{c}\mathbf{I}\right)^{-1}\mathbf{X}(\mathbf{y} - b\mathbf{1}), \tag{7}$$

where $\mathbf{y} = [y_1, \ldots, y_N]^T$, $\mathbf{1} \in \mathbb{R}^N$ is a vector of ones, $\mathbf{B} = \mathbf{X}^T\left(\mathbf{X}\mathbf{X}^T + \frac{1}{c}\mathbf{I}\right)^{-1}\mathbf{X}$.

B. Graph Embedding

The Graph Embedding (GE) framework [12] assumes that the training data $\mathbf{x}_i$, $i = 1, \ldots, N$ are employed in order to form the vertex set of an undirected weighted graph $\mathcal{G} = \{\mathbf{X}, \mathbf{V}\}$, where $\mathbf{V} \in \mathbb{R}^{N \times N}$ is a similarity matrix whose elements denote the relationships between the graph vertices $\mathbf{x}_i$. A penalty graph $\mathcal{G}^p = \{\mathbf{X}, \mathbf{V}^p\}$ can also be defined, whose weight matrix $\mathbf{V}^p \in \mathbb{R}^{N \times N}$ penalizes specific relationships between the graph vertices $\mathbf{x}_i$.

A linear transformation $s_i = \mathbf{w}^T\mathbf{x}_i$ is obtained by optimizing for:

$$\begin{aligned} \mathbf{w}^* &= \underset{\mathbf{w}^T\mathbf{X}\mathbf{C}\mathbf{X}^T\mathbf{w}=c}{min} \ \sum_{i,j=1}^{N}(\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j)^2 V_{ij} \\ &= \underset{\mathbf{w}^T\mathbf{X}\mathbf{C}\mathbf{X}^T\mathbf{w}=c}{argmin} \ \mathbf{w}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{w}, \end{aligned} \tag{8}$$

where $\mathbf{L} \in \mathbb{R}^{N \times N}$ is the graph Laplacian matrix defined as $\mathbf{L} = \mathbf{D} - \mathbf{V}$, $\mathbf{D}$ being the diagonal degree matrix having elements $D_{ii} = \sum_{j=1}^{N} V_{ij}$. $\mathbf{C} \in \mathbb{R}^{N \times N}$ is the graph Laplacian matrix of $\mathcal{G}^p$, that is $\mathbf{C} = \mathbf{L}^p = \mathbf{D}^p - \mathbf{V}^p$. In the case where no penalty criteria are taken into account, $\mathbf{C}$ can be set equal to a constraint matrix, e.g. a diagonal matrix for scale normalization, that is used in order to avoid trivial solutions.

The solution of (8) is obtained by solving the generalized eigenvalue decomposition problem $\mathbf{S}_i\mathbf{v} = \lambda\mathbf{S}_p\mathbf{v}$, where $\mathbf{S}_i = \mathbf{X}\mathbf{L}\mathbf{X}^T$ is a matrix expressing the data relationships that are subject to minimization and $\mathbf{S}_p = \mathbf{X}\mathbf{C}\mathbf{X}^T$ is a matrix expressing the data relationships that are subject to maximization. That is, $\mathbf{w}$ is the leading eigenvector of the matrix $\tilde{\mathbf{S}} = \mathbf{S}_p^{-1}\mathbf{S}_i$. In the case where the matrix $\mathbf{S}_p$ is singular, a regularized version is exploited, i.e. $\tilde{\mathbf{S}}_p = \mathbf{S}_p + r\mathbf{I}$, and eigenanalysis is performed to the matrix $\tilde{\mathbf{S}} = \tilde{\mathbf{S}}_p^{-1}\mathbf{S}_i$. r is a parameter that is used in order to exploit the dominant diagonal property of non-singular matrices. Within the Graph Embedding framework, several Discriminant Learning techniques can be described e.g. [13], [14], while it has also been exploited in classification schemes [15], [16], [17], [18].

III. Proposed Methods

From the above, it can be seen that the matrix $\tilde{\mathbf{S}} = \tilde{\mathbf{S}}_p^{-1}\mathbf{S}_i$ can be employed in order to describe both intrinsic and penalty relationships between the training data. We shall employ this matrix in order to embed discriminant information in maximum margin classification in Subsection III-A. In addition, we will use it in order to properly regularize the solution of LS-SVM classifier in Subsection III-B. Qualitative comparison of the proposed methods with the related ones and nonlinear extensions are provided in Subsection III-C.

A. Proposed Graph Embedded SVM

The proposed maximum margin classifier solves the following optimization problem:

$$\underset{\mathbf{w},b}{min} \; \frac{1}{2}\mathbf{w}^T\tilde{\mathbf{S}}\mathbf{w} + c\sum_{i=1}^{N}\xi_i, \tag{9}$$

subject to the constraints:

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \qquad i = 1,\dots,N. \tag{10}$$

The Lagrangian of (9) with the constraints in (10) is:

$$\begin{aligned}\mathcal{L} &= \frac{1}{2}\mathbf{w}^T\tilde{\mathbf{S}}\mathbf{w} + c\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\beta_i\xi_i \\ &- \sum_{i=1}^{N}\alpha_i[y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \xi_i]. \end{aligned}\tag{11}$$

By determining the saddle points of $\mathcal{L}$ with respect to $\mathbf{w}$, b and ξ_i, we obtain:

$$\frac{\theta\mathcal{L}}{\theta\mathbf{w}} = 0 \;\Rightarrow\; \tilde{\mathbf{S}}\mathbf{w} = \sum_{i=1}^{N}\alpha_i y_i\mathbf{x}_i, \tag{12}$$

$$\frac{\theta\mathcal{L}}{\theta b} = 0 \;\Rightarrow\; \sum_{i=1}^{N}\alpha_i y_i = 0, \tag{13}$$

$$\frac{\theta\mathcal{L}}{\theta\xi_i} = 0 \;\Rightarrow\; c - \alpha_i - \beta_i = 0. \tag{14}$$

By substituting (12), (13) and (14) in (11), the solution is obtained by solving the following quadratic convex optimization problem:

$$\underset{\boldsymbol{\alpha}}{max} \; \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\tilde{\mathbf{S}}^{-1}\mathbf{x}_j + \sum_{i=1}^{N}\alpha_i, \tag{15}$$

subject to the constraint $0 \leq \alpha_i \leq c$, $i = 1,\dots,N$.

By observing (15), it can be seen that the solution of the proposed classifier can be obtained by applying standard SVM classification on a transformed feature space. That is, it is equivalent to the application of standard SVM on $\tilde{\mathbf{x}}_i = \tilde{\mathbf{\Sigma}}^{-\frac{1}{2}}\tilde{\mathbf{U}}^T\mathbf{x}_i$, where $\tilde{\mathbf{\Sigma}}$ and $\tilde{\mathbf{U}}$ are matrices containing the $d \leq D$ smallest eigenvalues of $\tilde{\mathbf{S}}$ and the corresponding eigenvectors of $\tilde{\mathbf{S}}$. From the discussion in Subsection II-B, it can be easily seen that the first step of the proposed classification scheme is equivalent to Graph Embedding-based discriminant learning. By exploiting different intrinsic and penalty graphs, described in $\mathbf{S}_i$ and $\mathbf{S}_p$, respectively, the proposed classifier inherently exploits data (or class) relationships for maximum margin-based classification.

In order to exploit optimized SVM implementations [23], we can also exploit the equivalence of the proposed classifier with the following two-step classification scheme:

- Graph Embedding-based data projection: Discriminant data representation learning, i.e. $\tilde{\mathbf{x}}_i = \tilde{\mathbf{\Sigma}}^{-\frac{1}{2}}\tilde{\mathbf{U}}^T\mathbf{x}_i$, where $\tilde{\mathbf{\Sigma}}$ and $\tilde{\mathbf{U}}$ are matrices containing the $d \leq D$ smallest eigenvalues and the corresponding eigenvectors of $\tilde{\mathbf{S}}$.
- Standard SVM-based classification of discriminant data representations $\tilde{\mathbf{x}}_i$.

B. Proposed Graph Embedded LS-SVM

In order to incorporate discriminant information described in both intrinsic and penalty graphs in LS-SVM-based classification, we formulate the following optimization problem:

$$\underset{\mathbf{w},b}{min} \; \frac{1}{2}\mathbf{w}^T\tilde{\mathbf{S}}\mathbf{w} + c\sum_{i=1}^{N}\xi_i^2, \tag{16}$$

subject to the constraints:

$$\mathbf{w}^T\mathbf{x}_i + b = y_i - \xi_i, \quad i = 1,\dots,N. \tag{17}$$

Similarly to the SVM case described in Subsection III-A, the Lagrangian of (16) is given by:

$$\begin{aligned}\mathcal{L}_{LS-SVM} &= \frac{1}{2}\mathbf{w}^T\tilde{\mathbf{S}}\mathbf{w} + c\sum_{i=1}^{N}\xi_i^2 \\ &- \sum_{i=1}^{N}[\mathbf{w}^T\mathbf{x}_i + b - y_i + \xi_i]. \end{aligned}\tag{18}$$

The solution of the proposed LS-SVM classifier exploiting discriminant information described in both intrinsic and penalty graphs is obtained by determining the saddle points of the Lagrangian function $\mathcal{L}_{LS-SVM}$ with respect to $\mathbf{w}$ and b. By doing so, we obtain:

$$b = \frac{1}{N + \mathbf{1}^T\mathbf{B}\mathbf{1}}\mathbf{y}^T(\mathbf{I} - \mathbf{B})\mathbf{1} \tag{19}$$

$$\mathbf{w} = \left(\mathbf{X}\mathbf{X}^T + \frac{1}{c}\tilde{\mathbf{S}}\right)^{-1}\mathbf{X}(\mathbf{y} - b\mathbf{1}), \tag{20}$$

where $\mathbf{B} = \mathbf{X}^T\mathbf{A}^T\mathbf{X}$ and $\mathbf{A} = \left(\mathbf{X}\mathbf{X}^T + \frac{1}{c}\tilde{\mathbf{S}}\right)^{-1}$.

By comparing (6), (7) with (19) and (20), it can be seen that the exploitation of the intrinsic and penalty graphs in LS-SVM classification has an effect of regularization in the derived solution. This regularizer expresses both intrinsic and penalty data relationships.

TABLE II
PERFORMANCE (%) OF METHODS FOLLOWING THE SVM FORMULATION.

Dataset	KNN-Li	KNN-LiLp	LDA-Li	LDA-LiLp	CDA-Li	CDA-LiLp
Liver	57.62	**58.26**	57.27	**67.42**	57.33	**67.42**
Transfusion	76.18	**76.20**	76.18	**76.20**	76.18	**76.20**
Webcam (fc7)	86.26	**86.41**	86.41	**86.81**	86.49	**86.57**
Webcam (fc8)	82.88	**83.95**	83.57	**83.72**	83.25	**83.65**
DSLR (fc8)	88.17	**88.28**	88.06	**89.15**	88.07	**89.34**
DSLR+Webcam (SURF)	79.07	**83.40**	77.75	**81.23**	**80.53**	80.40

TABLE III
PERFORMANCE (%) OF METHODS FOLLOWING THE LS-SVM FORMULATION.

Dataset	KNN-Li	KNN-LiLp	LDA-Li	LDA-LiLp	CDA-Li	CDA-LiLp
Liver	66.49	**66.78**	**67.01**	66.03	65.85	**66.9**
Transfusion	**77.46**	77.22	77.27	77.27	77.27	**77.33**
Webcam (fc7)	80.32	**80.63**	80.47	**80.55**	**80.47**	80.32
Webcam (fc8)	80.17	**80.49**	80.17	**80.41**	80.02	**80.25**
DSLR (fc8)	82.51	**83.16**	82.83	**83.05**	82.17	**82.95**
DSLR+Webcam (SURF)	87.12	**87.26**	87.30	**87.44**	**87.31**	87.13

TABLE I
DATASETS USED IN OUR EXPERIMENTS.

Dataset	Source	#Samples	D	#Classes
Liver	UCI [20]	345	6	2
Transfusion	UCI [20]	768	8	2
Webcam (fc7)	DA [21]	259	4096 (8PCA)	11
Webcam (fc8)	DA [21]	259	1000 (7PCA)	11
DSLR (fc8)	DA [21]	186	1000 (7PCA)	10
DSLR+Webcam (SURF)	DA [21]	452	800 (45PCA)	10

C. *Discussion*

Here, we provide a qualitative comparison between the proposed method with the related methods described in Section II and we show that these methods can be considered to be special cases of the proposed one. In addition, we show that the proposed method can be easily extended to handle non-linear classification problems.

By comparing the solutions obtained for the previous methods (3) and the proposed one (15), we can see that the main difference lies on the use of a different pre-processing matrix $\mathbf{S}$ and $\tilde{\mathbf{S}}$, respectively. Having in mind that both approaches are equivalent to a two-step classification process, where the first one is a preprocessing step, it is expected that the adoption of both intrinsic and penalty graphs will lead to a more discriminant feature space for data projection and classification, while (as has been already explained in Section II) the use of only an intrinsic graph has the effect of whitening, which might not increase the discrimination ability of the obtained feature space. Moreover, the methods described in Section II can be considered as a special case of the proposed approach (i.e. by using the trivial graph structure described by the matrix $\mathbf{S}_p = \mathbf{I}$).

The proposed methods can be extended to non-linear ones by exploiting the representer theorem [1], [22], stating that the solution can be expressed as a linear combination of the training data when represented in the kernel space, i.e. $\mathbf{w} = \mathbf{\Phi}\gamma$, where $\mathbf{\Phi} = [\phi(\mathbf{x}_1), \dots, \phi(\mathbf{x}_N)]$ and $\phi(\cdot)$ is the so-called kernel function mapping the data from the input space to the kernel space.

IV. EXPERIMENTS

In this Section, we provide experiments conducted in order to compare the performance of the proposed methods with other, related, ones. We have employed 5 publicly available datasets to this end. Information regarding the datasets used is provided in Table I. On each dataset, we conducted five experiments. On each experiments we applied the 5-fold cross-validation process. On each fold, the optimal parameter values for each algorithm have been determined by applying 5-fold cross-validation on the data forming the training set and performance was measured on the test set (remaining fold). In multi-class classification problems, we followed the One-Versus-Rest approach, where multiple (equal to the number of classes) binary classifiers are learned to discriminate a class from the rest ones and the binary classification results are combined using a probabilistic approach [23].

Experimental results obtained by applying the SVM and the LS-SVM-based classification models are provided in Tables II and III, respectively. In these Tables, the graph types used for each method are provided, i.e., we have used the LDA [6], [7], CDA [9] and GE [11]. The cases where only the intrinsic graph is employed in the optimization problem are denoted by Li, while the cases where both the intrinsic and penalty graphs are employed in the optimization problem are denoted by LiLp. As can be seen, the exploitation of both intrinsic and penalty graphs has the potential of enhancing performance. While this enhancement might be small in some problems, it can be considerably big in some others. For instance, we can observe a big enhancement in performance on the Liver and

DSLR datasets in the cases where the LDA and CDA graphs are used.

V. Conclusions

In this paper, we proposed a variation of the binary Support Vector Machine (SVM) classifier that exploits intrinsic and penalty graphs in its optimization problem. We showed that the proposed method can be considered as a two-step process, where the first processing step defines an optimal data projection to a feature space of increased discrimination power, while the second one corresponds to the application of standard SVM classification. We showed that existing methods are special cases of the proposed approach. Moreover, we have employed the proposed approach for LS-SVM-based classification. When compared to existing related approaches, the proposed methods have shown to compare favourably in publicly available classification problems.

References

[1] B. Scholkpf and A.J. Smola, *Learning with Kernels*, MIT Press, 2001.

[2] V. Vapnik, *Statistical Learning Theory*, Wiley-Interscience, 2006.

[3] V. Blanz, B. Schlkopf, H. Blthoff, C. Burges, V. Vapnik and T. Vetter, *Comparison of view-based object recognition algorithms using realistic 3D models*, International Conference on Artificial Neural Networks, 1996.

[4] I. Mademlis, A. Iosifidis, A. Tefas, N. Nikolaidis and I. Pitas, *Exploiting Stereoscopic Disparity for Augmenting Human Activity Recognition Performance*, Mltimderia Tools and Applications, DOI: 10.1007/s11042-015-2719-x, 2015.

[5] C. Burges, *A Tutorial on Support Vector Machines for Pattern Recognition*, Data Mining and Knowledge Discoving, 1998.

[6] A. Tefas, C. Kotropoulos and I. Pitas, *Using Support Vector Machines to enhance the performance of Elastic Graph Matching for frontal face authentication*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 7, pp. 735-746, 2001.

[7] S. Zafeiriou, A. Tefas and I. Pitas, *Minimum Class Variance Support Vector Machines*, IEEE Transactions on Image Processing, vol. 16, no. 10, pp. 2551-2564, 2007.

[8] I. Kotsia, S. Zafeiriou and I. Pitas, *A Novel Class of Multiclass Classifiers based on the Minimization of Within-Class-Variance*, IEEE Transactions on Neural Networks, vol. 20, no. 1, pp. 14-34, 2009.

[9] G. Orfanidis and A. Tefas, *Exploiting subclass information in Support Vector Machines*, IEEE International Conference on Pattern Recognition, 2012.

[10] N. Vretos, A. Tefas and I. Pitas, *Using robust dispersion estimation in Support Vector Machines*, Pattern Recognition, vol. 46, no. 12, pp. 3441-3451, 2013.

[11] G. Arvanitidis and A. Tefas, *Exploiting Graph Embedding in Support Vector Machines*, IEEE International Workshop on Machine Learning for Signal Processing, 2012.

[12] S. Yan, D. Xu, B. Zhang and H.J. Zhang, *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.

[13] A. Iosifidis, A. Tefas and I. Pitas, *Kernel Reference Discriminant Analysis*, Pattern Recognition Letters, vol. 49, pp. 85-91, 2014.

[14] A. Iosifidis, A. Tefas and I. Pitas, *Class-specific Reference Discriminant Analysis with application in Human Behavior Analysis*, IEEE Transactions on Human-Machine Systems, vol. 45, no. 3, 315-326, 2015.

[15] A. Iosifidis, A. Tefas and I. Pitas, *Graph Embedded Extreme Learning Machine*, IEEE Transactions on Cybernetics, vol. 46, no. 1, pp. 311-324, 2016.

[16] A. Iosifidis and M. Gabbouj, *On the kernel Extreme Learning Machine speedup*, Pattern Recognition Letters, vol. 68, pp. 205-210, 2015.

[17] A. Iosifidis, A. Tefas and I. Pitas, *Regularized Extreme Learning Machine for Multi-view Semi-supervised Action Recognition*, Neurocomputing, vol. 145, pp. 250-262, 2014.

[18] A. Iosifidis, A. Tefas and I. Pitas, *Minimum Class Variance Extreme Learning Machine for Human Action Recognition*, IEEE Transactions on Circuits and Systems for Video Technology, vol. 23, no. 11, pp. 1968-1979, 2013.

[19] J.A.K. Suykens, T. Van Gestel, J. De Brabantter, B. De Moor and J. Vandewalle, *Least Squares Support Vector Machines*, World Scientific, 2002.

[20] K. Bache and M. Lichman, *UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science*, 2013.

[21] K. Saenko, B. Kulis,M. Fritz and T. Darrell *Adapting Visual Category Models to new Domaiins*, International Conference on Computer Vision, 2010.

[22] A. Argyriou, C.A. Micchelli and M. Pontil, *When is there a Representer Theorem? Vector versus Matrix regularizers*, Journal of Machine Learning Research, vol. 10, pp. 2507-2529, 2009.

[23] C.C. Chang and C.J. Lin, *LIBSVM: A library for support vector machines*, ACM Transactions on Intelligent Systems and Technology, vol. 2, no. 3, pp. 1-27, 2011.