Pattern Recognition 85 (2019) 70–77

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Exploiting multiplex data relationships in Support Vector Machines

Vasileios Mygdalis*, Anastasios Tefas, Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece

A R T I C L E   I N F O

Article history:
Received 25 August 2017
Revised 30 May 2018
Accepted 31 July 2018
Available online 1 August 2018

Keywords:
Multiplex data relationships
Support Vector Machine
Graph-based regularization
Multiple Kernel Learning

A B S T R A C T

In this paper, a novel method for introducing multiplex data relationships to the SVM optimization process is presented. Different properties about the training data are encoded in graph structures, in the form of pairwise data relationships. Then, they are incorporated to the SVM optimization problem, as modified graph-regularized basekernels, each highlighting a different property about the training data. The contribution of each graph-regularized kernel to the SVM classification problem, is estimated automatically. Thereby, the solution of the proposed modified SVM optimization problem lies in a regularized space, where data similarity is expressed by a linear combination of multiple single-graph regularized kernels. The proposed method exploits and extends the findings of Multiple Kernel Learning and graph-based SVM method families. It is shown that the available kernel options for the former can be broadened, and the exhaustive parameter tuning for the latter can be eliminated. Moreover, both method families can be considered as special cases of the proposed formulation, hereafter. Our experimental evaluation in visual data classification problems denote the superiority of the proposed method. The obtained classification performance gains can be explained by the exploitation of multiplex data relationships, during the classifier optimization process.

1. Introduction

Computer vision/visual analysis methods have found industrial applications in several areas such as in robotic systems e.g., unmanned aerial vehicles and virtual reality, and their growth over the past few years have been immense. Such visual analysis applications including face recognition, object recognition, human action recognition, human/object tracking and many other applications, are commonly addressed as classification problems [1,2]. One of the most widely studied classification methods in visual analysis applications is the Support Vector Machines (SVM) classifier. SVM-based methods and extensions have been employed in mathematical/engineering problems including one-class and multiclass classification, regression and semi-supervised learning [3–6]. In its simplest form, SVM learns from labeled data examples originating from two classes, the hyperplane that separates them with the maximum margin, at the training data input (or feature) space. After its first proposal, SVM has been extended to determine decision functions in feature spaces obtained by employing non-linear data mappings, where data similarity is implicitly expressed by a kernel function. The explicit data mapping is not required to be known, if the adopted kernel function satisfies Mercer conditions [7]. Common practices for determining a feature space where SVM provides satisfactory performance to a given classification/regression problem, involve selecting a kernel function from a set of widely adopted kernel functions e.g., polynomial, sigmoid, Radial Basis Function (RBF), and thereby tuning the corresponding hyperparameters using e.g., cross validation, based on previous knowledge about the problem at hand. In every case, the performance of SVM heavily depends on the adopted kernel function choice, since the optimal solution for each problem might lie in unknown feature spaces.

In order to determine the optimal feature space for SVM operation, Multiple Kernel Learning (MKL) methods have been proposed. Their basic assumption is that the optimal underlying data mapping, i.e., the optimal kernel function, is a weighted combination (either linear or non-linear) of Multiple Kernel functions, the so-called basekernels [8–11]. The participation of each kernel to the optimal solution is determined by a parameter vector, i.e., the basekernel weights. The weights of the basekernels are estimated in an automated fashion along with the SVM hyperplane, by following an additional optimization procedure (e.g., single-step sequential optimization, two-step optimization). Standard MKL methods employ L_p or L_1 loss functions for determining the kernel weights, with the latter producing sparse solutions and the former providing fast convergence [12,13]. Besides the important theoretical advancements of MKL methods, only few basekernel combinations have found to be successful in realistic applications, i.e.,

* Corresponding author.
E-mail address: vmigdal@aiia.csd.auth.gr (V. Mygdalis).

https://doi.org/10.1016/j.patcog.2018.07.032

MKL methods method might suffer from overfitting issues or limited performance gains [11–13].

A alternative approach for improving classification performance, are methods that introduce additional optimization options to the standard SVM optimization problem, exploiting discriminant/manifold learning criteria [6]. That is, slightly modified SVM-based optimization problems have been proposed, that lead to standard SVM solutions in regularized spaces, expressed by a geometric transformation of the derived SVM hyperplane with the adopted criteria. For example, employing discriminant learning information e.g., within-class variance information [14], promotes SVM hyperplanes that span along low data variance directions [15,16]. Alternatively, SVM-based methods have been proposed for semi-supervised learning case, by integrating SVM with manifold learning [6], by exploiting *k*NN graphs as additional regularization criteria. It has been shown that exploiting such criteria at the fully supervised learning case is also beneficial to the classification performance. Since advances in graph-theory allow several manifold/discriminant learning criteria to be expressed using generic graph-based representation [17], methods incorporating the underlying data geometry in the SVM optimization problem can be implemented through graph-based SVM methods [18–20]. The adoption of generic graph structures within the SVM optimization process, containing e.g., intrinsic (within-class), or between-class data relationships, promotes solutions that are less prone to over-fitting. The disadvantage of graph-based SVM methods is that deriving the optimal classification space requires the evaluation of different graph settings, as well as tuning the additional introduced hyperparameters.

In visual analysis applications, MKL and graph-based SVM methods have been successfully employed over the past few years. Their success can be mainly attributed to the exploitation of the multimodal/multiplex structure of images and video data [21], related to e.g., spatial and temporal information, information extracted by multiple descriptor types, or even noise generated by camera movement, multiple viewing angles and illumination changes. All this information cannot be efficiently encoded with a single kernel matrix. Our work was inspired by the successful exploitation of multiple graphs in related application scenarios, e.g., label propagation [22–26]. Therefore, we have devised a classification method that introduces multiple graphs to the SVM optimization problem, by exploiting the intuitions of both MKL and graph-based SVM method families.

In this paper, a novel classification method that incorporates multiplex data relationships to the SVM optimization process, is presented. Multiplex data relationships are encoded in the form of multiple graph structures, containing pairwise data relationships, each corresponding to a specific data property. We propose a modified SVM optimization problem, that incorporates this information. As an effect, the generated SVM hyperplane is driven to directions where the most discriminant training data properties are highlighted. From our derivations, it is shown that the solution of the proposed optimization problem lies in a modified space, where data similarity is explicitly determined by a linear combination of graph-regularized kernel matrices. Moreover, it is proven that both Multiple Kernel Learning and Graph-based SVM method families method families can be formulated as special cases of the proposed method, hereafter. Finally, the proposed method exploits and extends the findings of Multiple Kernel Learning and graph-based SVM method families, by broadening the available kernel options for the former, and eliminating exhaustive parameter tuning for the latter.

2. Related work

In this section, we overview the preliminary material required to introduce the proposed method. Section 2.1 contains the description of the generic MKL–SVM optimization problem and Section 2.2 contains an overview of the recently proposed Graph-Embedded Support Vector Machines, exploiting a single graph in its optimization problem for regularization purposes.

2.1. Multiple Kernel learning support vector machines

Let a set of labeled data $\mathcal{S} = \{\boldsymbol{x}_i, y_i\}, i = 1, \dots, N$ sampled from $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \in \mathbb{R}^D$ and $\mathcal{Y} \in \{-1, 1\}$, that is employed in order to train an SVM classifier. MKL–SVM methods optimize for implicitly determining the optimal feature space for solving the SVM optimization problem. Similarity in that space is reproduced by a linear or non-linear combination of Multiple Kernel functions [10,13,27–30]. Let M mapping functions $\phi_m(\cdot) \mapsto \mathcal{H}^m, m = 1, \dots, M$ that have been employed as base data mappings. Similarity in the respective spaces is reproduced by the associated basekernel function $\kappa_m(\cdot,\cdot) = \phi_m(\cdot)^T \phi_m(\cdot)$, and $\mathcal{H}^m$ is a Reproducing Kernel Hilbert Space (RKHS). Assuming M basekernels have been linearly combined, then the obtained space $\mathcal{H}$ is also a RKHS, reproduced by kernel $\kappa(\cdot,\cdot)$. Similarity in that space can be calculated explicitly by a weighted summation of the basekernels, as follows:

$$\kappa(\cdot,\cdot) = \sum_{m=1}^{M} \mu_m \kappa_m(\cdot,\cdot), \tag{1}$$

where κ_m is the mth kernel function weighted by a parameter $\mu_m \geq 0$.

In order to learn the kernel weighting parameters μ_m and the optimal SVM hyperplane at the same time, the MKL–SVM optimization problem is formed as a max-min optimization problem:

$$\max_{\boldsymbol{\alpha}} \min_{\boldsymbol{\mu}} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \sum_{m=1}^{M} \mu_m \kappa_m(\boldsymbol{x}_i, \boldsymbol{x}_j)$$
$$\text{s.t.} \quad 0 \leq \alpha_i \leq c \ \text{ and } \sum_{m=1}^{M} \mu_m^p = 1, \tag{2}$$

where $\boldsymbol{a}$ is the support vector coefficient vector and $p \geq 1$ is a parameter that affects the sparsity of the obtained basekernel combination. The above defined optimization problem can be solved sequentially or in an iterative manner, keeping $\boldsymbol{a}$ or $\boldsymbol{\mu}$ as constants in the respective optimization steps. Assuming that the kernel weighting parameters $\boldsymbol{\mu}$ have been determined, then $\boldsymbol{K} = \sum_{m=1}^{M} \mu_m \boldsymbol{K}_m$ is the kernel matrix that can be employed for solving the standard SVM classification problem. According to Representer Theorem [7], the relevant SVM hyperplane $\boldsymbol{w} = \boldsymbol{\Phi a}$ that lies in the RKHS $\mathcal{H}$, can be reconstructed by the determined support vector coefficient vector $\boldsymbol{a}$ and the arbitrary training data representations $\boldsymbol{\Phi} \in \mathcal{H}$. Data similarity in that space can only be reproduced by the basekernel combination, since the kernel $\boldsymbol{K}$ cannot be calculated, otherwise.

After training the classifier, a test sample $\boldsymbol{x}$ is classified to the positive or negative training class, according to the outputs of the following decision function:

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} y_i \alpha_i \sum_{m=1}^{M} \mu_m \kappa_m(\boldsymbol{x}_i, \boldsymbol{x}) + b, \tag{3}$$

where b is the standard SVM bias term. Finally, the test sample is classified to the positive class if $sign(f(\boldsymbol{x})) \geq 0$ or the negative class, otherwise.

2.2. Support Vector Machines exploiting geometric data relationships

Graph-based SVM methods exploit data relationships expressed by a single graph in the SVM optimization problem [18,20]. To this end, it is assumed that the training data $\mathcal{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$ have been embedded in an undirected weighted graph $\mathcal{G} = \{\mathcal{X}, \mathbf{W}\}$, where $\boldsymbol{W} \in \mathbb{R}^{N \times N}$ is the graph weight matrix. It should be noted that non-linear data relationships might be expressed as well, by