Pattern Recognition 78 (2018) 144–153

Contents lists available at ScienceDirect

Pattern Recognition

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

Indefinite kernel spectral learning

Siamak Mehrkanoon[a], Xiaolin Huang[b,*], Johan A.K. Suykens[a]

[a] Department of Electrical Engineering (ESAT-STADIUS), KU Leuven, Kasteelpark Arenberg 10, Leuven B-3001, Belgium
[b] Institute of Image Processing and Pattern Recognition, and the MOE Key Laboratory of System Control and Information Processing, Shanghai Jiao Tong University, Shanghai 200240, PR China

ARTICLE INFO

Article history:
Received 21 February 2017
Revised 18 October 2017
Accepted 14 January 2018

Keywords:
Semi-supervised learning
Scalable models
Indefinite kernels
Kernel spectral clustering
Low embedding dimension

ABSTRACT

The use of indefinite kernels has attracted many research interests in recent years due to their flexibility. They do not possess the usual restrictions of being positive definite as in the traditional study of kernel methods. This paper introduces the indefinite unsupervised and semi-supervised learning in the framework of least squares support vector machines (LS-SVM). The analysis is provided for both unsupervised and semi-supervised models, i.e., Kernel Spectral Clustering (KSC) and Multi-Class Semi-Supervised Kernel Spectral Clustering (MSS-KSC). In indefinite KSC models one solves an eigenvalue problem whereas indefinite MSS-KSC finds the solution by solving a linear system of equations. For the proposed indefinite models, we give the feature space interpretation, which is theoretically important, especially for the scalability using Nyström approximation. Experimental results on several real-life datasets are given to illustrate the efficiency of the proposed indefinite kernel spectral learning.

1. Introduction

Kernel-based learning models have shown great success in various application domains [1–3]. Traditionally, kernel learning is restricted to positive semi-definite (PSD) kernels as the properties of Reproducing Kernel Hilbert Spaces (RKHS) are well explored. However, many positive semi-definite kernels such as the sigmoid kernel [4] remain positive semi-definite only when their associated parameters are within a certain range, otherwise they become non-positive definite [5]. Moreover, the positive definite kernels are limited in some problems due to the need of non-Euclidean distances [6,7]. For instance in protein similarity analysis, the protein sequence similarity measures require learning with a non-PSD similarity matrix [8].

The need of using indefinite kernels in machine learning methods attracted many research interests on indefinite learning in both theory and algorithm. Theoretical discussions are mainly on Reproducing Kernel Kreĭn Spaces (RKKS, [9,10]), which is different to the RKHS for PSD kernels. In algorithm design, a lot of attempts have been made to cope with indefinite kernels by regularizing the non-positive definite kernels to make them positive semi-definite [11–14]. It is also possible to directly use an indefinite kernel in e.g., support vector machine (SVM) [4]. Though an indefinite kernel makes the problem non-convex, it is still possible to get a local optimum as suggested by Lin and Lin [15]. One important issue is that kernel trick is no longer valid when an indefinite kernel is applied in SVM and one needs new feature space interpretations to explain the effectiveness of SVM with indefinite kernels. The interpretation is usually about a pseudo-Euclidean (pE) space, which is a product of two Euclidean vector spaces, as analyzed in [10,16]. Notice that “indefinite kernels” literally covers asymmetric ones and complex ones. But this paper restricts “indefinite kernel” to the kernels that correspond to real symmetric indefinite matrices, which is consistent to the existing literature on indefinite kernel.

Indefinite kernels are also applicable to the least squares support vector machines [17]. In LS-SVM, one solves a linear system of equations in the dual and the optimization problem itself has no additional requirement on the positiveness of the kernel. In other words, even if an indefinite kernel is used in the dual formulation of LS-SVM, it is still convex and easy to solve, which is different from indefinite kernel learning with SVM. However, like in SVM, using an indefinite kernel in LS-SVM looses the traditional interpretation of the feature space and a new formulation has been recently discussed in [18].

Motivated by the success of indefinite learning for some supervised learning tasks, we in this paper introduce indefinite similarities to unsupervised as well as semi-supervised models that can learn from both labeled and unlabeled data instances. There have been already many efficient semi-supervised models, such as

* Corresponding author.
E-mail addresses: siamak.mehrkanoon@esat.kuleuven.be (S. Mehrkanoon), xiaolinhuang@sjtu.edu.cn (X. Huang), johan.suykens@esat.kuleuven.be (J.A.K. Suykens).

https://doi.org/10.1016/j.patcog.2018.01.014
0031-3203/

Laplacian support vector machine [19], which assumes that neighboring point pairs with a large weight edge are most likely within the same cluster. However, to the best of our knowledge, there is no work that extends unsupervised/semi-supervised learning to indefinite kernels.

Since using indefinite kernels in the framework of LS-SVM does not change the training problem, here we focus on multi-class semi-supervised kernel spectral clustering (MSS-KSC) model proposed by Mehrkanoon et al. [20]. MSS-KSC model and its extensions for analyzing large-scale data, data streams as well as multi-label datasets are discussed in [21–23] respectively. When one of the regularization parameters is set to zero, MSS-KSC becomes the kernel spectral clustering (KSC), which is an unsupervised learning algorithm introduced by Alzate and Suykens [24]. It is a special case of MSS-KSC. Due to the link to LS-SVM, it can be expected and also will be shown here that MSS-KSC with indefinite similarities are still easy to solve. However, the kernel trick is no longer valid and we have to find corresponding feature space interpretations. The purpose of this paper is to introduce indefinite kernels for semi-supervised learning as well as unsupervised learning as a special case. Specifically, we propose indefinite kernels in MSS-KSC and KSC models. Subsequently, we derive their feature space interpretation. Besides of theoretical interests, the interpretation allows us to develop algorithms based on Nyström approximation for large-scale problems.

The paper is organized as follows. Section 2 briefly reviews the MSS-KSC with PSD kernel. In Section 3, the MSS-KSC with an indefinite kernel is derived and the interpretation of the feature map is provided. As a special case of MSS-KSC, the KSC with an indefinite kernel and its feature interpretation is discussed in Section 4. In Section 5, we discuss the scalability of the indefinite KSC/MSS-KSC model on large-scale problems. The experimental results are given in Section 6 to confirm the validity and applicability of the proposed model on several real life small and large-scale datasets. Section 7 ends the paper with a brief conclusion.

2. MSS-KSC with PSD kernel

Consider training data

$$\mathcal{D} = \{\underbrace{x_1, \ldots, x_{n_{UL}}}_{\substack{\text{Unlabeled} \\ (\mathcal{D}_U)}}, \underbrace{x_{n_{UL}+1}, \ldots, x_n}_{\substack{\text{Labeled} \\ (\mathcal{D}_L)}}\}, \tag{1}$$

where $\{x_i\}_{i=1}^n \in \mathbb{R}^d$. The first n_{UL} points do not have labels whereas the last $n_L = n - n_{UL}$ points have been labeled. Assume that there are Q classes ($Q \leq N_c$), then the label indicator matrix $Y \in \mathbb{R}^{n_L \times Q}$ is defined as follows:

$$Y_{ij} = \begin{cases} +1 & \text{if the } i\text{th point belongs to the } j\text{th class}, \\ -1 & \text{otherwise}. \end{cases} \tag{2}$$

The primal formulation of multi-class semi-supervised KSC (MSS-KSC) described by Mehrkanoon et al. [20] is given as follows:

$$\begin{aligned} \min_{w^{(\ell)}, b^{(\ell)}, e^{(\ell)}} \quad & \frac{1}{2}\sum_{\ell=1}^{Q} w^{(\ell)^T} w^{(\ell)} - \frac{\gamma_1}{2}\sum_{\ell=1}^{Q} e^{(\ell)^T} V e^{(\ell)} + \\ & \frac{\gamma_2}{2}\sum_{\ell=1}^{Q} (e^{(\ell)} - c^{(\ell)})^T \tilde{A} (e^{(\ell)} - c^{(\ell)}) \\ \text{subject to} \quad & e^{(\ell)} = \Phi w^{(\ell)} + b^{(\ell)} 1_n, \; \ell = 1, \ldots, Q, \end{aligned} \tag{3}$$

where c^ℓ is the ℓth column of the matrix C defined as

$$C = [c^{(1)}, \ldots, c^{(Q)}]_{n \times Q} = \left[\begin{array}{c} 0_{n_{UL} \times Q} \\ \hline Y \end{array}\right]_{n \times Q}. \tag{4}$$

Here

$\Phi = [\varphi(x_1), \ldots, \varphi(x_n)]^T \in \mathbb{R}^{n \times h}$

where $\varphi(\cdot) : \mathbb{R}^d \to \mathbb{R}^h$ is the feature map and h is the dimension of the feature space which can be infinite dimensional. $0_{n_{UL} \times Q}$ is a zero matrix of size $n_{UL} \times Q$, Y is defined previously, and the right hand of (4) is a matrix consisting of $0_{n_{UL} \times Q}$ and Y. The matrix $\tilde{A}$ is defined as follows:

$$\tilde{A} = \left[\begin{array}{c|c} 0_{n_{UL} \times n_{UL}} & 0_{n_{UL} \times n_L} \\ \hline 0_{n_L \times n_{UL}} & I_{n_L \times n_L} \end{array}\right],$$

where $I_{n_L \times n_L}$ is the identity matrix of size $n_L \times n_L$. V is the inverse of the degree matrix defined as follows:

$$V = D^{-1} = \text{diag}\left(\frac{1}{d_1}, \cdots, \frac{1}{d_n}\right),$$

where $d_i = \sum_{j=1}^n K(x_i, x_j)$ is the degree of the ith data point.

As stated in [20], the object function in the formulation (3), contains three terms. The first two terms together with the set of constraints correspond to a weighted kernel PCA formulation in the least squares support vector machine framework given in [24] which is shown to be suitable for clustering and is referred to as kernel spectral clustering (KSC) algorithm. The last regularization term in (3) aims at minimizing the squared distance between the projections of the labeled data and their corresponding labels. This term enforces the projections of the labeled data points to be as close as possible to the true labels. Therefore by incorporating the labeled information, the pure clustering KSC model is guided so that it respects the provided labels by not misclassifying them. In this way, one could learn from both labeled and unlabeled instances. In addition thanks to introduced model selection scheme in [20], the MSS-KSC model is also equipped with the out-of-sample extension property to predict the labels of unseen instances.

It should be noted that, ignoring the last regularization term, or equivalently setting $\gamma_2 = 0$ and $Q = N_c - 1$, reduces the MSS-KSC formulation to kernel spectral clustering (KSC) described in [24]. Therefore, KSC formulation in the primal can be covered as a special case of MSS-KSC formulation. As illustrated by Mehrkanoon et al. [20], given Q labels the approach is not restricted to finding just Q classes and instead is able to discover up to 2^Q hidden clusters. In addition, it uses a low embedding dimension to reveal the existing number of clusters which is important when one deals with large number of clusters.

When the feature map φ in (3) is not explicitly known, in the context of PSD kernel, one may use the kernel trick and solve the problem in the dual. Elimination of the primal variables $w^{(\ell)}$, $e^{(\ell)}$ and making use of Mercer's Theorem result in the following linear system in the dual [20]:

$$\gamma_2 \left(I_n - \frac{R 1_n 1_n^T}{1_n^T R 1_n}\right) c^{(\ell)} = \alpha^{(\ell)} - R\left(I_n - \frac{1_n 1_n^T R}{1_n^T R 1_n}\right) \Omega \alpha^{(\ell)}, \tag{5}$$

where $R = \gamma_1 V - \gamma_2 \tilde{A}$. In (5), there are two coefficients, namely γ_1 and γ_2, which reflect the emphasis on unlabeled and labeled samples, respectively, as shown in (3). Besides, there could be one or multiple parameters in the kernel. All of these parameters could be tuned by cross-validation.

3. MSS-KSC with indefinite kernel

Traditionally, the kernel used in MSS-KSC is restricted to be positive semi-definite. When the kernel in (5) is indefinite, one still requires to solve a linear system of equations. However, the feature space has different interpretations compared to definite kernels. In what follows we establish and analyze the feature space interpretations for MSS-KSC.

Theorem 3.1. *Suppose that for a symmetric but indefinite kernel matrix K, the solution of the linear system (5) is denoted by $[\alpha_*, b_*]^T$.*