Indefinite Kernels in LS-SVM and kPCA

Xiaolin Huang, Andreas Maier, Joachim Hornegger, and Johan A. K. Suykens

***Abstract*—Because of several successful applications, indefinite kernels have attracted many research interests in recent years. This paper addresses indefinite learning in the framework of least squares support vector machines (LS-SVM). Unlike C-SVM with an indefinite kernel, which involves a non-convex problem, the indefinite LS-SVM is still easy to solve, but the kernel trick and primal-dual relationship for LS-SVM with a Mercer kernel is no longer valid. In this paper, we give a feature space interpretation for indefinite LS-SVM. In the same framework, kernel principal component analysis with an infinite kernel is discussed as well. In numerical experiments, LS-SVM with an indefinite kernel, called the truncated ℓ_1 distance kernel, for classification and kernel principal component analysis is evaluated. Its good performance together with the feature space interpretation given in this paper imply the potential use of indefinite LS-SVM in real applications.**

***Index Terms*—least squares support vector machine, indefinite kernel, classification, kernel principal components analysis**

I. Introduction

MERCER'S condition is the traditional requirement on the kernel applied in classical kernel learning methods, such as support vector machines (C-SVM, [1]), least squares support vector machines (LS-SVM, [2][3]), and kernel principal components analysis (kPCA, [4]). However, in practice, one may meet sophisticated similarity or dissimilarity measures which lead to kernels violating Mercer's condition. Since the kernel matrices induced by such kernels are not positive semi-definite, they are called *indefinite kernels* and the corresponding learning methodology is called *indefinite learning*.

Two important problems arise for indefinite learning. First, it lacks the classical feature space interpretation for a Mercer kernel, i.e., we cannot find a nonlinear feature mapping such that its inner-dot gives the value of an indefinite kernel function. Second, lack of positive definitiveness makes many learning algorithms become non-convex if an indefinite kernel is used. In the last decades, there has been continuous progress aiming at these issues. In theory, indefinite learning in C-SVM has been discussed in the Reproducing Kernel Kreĭn Spaces (RKKS), cf. [5]–[8]. The analysis for kPCA with an indefinite kernel can be found in [9] and [10], which is also discussed on RKKS. In algorithm design, the current mainstream is to find an approximate positive semi-definite (PSD) kernel and then apply classical kernel learning algorithm based on that PSD kernel. For example, it was proposed by [11] to set negative eigenvalues as zero to construct an approximate PSD kernel. Another method is to flip the negative eigenvalues, as discussed by [12]. The distance between the original indefinite kernel and the approximate PSD one can be measured by the Frobenius norm and pursuing the nearest PSD kernel was formulated as an optimization problem by [13].

In [14], the above methods transferring indefinite kernels to PSD ones have been reviewed and compared. That paper also discusses directly applying indefinite kernels in some classical kernel learning methods which are not sensitive to metric violations. As suggested by [15], one can use an indefinite kernel to replace the PSD kernel in the dual formulation of C-SVM and solve it by sequential minimization optimization [16] [17]. This kind of methods enjoy a similar computational efficiency as the classical learning methods and hence is more attractive in practical use.

Following a similar way of introducing indefinite kernels to C-SVM, we consider indefinite learning based on LS-SVM. Notice that using an indefinite kernel in C-SVM results in a non-convex problem but indefinite learning based on LS-SVM is still easy to solve. However, the kernel trick is no longer valid. We have to find a new feature space interpretation and to give a characterization in terms of primal and dual problems, which are the main targets of this paper. Since kPCA can be conducted in the framework of LS-SVM [18], we will discuss kPCA with an indefinite kernel as well.

This paper is organized as follows. Section II addresses LS-SVM with indefinite kernels and provides its feature space interpretation. Similar discussion on kPCA is given in Section III. Then the performance of indefinite learning based on LS-SVM is evaluated by numerical experiments in Section IV. Finally, Section V gives a brief conclusion.

Manuscript received 2015;

Xiaolin Huang is currently supported by the Humboldt research fellowship for postdoctoral researchers program. Johan Suykens acknowledges support of ERC AdG A-DATADRIVE-B (290923), KUL: GOA/10/09 MaNet, CoE PFV/10/002 (OPTEC), BIL12/11T; FWO: G.0377.12, G.088114N, SBO POM (100031); IUAP P7/19 DYSCO.

X. Huang, A. Maier, and J. Hornegger are with the Pattern Recognition Lab, the Friedrich-Alexander-Universität Erlangen-Nürnberg, 91058 Erlangen, Germany. (emails: huangxl06@mails.tsinghua.edu.cn, anreas.maier@fau.de, joachim.hornegger@fau.de). J.A.K. Suykens is with the Department of Electrical Engineering (ESAT-STADIUS), KU Leuven, B-3001 Leuven, Belgium. (e-mail: johan.suykens@esat.kuleuven.be).

II. LS-SVM with Indefinite Kernels

We start the discussion from C-SVM with a Mercer kernel. Given a set of training data $\{\mathbf{x}_i, y_i\}_{i=1}^m$ with $\mathbf{x}_i \in \mathbb{R}^n$ and $y_i \in \{-1, +1\}$, we are trying to construct a discriminant function $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ and use its sign for classification. Except of linearly separable problems, a nonlinear feature mapping $\phi(\mathbf{x})$ is needed and the discriminant function is usually formulated as $f(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b$. Then C-SVM trains $\mathbf{w}$ and b by the

following optimization problem:

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^T\phi(\mathbf{x}_i)+b) \geq 1-\xi_i, \quad \forall i \\ & \xi_i \geq 0, \quad \forall i. \end{aligned} \tag{1}$$

It is well known that the dual problem of (1) takes the formulation as below,

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y_i\alpha_i\mathcal{K}(\mathbf{x}_i,\mathbf{x}_j)\alpha_j y_j \\ \text{s.t.} \quad & \sum_{i=1}^{m} y_i\alpha_i = 0 \\ & 0 \leq \alpha_i \leq C, \quad \forall i, \end{aligned} \tag{2}$$

where $\mathcal{K}(\mathbf{x}_i,\mathbf{x}_j) = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$ is the kernel function. At the same time, for any kernel $\mathcal{K}$ which satisfies Mercer's condition, there is always a feature mapping ϕ such that $\mathcal{K}(\mathbf{x}_i,\mathbf{x}_j) = \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$. This kernel trick allows us to construct a classifier to maximize the margin in the feature space without explicitly knowing ϕ.

When an indefinite kernel is used in (2), the kernel matrix $\mathcal{K}$ with $\mathcal{K}_{ij} = \mathcal{K}(\mathbf{x}_i,\mathbf{x}_j)$ is not positive semi-definite. Therefore, (2) is no longer a convex problem, which makes many algorithms based on global optimality invalid. But some descent methods, like SMO, are still applicable, as suggested by [15]. A more crucial issue is that for an indefinite kernel, the kernel trick is not valid, which means that (1) and (2) are not a pair of primal-dual problems and then the solution of (2) cannot be explained as margin maximization in a feature space. In [19], solving (2) with an indefinite kernel has been explained as distance minimization in a pseudo-Euclidean (pE) space, which is denoted as $\mathbb{R}^{(p,q)}$ with non-negative integers p and q. This space is a product of two Euclidean vector spaces $\mathbb{R}^p \times i\mathbb{R}^q$. An element in $\mathbb{R}^{(p,q)}$ can be represented by its coordinate vector and the coordinate vector gives the inner product: $\langle \mathbf{u},\mathbf{v}\rangle_{\mathrm{pE}} = \mathbf{u}^T\mathbf{M}\mathbf{v}$, where $\mathbf{M}$ is a diagonal matrix with the first p components equal 1 and others equal to -1.

Similar to the discussion about C-SVM, indefinite kernels can also be used in the framework of least squares support vector machines. In the dual space, LS-SVM is to solve the following linear system [2]:

$$\begin{bmatrix} 0 & \mathbf{y}^T \\ \mathbf{y} & \mathcal{H} + \frac{1}{\gamma}\mathcal{I} \end{bmatrix} [b,\alpha_1,\ldots,\alpha_m]^T = \begin{bmatrix} 0 \\ \mathbf{1} \end{bmatrix}, \tag{3}$$

where $\mathcal{I}$ is an identity matrix, $\mathbf{1}$ is an all ones vector with the proper dimension, and $\mathcal{H}$ is given by

$$\mathcal{H}_{ij} = y_i y_j \mathcal{K}_{ij} = y_i y_j \mathcal{K}(\mathbf{x}_i,\mathbf{x}_j).$$

We assume that the matrix in (3) is full rank. Then its solution can be effectively obtained and the corresponding discriminant function is

$$f(\mathbf{x}) = \sum_{i=1}^{m} y_i\alpha_i\mathcal{K}(\mathbf{x},\mathbf{x}_i) + b.$$

The existing discussion about LS-SVM usually requires $\mathcal{K}$ to be positive semi-definite such that the kernel trick is applicable and the solution of (3) is related to Fisher discriminant analysis in feature space [20].

Now we propose to use an indefinite kernel in LS-SVM (3). Notice that even if $\mathcal{K}$ is indefinite, (3) is still easy to solve, which differs from C-SVM, where an indefinite kernel makes (2) non-convex. Though (3) with an indefinite kernel is easy to solve, the solution looses many properties of PSD kernels and its feature space interpretations have to be analyzed also in a pE space. This is based on the following proposition:

Proposition 1: Suppose α_i^*, b^* is the solution of (3) for a symmetric but indefinite kernel matrix $\mathcal{K}$. Then there exist two feature mappings ϕ_+ and ϕ_- such that

$$\mathbf{w}_+^* = \sum_{i=1}^{m}\alpha_i^*\phi_+(\mathbf{x}_i), \qquad \mathbf{w}_-^* = \sum_{i=1}^{m}\alpha_i^*\phi_-(\mathbf{x}_i),$$

which is a stationary point of the following primal problem:

$$\begin{aligned} \min_{\mathbf{w}_+,\mathbf{w}_-,b,\xi} \quad & \frac{1}{2}(\mathbf{w}_+^T\mathbf{w}_+ - \mathbf{w}_-^T\mathbf{w}_-) + \frac{\gamma}{2}\sum_{i=1}^{m}\xi_i^2 \\ \text{s.t.} \quad & y_i(\mathbf{w}_+^T\phi_+(\mathbf{x}_i) + \mathbf{w}_-^T\phi_-(\mathbf{x}_i) + b) = 1-\xi_i, \forall i. \end{aligned} \tag{4}$$

Meanwhile, the dual problem of (4) is given by (3), where

$$\mathcal{K}(\mathbf{x}_i,\mathbf{x}_j) = \mathcal{K}_+(\mathbf{x}_i,\mathbf{x}_j) - \mathcal{K}_-(\mathbf{x}_i,\mathbf{x}_j) \tag{5}$$

with two PSD kernels $\mathcal{K}_+$ and $\mathcal{K}_-$:

$$\mathcal{K}_+(\mathbf{x}_i,\mathbf{x}_j) = \phi_+(\mathbf{x}_i)^T\phi_+(\mathbf{x}_j), \tag{6}$$

and

$$\mathcal{K}_-(\mathbf{x}_i,\mathbf{x}_j) = \phi_-(\mathbf{x}_i)^T\phi_-(\mathbf{x}_j). \tag{7}$$

Proof: For an indefinite kernel $\mathcal{K}$, we can always find two PSD kernels $\mathcal{K}_+, \mathcal{K}_-$ and the corresponding feature mappings ϕ_+, ϕ_- to satisfy (5)–(7). Using ϕ_+ and ϕ_- in (4), its Lagrangian of (4) can be written as

$$\begin{aligned} \mathcal{L}(\mathbf{w}_+,\mathbf{w}_-,b,\xi;\alpha) = {} & \frac{1}{2}(\mathbf{w}_+^T\mathbf{w}_+ - \mathbf{w}_-^T\mathbf{w}_-) + \frac{\gamma}{2}\sum_{i=1}^{m}\xi_i^2 \\ & - \sum_{i=1}^{m}\alpha_i\left(y_i(\mathbf{w}_+^T\phi_+(\mathbf{x}_i) + \mathbf{w}_-^T\phi_-(\mathbf{x}_i) + b) - 1 + \xi_i\right). \end{aligned}$$

Then the condition of a stationary point yields

$$\begin{aligned} \frac{\partial\mathcal{L}}{\partial\mathbf{w}_+} &= \mathbf{w}_+ - \sum_{i=1}^{m}\alpha_i y_i\phi_+(\mathbf{x}_i) = 0, \\ \frac{\partial\mathcal{L}}{\partial\mathbf{w}_-} &= -\mathbf{w}_- - \sum_{i=1}^{m}\alpha_i y_i\phi_-(\mathbf{x}_i) = 0, \\ \frac{\partial\mathcal{L}}{\partial b} &= \sum_{i=1}^{m}\alpha_i = 0, \\ \frac{\partial\mathcal{L}}{\partial\xi_i} &= \gamma\xi_i - \alpha_i = 0, \\ \frac{\partial\mathcal{L}}{\partial\alpha_i} &= y_i(\mathbf{w}_+^T\phi_+(\mathbf{x}_i) + \mathbf{w}_-^T\phi_-(\mathbf{x}_i) + b) - 1 + \xi_i = 0. \end{aligned}$$

Eliminating the primal variables $\mathbf{w}_+, \mathbf{w}_-, \xi$, we get the optimality conditions: $\sum_{i=1}^m \alpha_i = 0$ and

$$y_i \Big(\sum_{j=1}^m \alpha_j \phi_+(\mathbf{x}_i)^T \phi_+(\mathbf{x}_j) - \sum_{j=1}^m \alpha_j \phi_-(\mathbf{x}_i)^T \phi_-(\mathbf{x}_j) \Big) - b - \frac{\alpha_i}{\gamma} = 0.$$

Substituting (5)–(7) into the above condition leads to (3). Therefore, (3) is the dual problem of (4). If α_i^*, b^* is the solution of (3), then b^* and

$$\mathbf{w}_+^* = \sum_{i=1}^m \alpha_i^* \phi_+(\mathbf{x}_i), \qquad \mathbf{w}_-^* = \sum_{i=1}^m \alpha_i^* \phi_-(\mathbf{x}_i)$$

satisfy the first-order optimality condition of (4), i.e., $\mathbf{w}_+^*$, $\mathbf{w}_-^*$, b^* is a stationary point of (4). ■

Proposition 1 gives a feature space interpretation for LS-SVM with an indefinite kernel that (3) is still pursing a small within-class scatter in a pE space $\mathbb{R}^{(p,q)}$. Notice that in both (6) and (7), the dimension of the feature mapping could be indefinite and so is the pE space. Based on this interpretation, one can feel free to use indefinite kernels in LS-SVM. In numerical experiments, the performance of an indefinite kernel for LS-SVM classifier will be evaluated.

III. Indefinite Kernel PCA

In the last section, we considered LS-SVM with an indefinite kernel for binary classification. The analysis is applicable to other tasks which can be solved in the framework of LS-SVM. In [18], the link between kernel principal component analysis and LS-SVM has been investigated. Accordingly, we can give the feature space interpretation for kernel PCA with an indefinite kernel.

For given data $\{\mathbf{x}_i\}_{i=1}^m$, the kernel PCA is to solve an eigenvalue problem:

$$\Omega \alpha = \lambda \alpha, \tag{8}$$

where the centered kernel matrix Ω is induced from a kernel $\mathcal{K}$ as follows,

$$\begin{aligned} \Omega_{ij} &= \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) - \frac{1}{m}\sum_{r=1}^m \mathcal{K}(x_i, x_r) \\ &\quad - \frac{1}{m}\sum_{r=1}^m \mathcal{K}(x_j, x_r) + \frac{1}{m^2}\sum_{r=1}^m\sum_{s=1}^m \mathcal{K}(x_r, x_s). \end{aligned}$$

Traditionally, $\mathcal{K}$ is limited to be a PSD kernel and then the kernel trick is valid and (8) maximizes the variance in the related feature space.

Following the same way of introducing an indefinite kernel in C-SVM or LS-SVM, we can directly use an indefinite kernel for kPCA (8). Notice that for an indefinite kernel, the eigenvalues will be positive and negative. All these eigenvalues will be still real for the use of a symmetric kernel. There is no difference on the problem itself and the projected variables can be calculated as the same. However, the feature space interpretation fundamentally changes, which is discussed in the following proposition.

Proposition 2: Suppose α_i^* is the solution of (8) for an indefinite kernel matrix $\mathcal{K}$. There are two feature mappings ϕ_+ and ϕ_- such that

$$\mathbf{w}_+^* = \sum_{i=1}^m \alpha_i^* (\phi_+(\mathbf{x}_i) - \hat{\mu}_{\phi_+}), \quad \mathbf{w}_-^* = \sum_{i=1}^m \alpha_i^* (\phi_-(\mathbf{x}_i) - \hat{\mu}_{\phi_-}),$$

which is a stationary point of the following primal problem:

$$\begin{aligned} \max_{\mathbf{w}_+, \mathbf{w}_-, \xi} \quad & \frac{\gamma}{2}\sum_{i=1}^m \xi_i^2 - \frac{1}{2}(\mathbf{w}_+^T\mathbf{w}_+ - \mathbf{w}_-^T\mathbf{w}_-) \\ \text{s.t.} \quad & \xi_i = \mathbf{w}_+^T(\phi_+(\mathbf{x}_i) - \hat{\mu}_{\phi_+}) \\ & \qquad + \mathbf{w}_-^T(\phi_-(\mathbf{x}_i) - \hat{\mu}_{\phi_-}), \forall i, \end{aligned} \tag{9}$$

where $\hat{\mu}_{\phi_+}, \hat{\mu}_{\phi_-}$ are the centering terms, i.e.,

$$\hat{\mu}_{\phi_+} = \frac{1}{m}\sum_{i=1}^m \phi_+(\mathbf{x}_i) \quad \text{and} \quad \hat{\mu}_{\phi_-} = \frac{1}{m}\sum_{i=1}^m \phi_-(\mathbf{x}_i).$$

Meanwhile, the dual problem of (9) is given by (8) if γ is chosen as $\gamma = \frac{1}{\lambda}$ and $\mathcal{K}$ as in (5)–(7).

Proof: Again, for an indefinite kernel $\mathcal{K}$, we can find two PSD kernels $\mathcal{K}_+, \mathcal{K}_-$, and the corresponding nonlinear feature mappings ϕ_+, ϕ_- to satisfy (5)–(7).

The Lagrangian of (9) can be written as

$$\begin{aligned} \mathcal{L}(\mathbf{w}_+, \mathbf{w}_-, \xi; \alpha) = & -\frac{\gamma}{2}\sum_{i=1}^m \xi_i^2 + \frac{1}{2}(\mathbf{w}_+^T\mathbf{w}_+ - \mathbf{w}_-^T\mathbf{w}_-) \\ & - \sum_{i=1}^m \alpha_i \big(\mathbf{w}_+^T(\phi_+(\mathbf{x}_i) - \hat{\mu}_{\phi_+}) \\ & \qquad + \mathbf{w}_-^T(\phi_-(\mathbf{x}_i) - \hat{\mu}_{\phi_-}) - \xi_i \big). \end{aligned}$$

Then from the conditions of a stationary point, we have

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{w}_+} &= \mathbf{w}_+ - \sum_{i=1}^m \alpha_i(\phi_+(\mathbf{x}_i) - \hat{\mu}_{\phi_+}) = 0, \\ \frac{\partial \mathcal{L}}{\partial \mathbf{w}_-} &= -\mathbf{w}_- - \sum_{i=1}^m \alpha_i(\phi_-(\mathbf{x}_i) - \hat{\mu}_{\phi_-}) = 0, \\ \frac{\partial \mathcal{L}}{\partial \xi_i} &= -\gamma\xi_i + \alpha_i = 0, \\ \frac{\partial \mathcal{L}}{\partial \alpha_i} &= \mathbf{w}_+^T(\phi_+(\mathbf{x}_i) - \hat{\mu}_{\phi_+}) + \\ & \qquad \mathbf{w}_-^T(\phi_-(\mathbf{x}_i) - \hat{\mu}_{\phi_-}) - \xi_i = 0. \end{aligned}$$

Elimination of the primal variables results in the following optimality condition,

$$\begin{aligned} & \frac{1}{\gamma}\alpha_i - \sum_{j=1}^m \alpha_j (\phi_+(\mathbf{x}_j) - \hat{\mu}_{\phi_+})^T(\phi_+(\mathbf{x}_i) - \hat{\mu}_{\phi_+}) \\ & \quad - \sum_{j=1}^m \alpha_j (\phi_-(\mathbf{x}_j) - \hat{\mu}_{\phi_-})^T(\phi_+(\mathbf{x}_i) - \hat{\mu}_{\phi_-}) = 0, \forall i. \end{aligned} \tag{10}$$

Applying the kernel trick (6) and (7), we know that

$$\begin{aligned}&(\phi_\pm(\mathbf{x}_i)-\hat{\mu}_{\phi_+})^T(\phi_\pm(\mathbf{x}_j)-\hat{\mu}_{\phi_\pm})\\ =\;& \mathcal{K}_\pm(\mathbf{x}_i,\mathbf{x}_j)-\frac{1}{m}\sum_{r=1}^{m}\mathcal{K}_\pm(\mathbf{x}_i,\mathbf{x}_r)\\ &-\sum_{r=1}^{m}\mathcal{K}_\pm(x_j,x_r)+\frac{1}{m^2}\sum_{r=1}^{m}\sum_{s=1}^{m}\mathcal{K}_\pm(x_r,x_s).\end{aligned}$$

Additionally with (5), the optimality condition (10) can be formulated as the eigenvalue problem (8). Therefore, (8) is the dual problem of (9) and gives a stationary solution for (9), which aims at having maximal variance in a pE space. ■

IV. Numerical Experiments

In the preceding sections, we discussed the use of indefinite kernels in the framework of LS-SVM for classification and kernel principal component analysis, respectively. The general conclusions are: i) on the computational load of LS-SVM, there is no difference between a PSD kernel and an indefinite kernel, i.e., using an definite kernel in LS-SVM will not additionally bring computational burden; ii) the feature space interpretation of LS-SVM for an indefinite kernel is extended to a pE space and only a stationary point can be obtained.

For a PSD kernel, there is no negative part in (5), i.e., $\mathcal{K}_- = \mathbf{0}$. Thus, indefinite kernels could be regarded as a generalization of PSD ones. If one can find a suitable indefinite kernel, its learning performance could be expected. In this section, we use the following indefinite kernel in LS-SVM:

$$\mathcal{K}(\mathbf{u},\mathbf{v}) = \max\{\rho - \|\mathbf{u}-\mathbf{v}\|_{\ell_1}, 0\}, \tag{11}$$

which is called *truncated ℓ_1 distance (TL1)* kernel and has been recently discussed in [21]. To make a comparison, we also use LS-SVM with the radial basis function (RBF) kernel:

$$\mathcal{K}(\mathbf{u},\mathbf{v}) = \exp\left(-\|\mathbf{u}-\mathbf{v}\|_{\ell_2}^2/\sigma^2\right).$$

The TL1 kernel and the RBF kernel both have one kernel parameter. The same toolbox, namely LS-SVMlab [22], is used to tune the parameters and train the classifiers.

We first consider binary classification problems, for which the data are downloaded from UCI Repository of Machine Learning Datasets [23]. For some datasets, there are both training and test data. Otherwise, we randomly pick half of the data for training and the rest for test. All training data are normalized to $[0,1]^n$ in advance. We focus on the classification accuracy on test data and then report the average accuracy over 10 trials in Table I, where the number of training set m and the problem dimension n are given as well.

Comparing the classification performance of the RBF kernel and TL1 kernel, we find that with the same cross-validation procedure, the TL1 kernel at least gives a similar accuracy as the RBF kernel. For some datasets (Monk1, Monk3, and Splice), the accuracy can be significantly improved if we use an indefinite kernel. Moreover, the TL1 kernel enjoys good stability of the parameter value, which makes its performance for a pre-given value, e.g., $\rho = 0.7n$, being satisfactory for many datasets. With one parameter less to tune, the training time can be largely saved.

TABLE I
Test Accuracy of LS-SVM with RBF and TL1 kernel

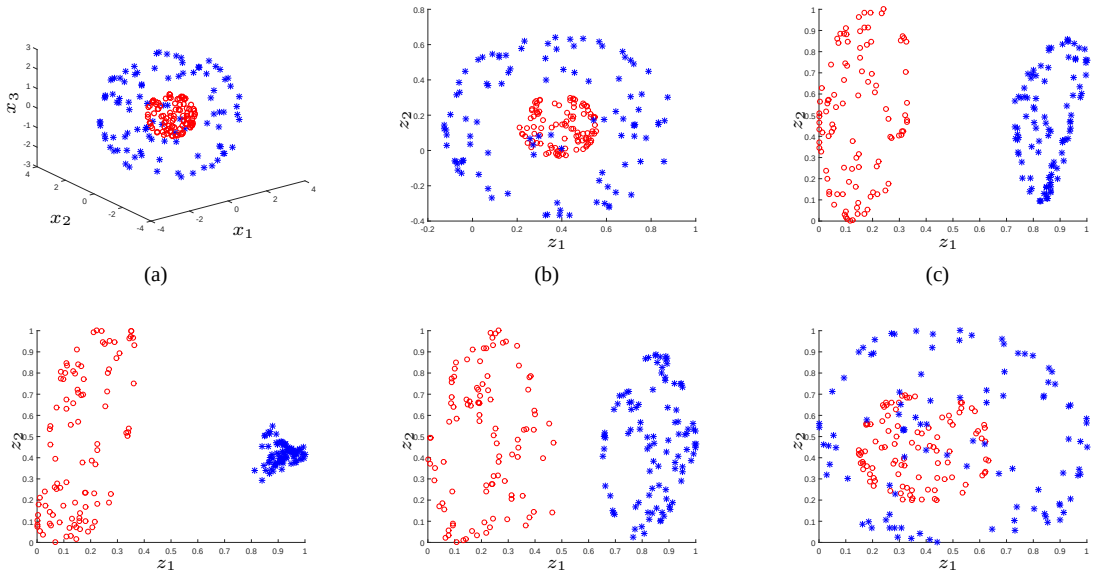
dataset	m	n	RBF σ by CV	TL1 $\rho = 0.7n$	TL1 ρ by CV
DBWords	32	242	84.3%	85.2%	84.4%
Fertility	50	9	86.7%	86.7%	87.8%
Planning	91	12	70.2%	70.6%	73.6%
Sonar	104	60	84.5%	84.3%	83.6%
Statlog	135	13	81.4%	83.8%	83.5%
Monk1	124	6	79.1%	73.4%	85.2%
Monk2	169	6	84.1%	53.4%	83.7%
Monk3	122	6	93.5%	97.2%	97.2%
Climate	270	20	93.2%	91.9%	92.0%
Liver	292	10	70.6%	70.4%	71.8%
Austr.	345	14	85.0%	86.0%	86.7%
Breast	349	10	96.4%	97.0%	97.1%
Trans.	374	4	78.3%	78.4%	78.4%
Splice	1000	121	89.4%	93.6%	94.9%
Spamb.	2300	58	93.1%	94.1%	94.2%
ML-prove	3059	51	72.5%	79.1%	79.3%

In the following, we evaluate the performance of the TL1 kernel in kernel principal competent analysis. Consider a 3-dimensional sphere problem: we have 100 data from the sphere with radius equal to 1 and other 100 data from the sphere with radius equal to 3. The data are shown in Fig.1(a). To reduce the dimension, we apply PCA, kPCA with the RBF kernel, and kPCA with the TL1 kernel, respectively, The obtained two dimensional data are displayed in Fig.1(b)–(f), which roughly imply that with a suitable selected parameter the TL1 kernel gives a good result for kernel principal component analysis.

To quantitatively evaluate kPCA with the TL1 kernel, we choose the problems of which the dimension is higher than 20 from Table I and then apply kPCA to reduce the data into n_r dimension. For the reduced data, linear classifiers, trained from linear C-SVM with libsvm [24], are used to classify the test data. The parameters, including the kernel parameter and the regularization constant in linear C-SVM, are tuned based on 10-fold cross-validation. In Table II, the average classification accuracy of 10 trials for different reduction ratios n_r/n is listed. This experiments again show the good performance of the TL1 kernel that it outperforms the RBF kernel for some datasets.

V. Conclusion

In this paper, we proposed to use indefinite kernels in the framework of least squares support vector machines. In the training problem itself, there is no difference between definite kernels and indefinite kernels. Thus, one can easily use an indefinite kernel in LS-SVM by simply changing the kernel valuation function, which allows us to use LS-SVMlab for the TL1 kernel. Numerically, the TL1 kernel achieves better performance than the RBF kernel for both classification and kernel principal component analysis. The good performance motivates us to investigate the feature space interpretation for an indefinite kernel in LS-SVM, which is the main theoretical



(a) (b) (c)

(d) (e) (f)

Fig. 1. (a) Data points of one class come from the unit sphere and are marked by red circles. The other data points, shown by blue stars, come from a sphere with radius 3. (b) This dataset is not linearly separable and thus linear PCA is not helpful for distinguishing the two classes. Instead, kernel PCA is needed and if the parameter is suitably chosen, the reduced data can be correctly classified by a linear classifier. (c) the RBF kernel with $\sigma = 0.05$; (d) the TL1 kernel with $\rho = 0.1n$; (e) the TL1 kernel with $\rho = 0.2n$; (f) the TL1 kernel with $\rho = 0.5n$, which is similar to linear PCA.

TABLE II
TEST ACCURACY BASED ON KPCA WITH DIFFERENT REDUCTION RATIOS

			Linear			RBF			TL1		
dataset	m	n	10%	30%	50%	10%	30%	50%	10%	30%	50%
Sonar	104	60	72.6%	73.1%	75.9%	75.6%	79.1%	80.7%	77.9%	80.4%	81.9%
Climate	270	21	90.4%	90.9%	91.6%	90.5%	90.8%	91.4%	90.5%	90.9%	91.9%
Qsar	528	41	74.4%	85.4%	85.9%	77.8%	86.4%	86.7%	78.8%	85.9%	86.2%
Splice	1000	60	83.7%	83.9%	84.1%	86.6%	87.7%	87.8%	91.9%	91.1%	91.3%
Spamb.	2300	57	84.7%	87.7%	90.7%	86.5%	89.9%	91.0%	88.4%	91.8%	92.8%
ML-prove	3059	51	59.2%	67.9%	70.2%	69.7%	70.3%	71.0%	70.1%	71.3%	75.5%

contribution of this paper. We hope that the theoretical analysis and good performance shown in this paper can attract research and application interests on LS-SVM with an indefinite kernel in the future.

REFERENCES

[1] V. N. Vapnik, *Statistical Learning Theory*. Wiley New York, 1998.

[2] J.A.K. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural Processing Letters*, vol. 9, no. 3, pp. 293–300, 1999.

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

[4] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[5] C. S. Ong, X. Mary, S. Canu, and A. J. Smola, "Learning with non-positive kernels," in proceeding of *the 21st International Conference on Machine Learning (ICML)*, pp. 639–646, 2004,

[6] Y. Ying, C. Campbell, and M. Girolami, "Analysis of SVM with indefinite kernels," in *Advances in Neural Information Processing Systems 22*, Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, Eds. pp. 2205–2213, 2009.

[7] S. Gu and Y. Guo, "Learning SVM classifiers with indefinite kernels." in proceedings of *the 26th AAAI Conference on Artificial Intelligence*, pp. 942–948, 2012.

[8] G. Loosli, S. Canu, and C. S. Ong, "Learning SVM in Kreĭn spaces." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, doi:10.1109/TPAMI.2015.2477830, 2015.

[9] E. Pekalska and B. Haasdonk, "Kernel discriminant analysis for positive definite and indefinite kernels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 6, pp. 1017–1032, 2009.

[10] S. Zafeiriou, "Subspace learning in Kreĭn spaces: complete kernel Fisher discriminant analysis with indefinite kernels," in proceedins of *European Conference on Computer Vision (ECCV) 2012*, pp. 488–501, 2012.

[11] E. Pekalska, P. Paclik, and R. P. Duin, "A generalized kernel approach to dissimilarity-based classification," *Journal of Machine Learning Research*, vol. 2, pp. 175–211, 2002.

[12] V. Roth, J. Laub, M. Kawanabe, and J. M. Buhmann, "Optimal cluster preserving embedding of nonmetric proximity data," *IEEE Transactions onPattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1540–1551, 2003.

[13] R. Luss and A. d'Aspremont, "Support vector machine classification with indefinite kernels," in *Advances in Neural Information Processing Systems 20*, J.C. Platt, D. Koller, Y. Singer and S.T. Roweis, Eds. pp. 953–960, 2008.

[14] F. Schleif and P. Tino, "Indefinite proximity learning: a review," *Neural Computation*, vol. 27, pp. 2039–2096, 2015.

[15] H.-T. Lin and C.-J. Lin, "A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods." Technical report, Department of Computer Science, National Taiwan University, 2003. http://www.csie.ntu.edu.tw/~cjlin/papers/tanh.pdf

[16] J.C. Platt, "Fast training of support vector machines using sequential

minimal optimization," *Advances in Kernel Methods - Support Vector Learning*, MIT press, pp. 185-208, 1999.

[17] R.-E. Fan, P.-H. Chen, and C.-J. Lin, "Working set selection using second order information for training support vector machines," *Journal of Machine Learning Research*, vol. 6, pp. 1889–1918, 2005.

[18] J.A.K. Suykens, T. Van Gestel, J. Vandewalle, and B. De Moor, "A support vector machine formulation to PCA analysis and its kernel version," *IEEE Transactions on Neural Networks*, vol. 14, no. 2, pp. 447–450, 2003.

[19] B. Haasdonk, "Feature space interpretation of SVMs with indefinite kernels," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 4, pp. 482–492, 2005.

[20] T. Van Gestel, J.A.K. Suykens, G. Lanckriet, A. Lambrechts, B. De Moor, and J. Vandewalle, "Bayesian framework for least-squares support vector machine classifiers, Gaussian processes, and kernel Fisher discriminant analysis," *Neural Computation*, vol. 14, no. 5, pp. 1115–1147, 2002.

[21] X. Huang, J.A.K. Suykens, S. Wang, A. Maier, and J. Hornegger, "Classification with truncated ℓ_1 distance kernel," *Internal Report 15-211, ESAT-SISTA, KU Leuven*, 2015.

[22] K. De Brabanter, P. Karsmakers, F. Ojeda, C. Alzate, J. De Brabanter, K. Pelckmans, B. De Moor, J. Vandewalle, and J.A.K. Suykens, "LS-SVMlab toolbox user's guide," *Internal Report 10-146, ESAT-SISTA, KU Leuven*, 2011.

[23] A. Frank and A. Asuncion, "UCI machine learning repository." School of Information and Computer Science, University of California, Irvine, CA, 2010. http://archive.ics.uci.edu/ml

[24] 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, p. 27, 2011.