Robust Kernel Approximation for Classification

Fanghui Liu[1], Xiaolin Huang[1], Cheng Peng[1], Jie Yang[1] ⋆ and Nikola Kasabov[2]

[1]Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China
[2]Knowledge Engineering and Discovery Research Institute, Auckland University of Technology, Auckland, New Zealand
lfhsgre@outlook.com,xiaolinhuang@sjtu.edu.cn,pynchon1899@gmail.com, jieyang@sjtu.edu.cn,nkasabov@aut.ac.nz

Abstract. This paper investigates a robust kernel approximation scheme for support vector machine classification with indefinite kernels. It aims to tackle the issue that the indefinite kernel is contaminated by noises and outliers, i.e. a noisy observation of the true positive definite (PD) kernel. The traditional algorithms recovery the PD kernel from the observation with the small Gaussian noises, however, such way is not robust to noises and outliers that do not follow a Gaussian distribution. In this paper, we assume that the error is subject to a Gaussian-Laplacian distribution to simultaneously dense and sparse/abnormal noises and outliers. The derived optimization problem including the kernel learning and the dual SVM classification can be solved by an alternate iterative algorithm. Experiments on various benchmark data sets show the robustness of the proposed method when compared with other state-of-the-art kernel modification based methods.

Keywords: robust kernel approximation, indefinite kernel learning, support vector machine

1 Introduction

Kernel methods [1], such as support vector machine (SVM), have been broadly applied in computer vision, bioinformatics, and so on. They often employ a so-called kernel function $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$ to intuitively compute the similarity between two samples $\mathbf{x}_i$ and $\mathbf{x}_j$. If the similarity matrix derived by the kernel function $\mathcal{K}$ is positive semi-definite (PSD), then it can be served as a kernel matrix $\mathbf{K}$[1] in standard kernel methods. To be specific, such positive definite kernel methods are applicable to the support vector machine (SVM) method with remarkable classification performance. In this case, the optimization problem of SVM can be formulated as a convex quadratic programing and well analysed with solid theoretical foundations in the Reproducing Kernel Hilbert Spaces (RKHS).

However, in real-life applications, many potential kernels could be indefinite, such as hyperbolic tangent kernels, a kernel within the Kullback-Leibler divergence, and the protein sequence similarity measures derived from Smith Waterman and BLAST score [2]. In these cases, the derived kernel matrix $\mathbf{K}$

⋆ Corresponding author.

[1] The kernel matrix $\mathbf{K}$ associated to a positive definite kernel $\mathcal{K}$ is PSD.

is no longer guaranteed to be PSD due to the following two reasons. First, the kernel matrix generated by a certain similarity measure takes advantage of domain-specific structure in data and often display excellent empirical classification performance without any positiveness requirement of the kernel matrix. Second, the similarity measurements are easily affected by noises and outliers, which often results in an indefinite kernel matrix. In the next we will introduce the way to tackle such indefinite kernels in these situations, especially for SVM.

To fully exploit the implicit information carried by indefinite kernels in SVM, many algorithms have been proposed in the literature to solve it. One widely used method is kernel approximation, which aims to convert the indefinite kernel matrix to a PSD one by the spectrum modification scheme. For example, "flip" [3] takes the absolute value of all eigenvalues, "clip" [4] neglects the nonnegative eigenvalues and sets them to zero, and in the "shift" [5] method, all eigenvalues plus a positive constant until the smallest one is zero. However, the above three unsophisticated methods actually change the infinite kernel a lot, and thus some important information involved within it might be lost. Comparatively, a substitutable method [6] aims to seek for a PSD kernel matrix $\mathbf{K}$ as the optimal approximation to the indefinite one $\mathbf{K}_0$. In that sense, the indefinite kernel can be considered as a noise-distributed realization of a positive definite kernel. In [6,7], a joint optimization formulation is proposed to simultaneously learn a proxy kernel and the (dual) SVM classification problem. The corresponding objective function in the above two methods adopts a well-known loss, namely the least square error function to obtain a PSD approximation. The implicit rationale of using such loss is that the noise is subject to a Gaussian distribution with small mean and variance. Nevertheless, real data may contain many undesirable noises and outliers not limited to a Gaussian distribution, which makes such loss not appropriate to accommodate the practical use.

Motivated by the above issue, this paper introduces a robust PSD kernel approximation into indefinite learning. The main contributions of our method are summarized as follows:

1. A robust PSD kernel approximation is incorporated into the indefinite SVM framework within a Gaussian-Laplacian distribution noise assumption.
2. An alternate iterative algorithm is proposed to learn a robust PSD kernel approximation and solve the (dual) SVM problem with convergence guarantees.

Numerous experiments on various data sets demonstrate the effectiveness of the proposed robust PSD kernel approximation method.

2 Robust Approximation in SVM with Indefinite Kernels

In this section, we firstly review the regularized formulation of indefinite SVM presented in [6], and then introduce the proposed robust approximation method.

2.1 Review: Indefinite SVM

Let $\mathbf{K} \in \mathbb{S}^n$ be a given kernel matrix and $\mathbf{y} \in \{-1, +1\}^n$ be the vector of labels, with the label matrix $\mathbf{Y} = \text{diag}(\mathbf{y}) \in \mathbb{R}^{n \times n}$, where $\mathbb{S}^n$ is the set of symmetric matrices. Based on the above definitions, the SVM dual form is defined by:

$$\begin{aligned} &\min_{\boldsymbol{\alpha}} \frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{YKY}\boldsymbol{\alpha} - \mathbf{1}^\top \boldsymbol{\alpha} \\ &s.t.\ \ 0 \leq \boldsymbol{\alpha} \leq C\mathbf{1},\ \boldsymbol{\alpha}^\top \mathbf{y} = 0 \end{aligned}, \tag{1}$$

with the variable $\boldsymbol{\alpha} \in \mathbb{R}^n$, where C is the fixed tradeoff parameter and $\mathbf{1} \in \mathbb{R}^n$ denotes the all-one vector. Suppose that only an indefinite kernel matrix $\mathbf{K}_0 \in \mathbb{S}^n$ is given, Luss and d' Aspremont [6] proposed the following max-min method to simultaneously learn a proxy PSD kernel matrix $\mathbf{K}$ for $\mathbf{K}_0$ and the SVM classification problem:

$$\begin{aligned} &\max_{\boldsymbol{\alpha}}\ \min_{\mathbf{K} \succeq 0} \mathbf{1}^\top \boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{YKY}\boldsymbol{\alpha} + \rho\|\mathbf{K} - \mathbf{K}_0\|_F^2 \\ &s.t.\ \ \boldsymbol{\alpha}^\top \mathbf{y} = 0,\ 0 \leq \boldsymbol{\alpha} \leq C\mathbf{1} \end{aligned}, \tag{2}$$

where ρ is a regularization parameter. Observe that, the inner minimization problem is a convex conic program on $\mathbf{K}$, and the outer optimization problem is also convex. As a result, Eq. (2) is a concave maximization problem subject to linear constraints and thus is a convex problem of $\boldsymbol{\alpha}$. Specifically, the inner kernel learning optimization problem can be equivalent to a projection to a semi-definite cone, which arrives at:

$$\min_{\mathbf{K} \succeq 0} -\frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{YKY}\boldsymbol{\alpha} + \rho\|\mathbf{K} - \mathbf{K}_0\|_F^2 \tag{3}$$

Given $\boldsymbol{\alpha}$, the optimal solution to this problem is then given by:

$$\mathbf{K}^* = \big(\mathbf{K}_0 + \frac{1}{4\rho}\mathbf{Y}\boldsymbol{\alpha}\boldsymbol{\alpha}^\top \mathbf{Y}\big)_+ \tag{4}$$

where $\mathbf{K}_+$ is the positive part of the matrix $\mathbf{K}$, i.e., $\mathbf{K}_+ = \sum_i \max(0, \lambda_i)\mathbf{p}_i\mathbf{p}_i^\top$, where λ_i and $\mathbf{p}_i$ are the ith eigenvalue and eigenvector of the matrix $\mathbf{K}$, respectively. And then, plugging this solution into Eq. (2), we can get the optimization problem associated with $\boldsymbol{\alpha}$. Thereby, the learning proxy PSD kernel matrix $\mathbf{K}$ and SVM classification problem can be simultaneously learned by solving $\boldsymbol{\alpha}$.

2.2 Robust Approximation

In another view, the kernel learning in Eq. (3) aims to fit a PSD kernel $\mathbf{K}$ to a noisy observation $\mathbf{K}_0$, namely $\mathbf{K}_0 = \mathbf{K} + \mathbf{E}$, where $\mathbf{E}$ is defined as the error or residual term. The used Frobenius norm in Eq. (3) indicates that the error $\mathbf{E} = \mathbf{K}_0 - \mathbf{K}$ follows the Gaussian distribution ($e_{ij} \sim \mathcal{N}(0, \sigma_N^2)$[2]). However, such

[2] The probability density function of a Gaussian random variable x is defined as $f_{\mathcal{N}}(x) = \frac{1}{\sqrt{2\pi}\sigma_N} \exp\big(-\frac{\sqrt{2}x^2}{\sigma_N^2}\big)$.

solution only resists on Gaussian noises and can hardly tackle outliers and other undesirable noises. To remedy this defect, in our method, we assume that the error $\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2$ is modeled as an additive combination of two independent components: an i.i.d Gaussian noise matrix $\mathbf{E}_1$ and an i.i.d Laplacian noise matrix $\mathbf{E}_2$ [3]), where the Gaussian component models small dense (non-sparse) noise and the Laplacian (sparse) one aims to handle outliers [8]. Therefore, we incorporate the error term $\mathbf{E}$ into Eq. (3) with a uniform framework:

$$\min_{\mathbf{K}\succeq 0,\mathbf{E}} -\frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{Y}\mathbf{K}\mathbf{Y}\boldsymbol{\alpha} + \rho\|\mathbf{K}-\mathbf{K}_0-\mathbf{E}\|_F^2 + \gamma\|\mathbf{E}\|_1\,, \tag{5}$$

where γ controls the sparsity of the error matrix $\mathbf{E}$. By such modeling, the error matrix $\mathbf{E}$, as a mixture Gaussian-Laplacian distribution, comprehensively considers the diversity of noises and outliers in real-life data. Accordingly, by combining the kernel approximation problem demonstrated in Eq. (5) and the dual SVM classification problem, the final optimization problem can be formulated as:

$$\begin{aligned} &\max_{\boldsymbol{\alpha}} \ \min_{\mathbf{K}\succeq 0,\mathbf{E}} \mathbf{1}^\top\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^\top\mathbf{Y}\mathbf{K}\mathbf{Y}\boldsymbol{\alpha} + \rho\|\mathbf{K}-\mathbf{K}_0-\mathbf{E}\|_F^2 + \gamma\|\mathbf{E}\|_1 \\ &s.t.\ \ \boldsymbol{\alpha}^\top\mathbf{y} = 0,\ 0\le\boldsymbol{\alpha}\le C\mathbf{1} \end{aligned}\,. \tag{6}$$

Several optimization algorithms for such convex optimization problem have been well investigated, such as the analytic center cutting plane method [6] and the projection gradient method with Nesterov's smooth optimization [7]. However, due to the non-smooth regularization term $\|\cdot\|_1$, the above gradient based methods cannot be directly applied to solve this optimization problem. In this paper, we introduce an alternate iterative algorithm to tackle this non-smooth term and then solving the inner optimization problem. To be specific, Eq. (5) can be reformulated as the following formula:

$$\min_{\mathbf{K}\succeq 0,\mathbf{E}} \mathcal{O}(\mathbf{K},\mathbf{E}) = \|\mathbf{K} - (\mathbf{K}_0 + \frac{1}{4\rho}\mathbf{Y}\boldsymbol{\alpha}\boldsymbol{\alpha}^\top\mathbf{Y}^\top) - \mathbf{E}\|_F^2 + \gamma\|\mathbf{E}\|_1\,. \tag{7}$$

Given the solution $\mathbf{E}^{(t)}$ at t-th iteration, the solution $\mathbf{K}^{(t+1)}$ can be solved by a semi-definite programming, which arrives at:

$$\mathbf{K}^{(t+1)} = \big(\mathbf{K}_0 + \frac{1}{4\rho}\mathbf{Y}\boldsymbol{\alpha}\boldsymbol{\alpha}^\top\mathbf{Y} + \mathbf{E}^{(t)}\big)_+\,. \tag{8}$$

Given the learned kernel $\mathbf{K}^{(t)}$ at t-th iteration, the optimal error matrix $\mathbf{E}^{(t+1)}$ can be solved by the shrinkage thresholding algorithm [9]:

$$[\mathbf{E}^{(t+1)}]_{ij} = \mathcal{S}\Big(\frac{\gamma}{2}, \big[\mathbf{K}^{(t)} - \mathbf{K}_0 - \frac{1}{4\rho}\mathbf{Y}\boldsymbol{\alpha}\boldsymbol{\alpha}^\top\mathbf{Y}\big]_{ij}\Big)\,, \tag{9}$$

where the shrinkage operator is defined as $\mathcal{S}(\varepsilon, x) = \mathrm{sgn}(x)\cdot\max(|x|-\varepsilon, 0)$, and $\mathrm{sgn}(\cdot)$ is a sign function. Finally, the algorithm to learn the PSD kernel $\mathbf{K}$ and the error $\mathbf{E}$ is summarized in **Algorithm 1**.

[3] The probability density function of a Laplacian random variable x is defined as $f_{\mathcal{L}}(x) = \frac{1}{\sqrt{2}\sigma_L}\exp\big(-\frac{\sqrt{2}|x|}{\sigma_L}\big)$.

Algorithm 1: Algorithm for $\mathbf{K}^*$ and $\mathbf{E}^*$.

Input: A given $\boldsymbol{\alpha}$, the indefinite kernel matrix $\mathbf{K}_0$, the regularization parameters are $\rho = 10$ and $\gamma = 1$.

Output: The optimal $\mathbf{K}^*$ and $\mathbf{E}^*$.

Set: stop error $\varepsilon = 10^{-4}$.

Initialize $i = 0$, $\mathbf{E}$ with random positive values, and a symmetric matrix $\mathbf{K}$.

Compute the objective function value $\mathcal{O}^{(i)}(\mathbf{K}^{(i)}, \mathbf{E}^{(i)})$.

Repeat

$\mathbf{K}^{(i+1)} := \left(\mathbf{K}_0 + \frac{1}{4\rho}\mathbf{Y}\boldsymbol{\alpha}\boldsymbol{\alpha}^\top\mathbf{Y} + \mathbf{E}^{(i)}\right)_+$;

$[\mathbf{E}^{(i+1)}]_{ij} = \mathcal{S}\left(\frac{\gamma}{2}, \left[\mathbf{K}^{(i)} - \mathbf{K}_0 - \frac{1}{4\rho}\mathbf{Y}\boldsymbol{\alpha}\boldsymbol{\alpha}^\top\mathbf{Y}\right]_{ij}\right)$;

Compute the current objective function value $\mathcal{O}^{(i+1)}(\mathbf{K}^{(i+1)}, \mathbf{E}^{(i+1)})$;

$i := i + 1$;

Until $\frac{\|\mathcal{O}^{(i+1)} - \mathcal{O}^{(i)}\|_2}{\|\mathcal{O}^{(i)}\|_2} \leq \varepsilon$;

Next we discuss the convergence analysis of **Algorithm 1**. The optimization algorithm for minimizing the objective function $\mathcal{O}(\mathbf{K}, \mathbf{E})$ is essentially iterative. In order to prove the convergence, it is necessary to show that $\mathcal{O}(\mathbf{K}, \mathbf{E})$ is non-increasing under the optimization steps listed in **Algorithm 1**. It is clear that the objective function $\mathcal{O}(\mathbf{K}, \mathbf{E})$ satisfies:

$$\begin{aligned}
&\mathcal{O}(\mathbf{K}^{(i+1)}, \mathbf{E}^{(i)}) = \operatorname*{argmin}_{\mathbf{K}} \mathcal{O}(\mathbf{K}, \mathbf{E}^{(i)}) \leq \mathcal{O}(\mathbf{K}^{(i)}, \mathbf{E}^{(i)}), \\
&\mathcal{O}(\mathbf{K}^{i+1}, \mathbf{E}^{(i+1)}) = \operatorname*{argmin}_{\mathbf{E}} \mathcal{O}(\mathbf{K}^{(i+1)}, \mathbf{E}) \leq \mathcal{O}(\mathbf{K}^{(i+1)}, \mathbf{E}^{(i)}) \leq \mathcal{O}(\mathbf{K}^{(i)}, \mathbf{E}^{(i)}).
\end{aligned} \tag{10}$$

which completes the proof. After obtaining the optimal $\mathbf{K}^*$ and $\mathbf{E}^*$, the learned PSD kernel $\mathbf{K}^*$ can be used for solving the dual variable $\boldsymbol{\alpha}$. Therefore, the algorithm for learning $\mathbf{K}^*$ and $\boldsymbol{\alpha}$ is summarized in **Algorithm 2**. Specifically, such iteration algorithm converges very fast, usually within 15 iterations.

Algorithm 2: Algorithm for $\boldsymbol{\alpha}^*$, $\mathbf{K}^*$ and $\mathbf{E}^*$.

Input: The training set label $\mathbf{Y}$, the indefinite kernel matrix $\mathbf{K}_0$

Output: The optimal $\boldsymbol{\alpha}^*$

Set the maximum iteration number $T = 3$.

Initialize $i = 0$, $\mathbf{E}$ with random positive values, a symmetric matrix $\mathbf{K}$, and a random nonnegative vector $\boldsymbol{\alpha}$.

Repeat

Obtain $\mathbf{K}^{(i+1)}$ and $[\mathbf{E}^{(i+1)}]_{ij}$ by **Algorithm 1**;

Solve $\boldsymbol{\alpha}$ with the learned kernel $\mathbf{K}^{(i+1)}$ by SMO algorithm [10];

$i := i + 1$;

Until $i \geq T$;

Remark: If the outer iteration number T is fixed to 1 (i.e., without iteration in the outer loop), **Algorithm 2** with a warm start outputs the similar optimal result with the iteration version. Hence, such setting does not largely decrease the final performance with a high computational efficieny.

3 Experiments

In this section, we compare the proposed method with other methods using SVM given an indefinite similarity measure. These algorithms are tested on several benchmark data sets from the UCI repository [11] including *Monks1*, *Monks3*, and *SPECT*, and two data sets from USPS handwritten digits dataset [12] using the indefinite Simpson score (SS).

3.1 Experiments Setup

The compared kernel approximation algorithms includes three spectrum modification methods *flip*, *clip*, and *shift*, and two PSD kernel learning based approaches SVM-PG [13] and SVM-SMM [7]. Specifically, SVM with the original indefinite kernel, as a baseline method, is also taken into comparison (In this case, SVM would converge but the solution is only a stationary point and is not guaranteed to be optimal).

For each data set, we randomly pick up the half of the data for training and the rest for test. Specifically, for all methods, the parameter C is tuned by five-fold cross-validation on the training set: one of these five subsets is used for validation in turn and the remaining ones for training. The stopping error is set to 10^{-4}.

3.2 Generalization on the Noisy Kernel

To verify the effectiveness of the proposed method robust to kernels with noises, an indefinite kernel is added with small noises, i.e. $\mathbf{K}_0 := \mathbf{K} - 0.1\hat{\mathbf{E}}$, where the noisy matrix $\hat{\mathbf{E}}$ is randomly generated by zero mean and identity covariance matrix, and specifically, we randomly select some elements fixed to a large number (i.e. 10000). For USPS-3-5-SS and USPS-4-6-SS, the kernel matrix $\mathbf{K}$ is derived from the indefinite Simpson score (SS) in [13]. For *Monks1*, *Monks3*, and *SPECT* data sets, the kernel function $\mathcal{K}$ is chosen as a Gaussian kernel, which follows with the same setting in [7].

Tab. 1 provides statistics including the minimum and maximum eigenvalues of the training kernels, i.e. $\lambda_{\min}$ and $\lambda_{\max}$. Observe that, the USPS data uses highly indefinite kernels while the UCI data use kernels that are nearly positive semi-definite. The classification performance of our method and other algorithms are evaluated by accuracy[4] and recall[5] as shown in Tab. 1. One can see that the

[4] Accuracy is defined as the percentage of total instances predicted correctly.
[5] Recall is the percentage of true positives that were correctly predicted positive.

Table 1. Data set statistics and performance on the noisy kernel. The best scores are highlighted by **bold**.

Dataset	USPS-3-5-SS		USPS-4-6-SS		Monks1		Monks3		SPECT	
$\lambda_{\min}$, $\lambda_{\max}$	-34.76	453.6	-37.30	413.2	-0.72	11.43	-0.74	11.93	-0.53	9.16
Measure	Accuracy	Recall	Accuracy	Recall	Accuracy	Recall	Accuracy	Recall	Accuracy	Recall
SVM	74.90	72.73	90.08	88.49	51.61	62.07	57.38	55.17	74.47	79.36
Flip	95.73	95.45	97.90	98.65	58.73	46.88	62.90	54.55	71.74	76.74
Clip	95.47	94.50	97.78	98.42	55.56	48.39	56.45	65.63	70.53	72.94
Shift	90.43	92.11	94.28	93.68	52.38	47.50	53.23	55.17	73.68	76.09
SVM-PG [6]	96.25	96.65	97.90	**98.87**	**61.29**	79.31	70.49	58.62	67.37	68.18
SVM-SMM [7]	94.67	92.67	93.67	96.67	59.68	**88.89**	67.21	45.95	68.09	70.93
Ours	**97.67**	**98.33**	**98.0**	**98.87**	52.38	62.96	**70.97**	**82.14**	**74.74**	**81.61**

proposed method achieves a promising performance on the USPS data, *Monks3*, and *SPECT* with the highest accuracy and recall. The results on these data sets demonstrate that the proposed method is robust to noises and outliers, and it tackles the highly indefinite kernel better than the nearly PSD one.

Discussion: Apart from kernel approximation, there are also two kinds of algorithms to tackle such indefinite kernels. One approach is to directly solve the corresponding non-convex problem via some non-convex optimization algorithms. For example, in [14], the authors utilize the concave-convex procedure (CCCP) [15] algorithm for SVM with indefinite kernels. The other solution to such problem is to learn indefinite kernels in the Reproducing Kernel Kreĭn Spaces (RKKS) [16,17,18] with theoretical guarantees.

4 Conclusion

This paper proposes a robust PSD kernel approximation scheme in indefinite kernel learning. The Gaussian and Laplacian noise assumption makes our method more flexible to tackle the indefinite kernel with a noisy instance of a true kernel. The corresponding robust kernel learning problem can be solved by an alternate iterative algorithm with a semi-definite programming and a soft-threshold operator with theoretical guarantees. Quantitative comparisons with other state-of-the-art kernel approximation based methods on several data sets have demonstrated the effectiveness and robustness of the proposed method.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 61572315, Grant 6151101179, and Grant 61603248, in part by 863 Plan of China under Grant 2015AA042308.

References

1. Schölkopf, B., Smola, A.J.: Learning with kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press (2003)
2. Saigo, H., Vert, J.P., Ueda, N., Akutsu, T.: Protein homology detection using string alignment kernels. Bioinformatics 20(11), 1682–1689 (2004)
3. Graepel, T., Herbrich, R., Bollmann-Sdorra, P., Obermayer, K.: Classification on pairwise proximity data. In: Proceedings of Advances in Neural Information Processing Systems. vol. 11, pp. 438–444 (1999)
4. Pekalska, E., Paclik, P., Duin, R.P.W.: A generalized kernel approach to dissimilarity-based classification. Journal of Machine Learning Research 2(2), 175–211 (2002)
5. Roth, V., Laub, J., Kawanabe, M., Buhmann, J.: Optimal cluster preserving embedding of nonmetric proximity data. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(12), 1540–1551 (2003)
6. Luss, R., DAspremont, A.: Support vector machine classification with indefinite kernels. In: Proceedings of Advances in Neural Information Processing Systems. pp. 953–960 (2008)
7. Ying, Y., Campbell, C., Girolami, M.: Analysis of SVM with indefinite kernels. In: Proceedings of Advances in Neural Information Processing Systems. pp. 2205–2213 (2009)
8. Liu, F., Liu, M., Zhou, T., Qiao, Y., Yang, J.: Incremental robust nonnegative matrix factorization for object tracking. pp. 611–619. Proceedings of the International Conference on Neural Information Processing (2016)
9. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM journal on imaging sciences 2(1), 183–202 (2009)
10. Platt, J.C.: ℓ_2 fast training of support vector machines using sequential minimal optimization. Advances in kernel methods (1999)
11. Blake, C., Merz, C.J.: UCI repository of machine learning databases (1998), `http://archive.ics.uci.edu/ml/`
12. Hull, J.J.: A database for handwritten text recognition research. IEEE Transactions on Pattern Analysis and Machine Intelligence 16(5), 550–554 (1994)
13. Huang, X., Suykens, J.A., Wang, S., Hornegger, J., Maier, A.: Classification with truncated ℓ_1 distance kernel. IEEE Transactions on Neural Networks and Learning Systems (2017)
14. Xu, H., Xue, H., Chen, X., Wang, Y.: Solving indefinite kernel support vector machine with difference of convex functions programming. In: Proceedings of AAAI Conference on Artificial Intelligence. pp. 1610–1616 (2017)
15. Yuille, A.L., Rangarajan, A.: The concave-convex procedure. Neural Computation 15(4), 915–936 (2003)
16. Ong, C.S., Mary, X., Smola, A.J.: Learning with non-positive kernels. In: Proceedings of International Conference on Machine Learning. pp. 81–89 (2004)
17. Loosli, G., Canu, S., Cheng, S.O.: Learning SVM in Kreĭn spaces. IEEE Transactions on Pattern Analysis and Machine Intelligence 38(6), 1204–1216 (2016)
18. Huang, X., Maier, A., Hornegger, J., Suykens, J.A.K.: Indefinite kernels in least squares support vector machines and principal component analysis. Applied and Computational Harmonic Analysis 43(1), 162–172 (2017)