Proceedings of Machine Learning Research vol 278:1–8, 2025 2025 2nd International Conference on Machine Learning and Intelligent Computing

Semi-Supervised L2KC (S-L2KC) Classifier

Yue Lv LY1251961956@163.COM
Jiangnan University, Wuxi, Jiangsu, China

Editors: Nianyin Zeng, Ram Bilas Pachori and Dongshu Wang

Abstract

Building upon the density difference paradigm, a novel kernel classifier distinct from Support Vector Machines (SVM) — the L2-norm Kernel Classifier (L2KC) — has been developed. This methodology establishes an integrated squared error(ISE) criterion to estimate the true $d_\gamma(x)$ through minimizing the L2-distance between $d_\gamma(x)$ and $\widehat{d}_\gamma(x)$, thereby achieving classification via explicit density difference representation. While L2KC demonstrates comparable accuracy to SVM with enhanced decision efficiency, its performance on real-world semi-supervised datasets requires improvement. To address this limitation, we propose the Semi-supervised L2KC (S-L2KC) by incorporating a locality-preserving projection (LPP) based manifold regularization term into the L2KC objective function. This integration effectively enforces the manifold assumption. Experimental results on benchmark datasets from the UCI and LIBSVM demonstrate that compared to L2KC, the proposed S-L2KC exhibits superior generalization capability, characterized by higher mean test accuracy with comparable or even smaller variance.

Keywords: L2KC; Semi-supervised Learning; Manifold Learning; Local Preserving Projections

1. Introduction

The L2-norm kernel classifier (L2KC) proposed by Kim and Scott (2010) is a classification algorithm analogous to Support Vector Machines (SVM) (Cortes and Vapnik, 1995). Consider a binary classification problem with samples and their corresponding class labels denoted as $(\mathbf{x}_1, y_1),\ldots,(\mathbf{x}_n, y_n)$, where $\mathbf{x}_i \in R^d$ represents a d-dimensional sample and $y_i \in \{1,-1\}$ its class label. Conventional kernel classifiers such as SVM employ a decision function (without offset term):

$$g(\mathbf{x}) = sign\left\{\sum_{i=1}^{n} \alpha_i y_i k_\sigma(\mathbf{x}, \mathbf{x}_i)\right\} \tag{1}$$

In contrast, the L2KC proposed by Kim and Scott (2010) adopts a density difference paradigm. Let $f_+(\mathbf{x})$ and $f_-(\mathbf{x})$ denote the conditional probability density functions for the positive and negative class sample sets, respectively. The density difference is formulated as $\widehat{d}_\gamma(\mathbf{x};\alpha) = f_+(\mathbf{x}) - \gamma f_-(\mathbf{x})$. According to statistical decision theory, the optimal classifier takes the form:

$$g^*(\mathbf{x}) = sign\{f_+(\mathbf{x}) - \gamma f_-(\mathbf{x})\} \tag{2}$$

where γ denotes a user-defined fixed parameter that accounts for the class prior probabilities.

Kim and Scott (2010) redefine the class labels $y_i \in \{-\gamma, 1\}$ and, based on the kernel density estimation (KDE) method, formulate the density functions for positive and negative class samples with weighting factors $\alpha = (\alpha_1, \alpha_2, ..., \alpha_n)$ as:

$$\begin{aligned} \widehat{f}_+(\mathbf{x};\alpha) &= \sum_{i\in I_+} \alpha_i k_\sigma(\mathbf{x}, \mathbf{x}_i) \\ \widehat{f}_-(\mathbf{x};\alpha) &= \sum_{i\in I_-} \alpha_i k_\sigma(\mathbf{x}, \mathbf{x}_i) \end{aligned} \tag{3}$$

Then, an ISE function was formulated to estimate $d_\gamma(\mathbf{x})$ via $\hat{d}_\gamma(\mathbf{x};\alpha) = \hat{f}_+(\mathbf{x}) - \gamma\hat{f}_-(\mathbf{x})$, with parameter α estimated by minimizing the L2 distance between $\hat{d}_\gamma(\mathbf{x})$ and $d_\gamma(\mathbf{x})$. The mathematical expression is given by:

$$\begin{aligned} ISE(\alpha) =& \| \widehat{d}_\gamma(\mathbf{x};\alpha) - d_\gamma(\mathbf{x}) \|_{L_2}^2 \\ =& \int \left(\widehat{d}_\gamma(\mathbf{x};\alpha) - d_\gamma(\mathbf{x})\right)^2 d\mathbf{x} \\ =& \int \widehat{d}_\gamma^2(\mathbf{x};\alpha) d\mathbf{x} - 2\int \widehat{d}_\gamma(\mathbf{x};\alpha) d_\gamma(\mathbf{x}) d\mathbf{x} + \int d_\gamma^2(\mathbf{x}) d\mathbf{x} \end{aligned} \tag{4}$$

The optimization problem in Eq. (4) can be reformulated as a quadratic programming (QP) problem:

$$\hat{\alpha} = \arg\min_{\alpha\in A} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j k_{\sqrt{2}\sigma}(\mathbf{x}_i,\mathbf{x}_j) - \sum_{i=1}^{n} c_i\alpha_i \tag{5}$$

The corresponding dual formulation is derived as:

$$\begin{aligned} &\min_{w,\zeta_+,\zeta_-} J(w) = \frac{1}{2}\mathbf{w}^2 + \zeta_+ + \zeta_- \\ &s.t.\ y_i\mathbf{w}^T\phi_{\sqrt{2}\sigma}(\mathbf{x}_i) \geq c_i - \zeta_+, i \in I_+ \\ &s.t.\ y_i\mathbf{w}^T\phi_{\sqrt{2}\sigma}(\mathbf{x}_i) \geq c_i - \zeta_-, i \in I_- \end{aligned} \tag{6}$$

where $\phi_\sigma(\mathbf{x})$ denotes the kernel-induced mapping in the Gaussian reproducing kernel Hilbert space (RKHS). Differing from SVM, the L2-kernel classifier replaces the unit margin constraint with a flexible parameter c_i, implying that sufficiently large c_i values prioritize classification correctness of training samples $\mathbf{x}_i$. Furthermore, it employs a single smoothing factor ζ_+, ζ_- per class without requiring sign-specific constraints.

2. S-L2KC

Currently, there are many semi-supervised learning algorithms, such as:

- Semi-Supervised Support Vector Machines (S3VMs) (Bennett and Demiriz, 1998).
- MixMatch, which combines consistency regularization and entropy minimization to improve pseudo-labeling (Berthelot et al., 2019).
- SimPLE, which optimizes classification networks using three training objectives and introduces pairwise loss for better feature learning (Hu and et al., 2021).
- UPS, which incorporates negative samples to filter out uncertain labels during training (Rizve et al., 2021).
- FreeMatch, which dynamically adjusts confidence thresholds based on the model's learning state (Wang and et al., 2022).

- PRCL, which enhances the Robustness of Contrastive Learning through Probabilistic Representation (Xie and et al., 2024).
- FixMatch, which integrates Consistency Regularization and Pseudo-Labeling Techniques (Sohn and et al., 2020).
- UniMatch, which extends FixMatch by strengthening the perturbation approach through unified image-level and feature-level perturbations (Yang et al., 2023).
- FlexMatch, which proposes a Curriculum Learning Approach – Curriculum Pseudo-Labeling (Zhang and et al., 2021).
- Consistent-Teacher: Triple Solutions for Stable Pseudo-labels (ASA + FAM-3D + GMM-Thresh) (Wang and et al., 2023).

To enhance the semi-supervised classification capability of L2KC, we incorporate a manifold regularization term based on locality-preserving projection (LPP) (He and Niyogi, 2005). The standard LPP objective function is formulated as:

$$\mathbf{w}^T \sum_{i=1}^{n} \sum_{j=1}^{n} (\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j))(\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j))^T \mathbf{w} \tag{7}$$

The Gram matrix $(\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j))(\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j))^T$, constructed from unlabeled instances, is incorporated as a manifold regularization term in the objective function to preserve the intrinsic geometric structure embedded in the data distribution.

By integrating the aforementioned manifold hypothesis into the optimization objective of S-L2KC through manifold regularization (Wang and et al., 2022), the enhanced objective function can be formulated as:

$$\begin{aligned} &\min_{w,\zeta_+,\zeta_-} J(w) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{\mu}{2}\mathbf{w}^T A\mathbf{w} + \zeta_+ + \zeta_- \\ &s.t.\ y_i \mathbf{w}^T \phi_{\sqrt{2}\sigma}(\mathbf{x}_i) \geq c_i - \zeta_+, i \in I_+ \\ &s.t.\ y_i \mathbf{w}^T \phi_{\sqrt{2}\sigma}(\mathbf{x}_i) \geq c_i - \zeta_-, i \in I_- \end{aligned} \tag{8}$$

Formulating the Lagrangian function for the constrained optimization problem stated above leads to the following primal-dual relationship:

$$\begin{aligned} &L(\mathbf{w}, \zeta_+, \zeta_-, \alpha, \beta) = \hat{J}(w) \\ &- \sum_{i \in M, i \in I_+} \alpha_i (y_i \mathbf{w}^T \phi_{\sqrt{2}\sigma}(\mathbf{x}_i) - c_i + \zeta_+) \\ &- \sum_{i \in M, i \in I_-} \beta_i (y_i \mathbf{w}^T \phi_{\sqrt{2}\sigma}(\mathbf{x}_i) - c_i + \zeta_-) \end{aligned} \tag{9}$$

Through differentiation of the Lagrangian with respect to variables, the KKT optimality conditions enforce the following system of stationary equations:

$$\begin{aligned}
&\frac{\partial L}{\partial w} = 0 \rightarrow \mathbf{w} = \sum_{i\in M, i\in I_+} \alpha_i y_i (I+\mu A)^{-1} \phi_{\sqrt{2}\sigma}(\mathbf{x}_i)) + \sum_{i\in M, i\in I_-} \beta_i y_i (I+\mu A)^{-1} \phi_{\sqrt{2}\sigma}(\mathbf{x}_i)) \\
&\frac{\partial L}{\partial \zeta_+} = 0 \rightarrow \sum_{i\in M, i\in I_+} \alpha_i = 1 \\
&\frac{\partial L}{\partial \zeta_-} = 0 \rightarrow \sum_{i\in M, i\in I_-} \beta_i = 1 \\
&\frac{\partial L}{\partial \alpha} = 0 \rightarrow c_i - \zeta_+ = y_i \mathbf{w}^T \phi_{\sqrt{2}\sigma}(\mathbf{x}_i) \\
&\frac{\partial L}{\partial \beta} = 0 \rightarrow c_i - \zeta_- = y_i \mathbf{w}^T \phi_{\sqrt{2}\sigma}(\mathbf{x}_i)
\end{aligned} \tag{10}$$

Through algebraic manipulation and substitution of the stationary conditions, the optimization problem reduces to the following canonical form:

$$\begin{aligned}
&B = \{\alpha | \sum_{i\in I_+} \alpha_i = \sum_{i\in I_-} \alpha_i = 1, i \in N\} \\
&\frac{1}{2} \sum_{i\in N} \sum_{j\in M} \alpha_i \alpha_j y_i y_j \phi_{\sqrt{2}\sigma}(\mathbf{x}_i)^T (I+\mu A)^{-1} \phi_{\sqrt{2}\sigma}(\mathbf{x}_j)) = \sum_{i\in N} \alpha_i c_i - \zeta
\end{aligned} \tag{11}$$

The primal optimization problem is thus reduced to a convex quadratic programming problem, formally expressed as:

$$\hat{\alpha} = \arg\min_{\alpha\in B} \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{M} \alpha_i \alpha_j y_i y_j \phi_{\sqrt{2}\sigma}(\mathbf{x}_i)^T (I+\mu A)^{-1} \phi_{\sqrt{2}\sigma}(\mathbf{x}_j) - \sum_{i=1}^{n} c_i \alpha_i \tag{12}$$

Solving the quadratic programming problem yields the dual coefficients , which can be substituted into the decision function to obtain the final S-L2KC classifier formulation:

$$g^*(\mathbf{x}) = sign \left\{ \sum_{i=1}^{N} \alpha_i y_i \phi_{\sqrt{2}\sigma}(\mathbf{x}_i)^T (I+\mu A)^{-1} \phi_{\sqrt{2}\sigma}(\mathbf{x}) \right\} \tag{13}$$

3. Experimental results and analysis

This paper utilizes several datasets from the UCI and LIBSVM databases, with the sample sizes and feature information of the relevant datasets detailed in Table 1. The study employs a 10-folds cross-validation approach, where the preprocessed dataset is divided into ten parts. Nine parts are selected as the training set, and the remaining one serves as the test set. To more accurately demonstrate the classification performance of the algorithm, we use the average accuracy and variance obtained from performing 50 iterations of ten-fold cross-validation on different datasets as the performance metrics.

Table 1: Dataset details

Dataset	Sample Size	Feature Count
Diabetes	768	8
German	1000	20
Heart	270	13
Image	2310	18
Ionosphere	351	34
Breast-cancer	277	9
sonar	208	60
Waveform	5000	21
Splice	3175	60
Ringnorm	7400	20

Based on real-world experience, the proportion of unlabeled samples in a dataset is generally high. However, considering extreme conditions is not conducive to the algorithm's learning in semi-supervised scenarios. Therefore, in the experimental design, this paper sets the ratio (r) of unlabeled samples to total samples as (0.6, 0.7, 0.8) to test the algorithm's performance under reasonable proportions of unlabeled data and examines the algorithm's robustness by varying this ratio. Table 2 presents the relevant parameter settings for this experiment. During the experiment, the hyperparameters of the L2KC and S-L2KC algorithms were determined through grid search. A Gaussian kernel function was employed, with the hyperparameter (σ) ranging from 10^2 to 10^{-2} to avoid kernel matrix sparsity caused by excessively small values, while allowing σ to be increased when unlabeled samples dominate for enhanced global stability. μ controls the strength of the manifold regularization term, with a range from 10^2 to 10^{-2}. When μ is too small, it results in weak manifold constraints, making the model approach supervised learning. Conversely, when μ is too large, it enforces overly rigid geometric structure preservation.

Table 2: Dataset details

Parameter	Parameter Value
r	0.6, 0.7, 0.8
σ	10^{-2} to 10^2
μ	10^{-2} to 10^2

Tables 3, 4, and 5 present the training accuracy (expressed as mean $\pm$ standard deviation) of the L2KC and S-L2KC algorithms on the aforementioned UCI and LIBSVM datasets. In these tables, values where the S-L2KC algorithm achieves an average accuracy improvement greater than 0.6% are bolded, and values where the variance reduction exceeds 0.2% are underlined.

The experimental results demonstrate that S-L2KC generally achieves higher test accuracy and lower test variance compared to L2KC, with the improvement becoming more pronounced as the proportion of unlabeled samples increases (from 60% to 70%). However, when the proportion of unlabeled samples further increases, the performance gain of S-L2KC diminishes slightly, though it still maintains an advantage over L2KC. Additionally, the CPU time of both algorithms remains very close (less than 1ms).

Table 3: Average test accuracies and standard deviations ($r = 0.6$)

Dataset	L2KC	S-L2KC
Diabetes	72.725±2.374	**73.348±2.301**
German	70.104±1.951	70.483±1.733
Heart	82.152±4.137	82.471±4.109
Image	70.972±6.433	71.318±6.201
Ionosphere	70.011±3.730	70.425±3.703
Breast-cancer	73.384±4.342	73.377±4.293
sonar	81.949±3.551	**83.072±3.216**
Waveform	84.573±0.733	**85.541±0.701**
Splice	59.443±4.307	**60.172±4.103**
Ringnorm	97.479±0.239	97.470±0.225

Table 4: Average test accuracies and standard deviations ($r = 0.7$)

Dataset	L2KC	S-L2KC
Diabetes	72.304±2.401	**73.222±2.189**
German	69.696±1.961	**70.384±1.863**
Heart	81.183±4.132	81.506±3.891
Image	70.505±6.403	**71.248±6.199**
Ionosphere	69.991±3.730	70.375±3.582
Breast-cancer	73.408±4.331	73.414±4.248
sonar	82.079±3.527	**82.971±3.321**
Waveform	84.907±0.744	85.319±0.729
Splice	59.731±4.291	**61.231±4.031**
Ringnorm	97.239±0.241	97.235±0.243

Table 5: Average test accuracies and standard deviations ($r = 0.8$)

Dataset	L2KC	S-L2KC
Diabetes	72.337±2.351	**72.941±2.348**
German	69.005±1.922	69.283±1.917
Heart	81.037±4.107	**81.701±4.008**
Image	70.672±6.389	**71.342±6.388**
Ionosphere	69.759±3.691	70.048±3.696
Breast-cancer	73.384±4.342	73.847±4.339
sonar	81.992±3.545	**83.211±3.541**
Waveform	84.986±0.740	84.887±0.741
Splice	62.358±4.288	**63.517±4.207**
Ringnorm	97.171±0.244	97.159±0.240

4. Conclusion

This paper presents an introduction to L2KC while examining its theoretical foundations and unresolved challenges in semi-supervised learning scenarios. To enhance L2KC's capability for semi-

supervised learning, we propose a manifold regularization term based on Locality Preserving Projections (LPP) in Gram matrix form. By incorporating this regularization term into the original L2KC framework, we develop an improved algorithm termed S-L2KC, complete with theoretical analysis and predictive functions. A series of experiments are then conducted to evaluate its semi-supervised learning performance. The experimental results demonstrate that on selected UCI and LIBSVM datasets, S-L2KC achieves either superior test accuracy or comparable accuracy with reduced variance compared to L2KC, indicating enhanced generalization capability. Detailed analysis reveals that the performance improvement becomes more significant when the proportion of unlabeled samples increases from 60% to 70%. While further increasing the unlabeled sample ratio to 80% leads to slightly diminished effects, S-L2KC maintains consistent advantages in both accuracy and variance metrics. These findings confirm that S-L2KC effectively leverages information from unlabeled samples to improve classification performance and generalization ability in semi-supervised learning scenarios when compared to the original L2KC approach.

However, the performance of S-L2KC is highly dependent on the adaptability of the selected kernel function. If the kernel function fails to effectively capture the underlying structure of the data, the model's performance may degrade significantly. Additionally, in scenarios involving high noise or sparse data, the manifold assumption may not hold.

References

Kristin P. Bennett and Ayhan Demiriz. Semi-supervised support vector machines. In *Advances in Neural Information Processing Systems*, pages 368–374, 1998. doi: 10.1007/978-0-387-77501-2_7.

David Berthelot, Nicholas Carlini, and et al. Mixmatch: A holistic approach to semi-supervised learning. *arXiv preprint arXiv:1905.02249*, 2019. doi: 10.48550/arXiv.1905.02249.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. doi: 10.1023/A:1022627411411.

Xiaofei He and Partha Niyogi. Locality preserving projections. *Neural Information Processing Systems*, pages 153–160, 2005. doi: 10.1007/978-3-319-73830-7_13.

Zijian Hu and et al. Simple: Similar pseudo label exploitation for semi-supervised classification. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10428–10437, 2021. doi: 10.48550/arXiv.2103.16725.

Joo Seuk Kim and C. Scott. l_2 kernel classification. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 32(10):1822–1831, 2010. doi: 10.1109/TPAMI.2009.188.

Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*, 2021. doi: 10.48550/arXiv.2101.06329.

Kihyuk Sohn and et al. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020. doi: 10.48550/arXiv.2001.07685.

X. Wang and et al. Consistent-teacher: Towards reducing inconsistent pseudo-targets in semi-supervised object detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3240–3249, 2023. doi: 10.48550/arXiv.2209.01589.

Yidong Wang and et al. Freematch: Self-adaptive thresholding for semi-supervised learning. *arXiv preprint arXiv:2205.07246*, 2022. doi: 10.48550/arXiv.2205.07246.

Haoyu Xie and et al. Prcl: Probabilistic representation contrastive learning for semi-supervised semantic segmentation. *International Journal of Computer Vision*, 132(10):2495–2514, 2024. doi: 10.1007/s11263-024-02016-8.

L. Yang, L. Qi, L. Feng, and et al. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7236–7246, 2023. doi: 10.48550/arXiv.2208.09910.

Bowen Zhang and et al. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *arXiv preprint arXiv:2110.08263*, 2021. doi: 10.48550/arXiv.2110.08263.