Machine Learning Course Project

SKiP : SVM with K–nearest neighbor and Probabilistic weighting

Jeonghoon Park, Kangjun Lee, Jaemin Kim, Doyeol Oh

Department of Computer Science and Engineering, UNIST

Link for Code, Full paper

Introduction

Existing Support Vector Machines (SVMs) are highly vulnerable to decision boundary distortion when encountering outliers, as they penalize all training samples equally.
To address this challenge, we propose SKiP, a novel additive weighting approach that combines global class probability and local K–Nearest–Neighbor consistency to effectively mitigate both types of outliers and improve overall model robustness.

Method

Noise definition

We categorize noise into two distinct types.

- **Feature Outliers** : Correctly labeled samples located far from the main data distribution.
- **Label Outliers** : Mislabeled samples near the decision boundary.

Prob–SVM

Motivated by the assumption that the data follows a Gaussian distribution, we modify the SVM objective by introducing a probabilistic term p_i that assigns lower weights to samples far from the main distribution, reducing the impact of feature outliers.

$$\min_{w,b} \frac{1}{2}\|w\|^2 + \sum_i C\, p_i\, \xi_i. \quad p(y_i|x_i) \sim \mathcal{N}(x_i \mid \mu_{y_i}, \Sigma_{y_i}).$$

KNN–SVM

To address **label outliers**, we incorporate a KNN–based consistency term n_i into the SVM objective, assigning lower weights to samples that disagree with their neighborhood, reducing sensitivity to mislabeled data.

$$\min_{w,b} \frac{1}{2}\|w\|^2 + \sum_i C\, n_i\, \xi_i. \quad n_i = \frac{1}{K} \sum_{x_k \in \mathcal{N}_K(x_i)} \mathbb{I}(y_k = y_i),$$

SKiP

To jointly handle both types of outliers, SKiP integrates these two complementary measures. By additively combining the weights p_i and n_i, it achieves stable robustness against both feature and label corruption.

$$\min_{w,b} \frac{1}{2}\|w\|^2 + \sum_i C\,(\frac{p_i + n_i}{2})\xi_i$$

Result

We conducted experiments by injecting varying levels of label and feature noise into three datasets: Iris, Wine, and Titanic. Experimental results demonstrate that SKiP achieves the best performance across nearly all outlier configurations on Iris and Wine, which are Gaussian–like datasets. Furthermore, despite the Titanic dataset deviating substantially from a Gaussian distribution, SKiP still exhibited stable and competitive performance.

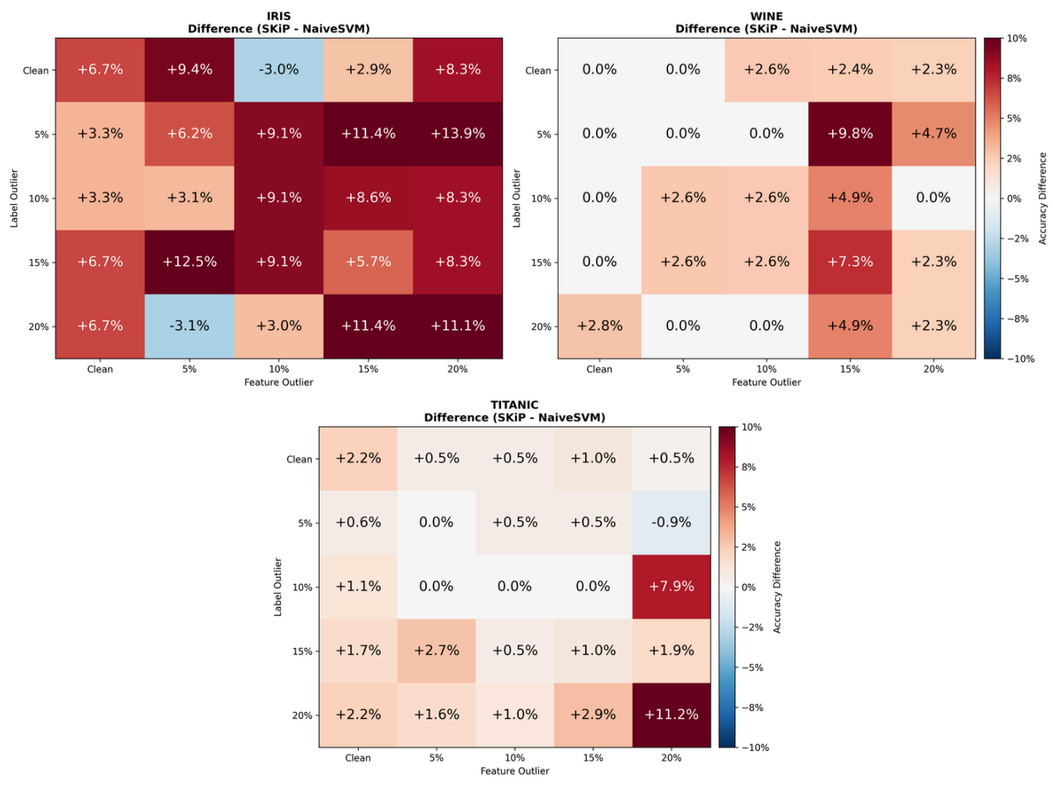

Figure 1. Accuracy difference between SKiP, NaiveSVM on Iris, Wine, Titanic dataset.

Table 1: Classification accuracy under label and feature outliers on Iris and Wine datasets.

Dataset	Outlier (%)		Accuracy (%)			
	Label	Feature	Naive-SVM	Prob-SVM	KNN-SVM	SKiP
Iris	0	0	86.7	86.7	90.0	**93.3**
		10	**87.9**	75.8	81.8	84.8
		20	80.6	86.1	**91.7**	88.9
	10	0	80.0	80.0	**83.3**	**83.3**
		10	78.8	69.7	81.8	**87.9**
		20	75.0	66.7	81.8	**83.3**
	20	0	63.3	**70.0**	**70.0**	**70.0**
		10	69.7	66.7	72.7	**75.0**
		20	63.9	66.7	72.7	**75.0**
Wine	0	0	**100.0**	**100.0**	**100.0**	**100.0**
		10	**97.4**	89.7	94.9	**97.4**
		20	**93.0**	88.4	90.7	**93.0**
	10	0	83.3	83.3	**86.1**	83.3
		10	92.3	84.6	86.0	**94.9**
		20	**93.0**	86.0	89.7	**93.0**
	20	0	83.3	80.6	**86.1**	**86.1**
		10	**79.5**	70.5	74.4	**79.5**
		20	**74.4**	72.1	72.1	**74.4**

Table 2: Classification accuracy under label and feature outliers on Titanic datasets.

Dataset	Outlier (%)		Accuracy (%)			
	Label	Feature	Naive-SVM	Prob-SVM	KNN-SVM	SKiP
Titanic	0	0	77.7	77.7	79.3	**79.9**
		10	74.0	73.5	**74.5**	**74.5**
		20	72.9	72.0	72.9	**73.4**
	10	0	72.1	72.1	72.6	**73.2**
		10	**68.4**	63.8	**68.4**	**68.4**
		20	59.8	57.5	**67.8**	**67.8**
	20	0	65.9	65.9	**69.3**	68.2
		10	63.8	60.7	**65.8**	64.8
		20	55.1	55.1	**66.4**	**66.4**