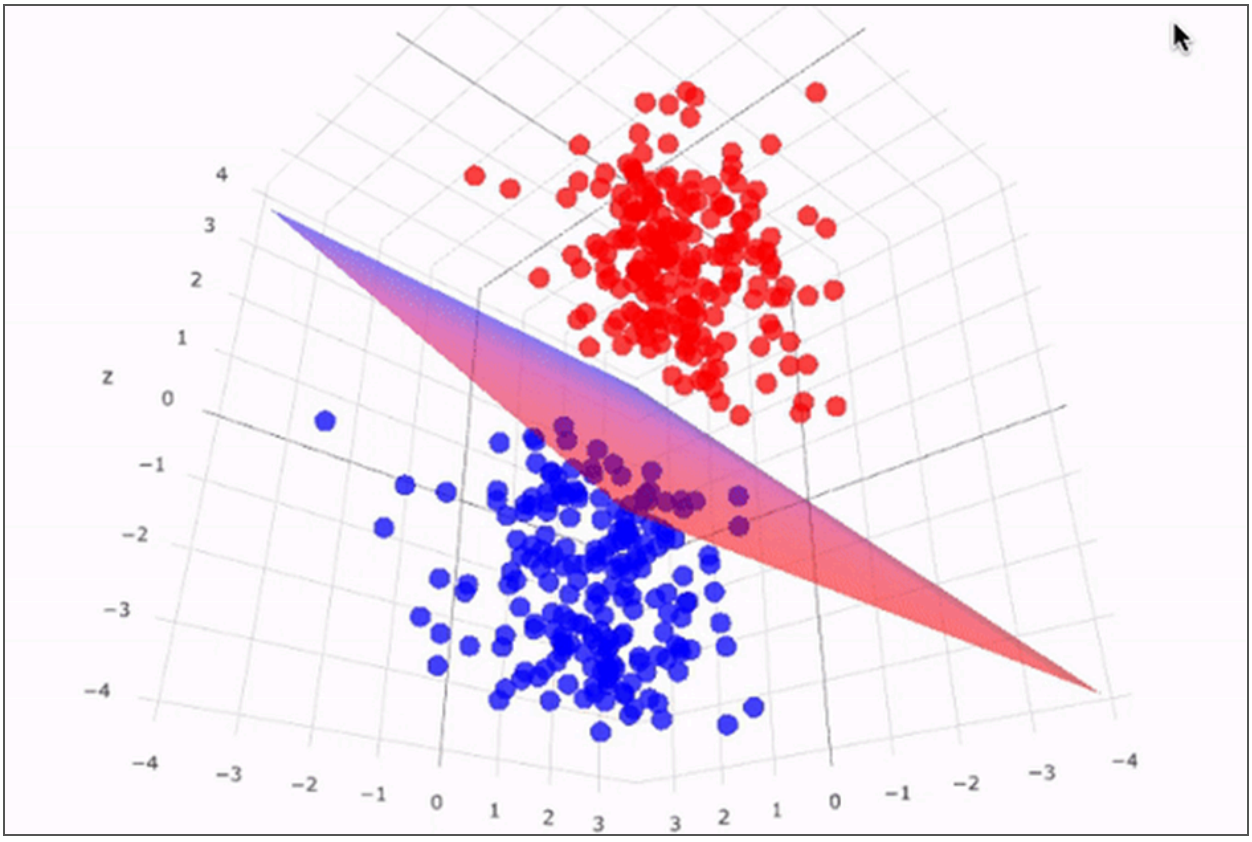

Density-Weighted Support Vector Machines for Binary Class Imbalance Learning

March 26, 2024

Priyansh Jaseja - 200001063
Kirtika Zanzan - 200002085
Atharva Mohite - 200003016
Tanishq Selot - 200003076
Bhavya Gupta - 200004008

Introduction

In real-world binary classification problems, the entirety of samples belonging to each class varies. These problems where the majority class is notably more significant than the minority class can be called a **Class Imbalance Learning (CIL)** problem. Due to the CIL problem, model performance degrades heavily. The data is not always correctly separable, which results in classification errors. Moreover, the problems that make data processing more difficult occur due to uncertain and **imbalanced datasets**. Uncertainty is ubiquitous in almost all datasets, and it is because of data collection, partial knowledge, approximate modelling, and uncertainty about the future.

In classification problems such as big data classification, fault detection, disease diagnosis, and fraud detection, the unwanted effects of imbalanced class sizes should be eliminated as they may cause classification bias. Also, while dealing with imbalanced classification issues, the learned classification boundary of the SVM tends toward the majority class due to the dominant influence of the majority class on learning, thus reducing the generalization performance of the classifier.

Problem statement

To propose a novel SVM model that guarantees that the learned classification boundary of the SVM does not tend toward the majority class, i.e., the class with a large number of samples, causing classification bias. Also, SVM has a high computational cost as it has to solve the quadratic programming problem (QPP) to find the optimal hyperplane. Hence, an SVM model to optimize the computational cost and time complexity is required. We also aim to study the different methods of assigning weights to the minority class and its subsequent effect on classification results. These findings aim to improve the following factor(s):

1. Versatility - Improved performance on imbalanced data will highlight the adaptability of SVM with application in more real-world cases. This enhancement makes the SVM suitable for a wider variety of data distributions.
2. Computation - Fast computation makes it even more adaptable for industrial applications.

Proposal

I. Implementation of the proposed algorithm:

We plan to try various ways of overcoming the class imbalance problem along with multiple versions of SVM mentioned in Hazarika et al.[1] and compare the results with the proposed models - DSVM-CIL and IDLSSVM-CIL. A few ways that involve tweaking the data distribution to handle CIL are discussed below.

II. Oversampling

A standard way is to oversample the minority class. This method is based on changing the data distribution to make the minority class equally influential in deciding the classification boundary of SVM. A few notable ways include:

a. SMOTE[2] (Synthetic Minority Oversampling Technique)

As the name suggests, it involves synthesizing new examples from existing minority class examples. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together.

b. Other versions of SMOTE

We plan to try other versions of SMOTE like Borderline SMOTE, KMeans SMOTE, ADASYN[3] (Adaptive Synthetic Sampling) etc.

III. Undersampling

We also plan to use some undersampling techniques to reduce the majority class size, like Near-Miss[4] , Condensed nearest neighbor rule and TomekLinks.

We plan to investigate other SVM versions like Least squares SVM, Fuzzy SVM, Improved 2-norm-based fuzzy least squares SVM, Entropy-based fuzzy least squares SVM and Affinity and class probability-based fuzzy SVM in a **comprehensive comparative study**.

Methodology

We plan to implement a new SVM model for the class imbalance learning (CIL) problem in the binary classification problem - Density-weighted SVM for binary class imbalance learning (DSVM-CIL). We also plan to implement an improved density-weighted least squares SVM for binary class imbalance learning (IDLSSVM-CIL) to counter the issue of slow training speed. The two SVM models were proposed by Hazarika et. al.

1. DSVM-CIL

   This variant involves multiplying the density weightage obtained from the KNN distance technique used in Cha et. al.[5] directly to the slack variable. The primal problem can hence be given as:

$$\min \frac{1}{2}\|w\|^2 + Cd\psi,$$
$$\text{s.t.,} \quad Y(\phi(X)w + eb) \geq e - \psi,\ \psi \geq 0.$$

   $\phi(x)^t w + eb = 0$ is the hyperplane

   Here, e and Ψ are vector of ones and slack variables. C > 0 is the trade-off parameter. d is the density weight derived using KNN. The unknown parameters w and b can be found by converting the above objective function to a dual problem using the Karush-Kuhn-Tucker condition.

2. IDLSSVM-CIL

   To reduce the computational cost of the proposed DSVM-CIL, we propose a least squares version of DSVM-CIL as IDLSSVM-CIL. The proposed IDLSSVM-CIL searches for a classifying hyperplane:

$$K(x^t, X^t)w + eb = 0$$

   Here, K is a nonlinear Kernel function.

   which can be obtained by solving the primal problem:

$$\min \frac{1}{2}(\|w\|^2 + b^2) + \frac{C}{2}(D\psi)^t(D\psi),$$
$$\text{s.t.,} \quad Y(\phi(X)w + eb) = e - \psi,$$

where D is a diagonal matrix obtained from affinity and class probabilities discussed in Tao et. al.[6]

It simply solves the system of linear equations by using the equality constraints and considering the 2-norm squared of the slack vector. Because of this, IDLSSVM-CIL is computationally faster compared to DSVM-CIL.

References

[1] Hazarika, B.B., Gupta, D. Density-weighted support vector machines for binary class imbalance learning. *Neural Comput & Applic* 33, 4243–4261 (2021). https://doi.org/10.1007/s00521-020-05240-8

[2] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

[3] Haibo He, Yang Bai, E. A. Garcia and Shutao Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, Hong Kong, 2008, pp. 1322-1328, doi: 10.1109/IJCNN.2008.4633969.

[4] Akira Tanimoto, So Yamada, Takashi Takenouchi, Masashi Sugiyama, Hisashi Kashima, Improving imbalanced classification using near-miss instances, Expert Systems with Applications, Volume 201, 2022, 117130, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2022.117130.

[5] Myungraee Cha, Jun Seok Kim, Jun-Geol Baek, Density weighted support vector data description, Expert Systems with Applications, Volume 41, Issue 7, 2014, Pages 3343-3350, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2013.11.025.

[6] Xinmin Tao, Qing Li, Chao Ren, Wenjie Guo, Qing He, Rui Liu, Junrong Zou, Affinity and class probability-based fuzzy support vector machine for imbalanced data sets, Neural Networks, Volume 122, 2020, Pages 289-307, ISSN 0893-6080, https://doi.org/10.1016/j.neunet.2019.10.016.