Pairwise Learning for Imbalanced Data Classification

1st Shu Liu
Computational and Data Science PhD Program
Middle Tennessee State University
Murfreesboro, TN 37132, USA
sl6b@mtmail.mtsu.edu

2nd Qiang Wu
Department of Mathematical Sciences
Middle Tennessee State University
Murfreesboro, TN 37132, USA
qwu@mtsu.edu

***Abstract*—Imbalanced data classification problems appear quite commonly in real-world applications and impose great challenges to traditional classification approaches which work well only on balanced data but usually perform poorly on the minority class when the data is imbalanced. Resampling preprocessing by oversampling the minority class or downsampling the majority class helps improve the performance but may suffer from overfitting or loss of information. In this paper we propose a novel method called pairwise robust support vector machine (PRSVM) to overcome the difficulty of imbalanced data classification. It adapts the non-convex robust support vector classification loss to the pairwise learning setting. In the training process, samples from the minority class and the majority class always appear as pairs. This automatically balances the impact of two classes. Simulations and real-world applications show that PRSVM is highly effective.**

***Index Terms*—pairwise robust support vector machine, imbalanced data, RSVC loss, pairwise learning**

I. Introduction

Imbalanced data appear ubiquitously in real-word applications. For instance, in computer-aided medical diagnosis patients with certain concerned disease such as cancer or diabetes count only a very small fraction of the screening population [1]–[4]. For spam detection system, the number of spam messages is usually far less than useful ones [5]. Intrusion detection requires detecting malicious and unauthorized activities that attack computer systems. Although rapid increase of such attacks has been seen along the popularization of computers, it is believed they are still "outliers" compared to normal activities [6].

Classification of imbalanced data is a challenging task. Approaches that are commonly used and effective on balanced data, such as support vector machines and logistic regression, could perform poorly on imbalanced data. One could get small total classification error as long as they classify data of the majority class correctly. The accuracy of classification of the minority class is usually low though more often it is the concerned class. The reason could be the absolute scarcity of minority samples [7], inappropriate evaluation method [8], the high degree of overlap of minority with the majority [9].

Researchers have been paying increasing attentions to the imbalanced data classification in the past two decades. The most common approaches are to preprocess the data via resampling techniques before training. The oversampling approach resamples or duplicates the minority class so that its sample size increases to be comparable with majority class [10]–[13] while the undersampling approach downsamples the majority class to have size comparable with the minority one [10], [11], [14], [15]. Training data becomes balanced after preprocessing and therefore traditional classification methods such as support vector machines and AdaBoost can be efficiently used. However, these preprocessing tricks have some known drawbacks: the oversampling method may increase the likelihood of overfitting due to duplicated data. On the contrary, the undersampling method may cause loss of information because some useful data present in the majority class might be eliminated. Moreover, when the minority class is too small, undersampling method will generate an undersized training set, which may decrease performance of classifiers. Menardi and Torelli [16] proposed random over-sampling examples (ROSE) by a smoothed bootstrap-based technique. It simultaneously oversamples the minority class while undersamples the majority class and more often show slightly better performance by reducing information loss and overfitting. In addition to these preprocessing tricks, researchers had also considered to modify classification algorithms or data characterizations to handle imbalanced data. For instance, [17] proposed to adjust the error costs of different classes in the training of support vector machines, which places a large error cost on minority class and a small error cost on majority class to balance the impact of skewed sample sizes. Later [18] proposed to combine the different error costs idea from [17] with synthetic minority oversampling technique from [19]. In [20] appropriate feature selection was combined with naive Bayes to handle imbalanced text data.

In this paper we will propose a new imbalanced data classification approach in the pairwise learning framework. Pairwise learning arises naturally from the metric learning, ranking, and information theoretic learning. When applied to binary classification problems, observations from the minority class and the majority class always appear as pairs for the loss evaluation. This automatically balances the impact of the two classes regardless of their sizes. It avoids the drawbacks of resampling techniques and therefore is potentially superior.

This work is partially supported by NSF (DMS-2110826).

978-1-6654-5841-2/21/$31.00 ©2021 IEEE
DOI 10.1109/CSCI54926.2021.00102

II. Pairwise robust support vector machine

Given a data of n observations $(x_i, y_i), i = 1, \ldots, n$, with $x_i \in \mathbb{R}^p$ and $y_i \in \{-1, 1\}$, the classification error of a real-valued classifier $f : \mathbb{R}^p \to \mathbb{R}$ is defined by the 0-1 loss:

$$\sum_{i=1}^{n} L_{\text{0-1}}(y_i f(x_i)) = \sum_{i=1}^{n} \mathbb{I}_{\{y_i f(x_i) < 0\}},$$

where $\mathbb{I}_{(\cdot)}$ is the indicator function taking values 1 when the condition is true and 0 otherwise. The optimization of classification error is known to be NP hard due to discontinuity of the 0-1 loss. The success of large margin classifiers such as support vector machines and AdaBoost lies on the use of surrogate loss. In support vector machines, hinge loss is used:

$$L_{\text{hinge}}(yf(x)) = (1 - yf(x))_+ = \max(0, 1 - yf(x)).$$

Support vector machines were extensively studied and successfully used in numerous fields; see e.g. [21]–[23] and references therein. Though, support vector machines could become less robust as outliers present. Among various robustification efforts, Feng et al [24] proposed robust support vector classifier (RSVC), which uses the smooth non-convex surrogate loss

$$L_{\text{RSVC}}(yf(x)) = \sigma^2 \left(1 - \exp\left(\frac{(1 - yf(x))_+^2}{\sigma^2}\right)\right)$$

to replace the hingle loss and handles outliers well, where σ is a tunable parameter. See Fig. 1 for a comparison of the three loss functions.

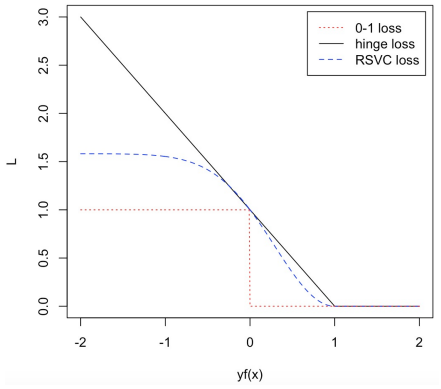

Fig. 1: 0-1 loss, hinge loss, and RSVC loss with $\sigma = 1$ (scaled so that all three losses have value 1 at $yf(x) = 0$ for comparison purpose.)

Our approach to handle imbalanced data classification is motivated by using RSVC loss in pairwise learning framework. A good real-valued classifier f should produce $\text{sign}(f(x_i)) = y_i$ as many as possible. Consequently, for each pair of observations (x_i, y_i) and (x_j, y_j), if $y_i = 1$ and $y_j = -1$, we would expect $f(x_i) - f(x_j) > 0$; if $y_i = -1$ but $y_j = 1$, we expect $f(x_i) - f(x_j) < 0$; when $y_i = y_j$ we have no expectations. Using the notation

$$y_{ij} = \begin{cases} 1, & \text{if } y_i = 1 \text{ and } y_j = -1; \\ -1, & \text{if } y_i = -1 \text{ and } y_j = 1, \end{cases}$$

we should have $y_{ij}(f(x_i) - f(x_j)) > 0$. We adapt the RSVC loss and define the pairwise loss as

$$L(f, (x_i, y_i), (x_j, y_j)) = \sigma^2 \left(1 - \exp\left(\frac{(1 - y_{ij}(f(x_i) - f(x_j)))_+^2}{\sigma^2}\right)\right).$$

In this paper we focus on linear classifiers $f(x) = w^\top x + b$ with $w \in \mathbb{R}^p$ and $b \in \mathbb{R}$. The optimization problem associated to the pairwise robust support vector classifier is

$$\min_w \sum_{i=1}^{n} \sum_{j: y_j \neq y_i} \sigma^2 \left(1 - \exp\left(\frac{(1 - y_{ij} w^\top (x_i - x_j))_+^2}{\sigma^2}\right)\right). \tag{1}$$

This can be solved by either the gradient descent algorithm or iterative least square method.

Note that the intercept b does not appear in (1) because it is canceled in calculation of the difference $f(x_i) - f(x_j) = w^\top (x_i - x_j)$. We need to figure out appropriate intercept b separately. This is done as follows: let $\hat{w}$ be the solution to the minimization problem (1). For $b \in \mathbb{R}$, let $\mathcal{E}_+(b)$ and $\mathcal{E}_-(b)$ denote the false positive rate (FPR) and false negative rate (FNR) of the classifier $\hat{w}^\top x + b$, respectively. Define

$$\mathcal{E}(b) = \max(\mathcal{E}_+(b), \mathcal{E}_-(b))$$

and the intercept is estimated by

$$\hat{b} = \min_{b \in \mathbb{R}} \mathcal{E}(b). \tag{2}$$

We call our method pairwise robust support vector machine (PRSVM). It has several advantages. First, the impact of the positive class and negative class is automatically balanced without need of resampling. This avoids overfitting or loss of information suffered by resampling techniques. Second, it inherits the robustness of the RSVC loss.

III. Simulations and applications

In this section, simulation studies and real-world applications are used to illustrate the effectiveness of PRSVM to handle imbalanced data. We will not only compare it with three traditional classification methods: support vector machine (SVM), AdaBoost, and Naive Bayes (NB), but also compare with the three resampling preprocessing techniques: Undersampling (US), Oversampling (OS), and Random Over-Sampling Examples (ROSE).

PRSVM is not very sensitive to the choice of the parameter σ. A moderate choice usually give sufficiently good results. We selected $\sigma = 1$ in all experiments. All other methods are implemented in R with standard packages. In particular, for SVM we used the `train` function in `caret` package. For fair comparison, method `svmlinear` is used to produce linear classifiers. Other parameters are kept default. AdaBoost used the function `boosting` from `adabag` package. The number of iterations is set as 200. Naive Bayes used the function `naiveBayes` from `e1071` package with default parameters. The implementation of three resampling methods used the `ROSE` package. For undersampling and oversampling

methods, the resampling sample size is set as twice of the amount of minority class or majority class, respectively, so that the resampling process stops when two classes have the same size. For ROSE all parameters are set as default. SVM linear classifier is used to classify data after the data is balanced.

For imbalanced data classification, the overall accuracy does not make much sense, especially when the majority class dominates. We will look at the FPR and FNR simultaneously. A balanced FPR and FNR implies the minority class has been equally addressed. We also evaluate the area under the ROC curve (AUC), which is widely accepted as a balanced accuracy metric for imbalanced data classification problems; see e.g. [25]. The larger AUC, the better.

A. Simulation studies

We assume the data come from $\mathbb{R}^2$. For the positive class, $x_i = (x_{i,1}, x_{i,2})$ has both features $x_{i,1}$ and $x_{i,2}$ sampled from normal distribution with mean 3 and variance 1. For the negative class, both features are sampled from normal distribution with mean 5 and variance 1. The positive class will be the minority class. In the experiment, we will first set an imbalance ratio, that is, the ratio between sample sizes of the positive class and the negative class, say 1:k with some $k > 1$. With training sample size n, we will generate $\frac{n}{k+1}$ samples for the positive class and $\frac{kn}{k+1}$ samples for the negative class. The performance of each classifier will then be evaluated on a test set which contains 1000 samples from the positive class and $1000k$ samples from the negative class.

To compare the performance of seven methods, we fix the number of training samples $n = 200$ and investigate the impact of imbalance ratio by varying it from 1:2 to 1:4 and then to 1:8. Each experiment is repeated 50 times. The average FPR, FNR and AUC for the seven classifiers are reported in Table I. The results indicate that SVM, AdaBoost and Naive Bayes deteriorate fast as imbalance ratio increases. FNR drops at the price of increasing FPR and AUC decreases. The three resampling techniques and PRSVM still give balanced FPR and FNR and AUC drop is not significant. In all scenarios PRSVM achieves the largest AUC.

B. Application I: Diabetes Data

Diabetes Data [4] (https://datahub.io/machine-learning/diabetes) is a multivariate data set with 8 attributes measuring patients' medical conditions. The response variable denotes whether a patient is tested positive $(+1)$ or negative (-1) for diabetes. The data contains 268 positive cases and 500 negative cases, leading to an imbalance ratio close to 1:2.

We randomly sample $1/3$ of positive cases and 1/3 of negative cases to form a training set so that the the original imbalance ratio is maintained in the training process. The remaining cases are used to test the performance of the seven classifiers. The experiments are repeated 50 times and the average FPR, FNR and AUC are reported in Table II. Similar to our simulation studies, SVM, AdaBoost and Naive Bayes have the poorest performance. The three resampling methods perform better. Although PRSVM has the total error (FPR+FNR) slightly larger than resampling methods, it has the best balanced FPR and FNR and highest AUC.

TABLE I: Classification performance of seven classifiers on simulated data with $n = 200$ and varying imbalance ratios

Imbalance Ratio	Method	FPR	FNR	AUC
1:2	SVM	0.0469	0.1328	0.9102
	ADABOOST	0.0724	0.1602	0.8837
	NB	0.0486	0.1266	0.9123
	US	0.0802	0.0845	0.9177
	OS	0.0824	0.0857	0.9159
	ROSE	0.0845	0.0858	0.9148
	PRSVM	0.0814	0.0844	0.9760
1:4	SVM	0.0272	0.1910	0.8909
	ADABOOST	0.0460	0.2177	0.8682
	NB	0.0284	0.1840	0.8938
	US	0.0867	0.0814	0.9159
	OS	0.0754	0.0897	0.9174
	ROSE	0.0848	0.0801	0.9176
	PRSVM	0.0765	0.0907	0.9766
1:8	SVM	0.0150	0.2722	0.8564
	ADABOOST	0.0287	0.3044	0.8334
	NB	0.0160	0.2563	0.8638
	US	0.0800	0.0938	0.9131
	OS	0.0735	0.0961	0.9152
	ROSE	0.0856	0.0841	0.9152
	PRSVM	0.0750	0.0976	0.9763

TABLE II: Classification performance on Diabetes Data

Method	FPR	FNR	AUC
SVM	0.1217	0.4444	0.7170
ADABOOST	0.1862	0.4228	0.6955
NB	0.1681	0.4124	0.7098
US	0.2339	0.2950	0.7356
OS	0.2257	0.3045	0.7349
ROSE	0.2422	0.2994	0.7292
PRSVM	0.2710	0.2863	0.7965

C. Application II: Wilt Data

Wilt Data Set [26] is a high-resolution remote sensing data. It consists of image segments generated by segmenting pan-sharpened images and contain spectral information from the Quickbird multispectral image bands and texture information from the panchromatic image band and was used to detect diseased trees. There are five features: GLCM-Pan, Mean-G, Mean-R, Mean-NIR and SD-Pan, meaning GLCM mean texture (Pan band), mean green value, mean red value, mean NIR value and standard deviation (Pan band), respectively. The response variable has two states, "diseased trees" or "other land cover".

On the UCI Machine Learning Repository webpage for Wilt Data (https://archive.ics.uci.edu/ml/datasets/wilt), we downloaded a training set of 4339 samples and a test set of 500 samples. They are combined together to form a single data set. The combined data set contains 261 samples labeled as "diseased trees" class and 4578 labeled as "other land cover" class. The imbalance ratio is high and exceeding 1:17.

In our study, we refer to the "diseased trees" as the positive class and "other land cover" as negative class. We randomly choose 1/3 of each class to form the training set and leave the other 2/3 as test set. This process is repeated

50 times. The mean performance metrics of seven classifiers are reported in Table III. SVM and Naive Bayes perform very poorly to predict the positive class due to the high imbalance ratio. Surprisingly AdaBoost works fine and even outperforms ROSE. All three resampling methods and PRSVM are able to overcome the highly imbalance problem and provide reasonable prediction for both classes. Oversampling gives the smallest FPR and FNR while PRSVM has the largest AUC.

TABLE III: Classification performance on Wilt Data

Method	FPR	FNR	AUC
SVM	0.0027	0.8881	0.5546
ADABOOST	0.0048	0.1478	0.8100
NB	0.0915	0.4538	0.7274
US	0.0939	0.0411	0.9325
OS	0.0697	0.0489	0.9407
ROSE	0.2749	0.1193	0.8029
PRSVM	0.0844	0.0567	0.9713

IV. Conclusions and future works

In this paper we proposed a new approach, the pairwise robust support vector machine, for imbalanced data classification. It automatically balances the impact of two imbalanced classes by pairing positive examples with negative ones. Simulations and real-world applications show that our new approach can effectively reduce the prediction error for the minority class. Compared with resampling techniques in the literature, it always has the largest AUC.

There are several open problems remaining for future research. First, we have focused on linear classifier in this paper. It is necessary to develop nonlinear PRSVM classifiers using kernel tricks or deep neural networks for nonlinear data classification which is also common in real applications. Second, we found through simulations that PRSVM is not very sensitive to the parameter σ and a moderate choice usually leads to sufficiently good results. We had fixed $\sigma = 1$ in all our experiments without tuning it for the best results. It is worth developing tuning strategies that helps improve the performance. Third, PRSVM shows to have largest AUC in all our experiments. But its prediction errors, namely FPR and FNR, are not the smallest. Note that AUC is independent of intercept while prediction errors do. A plausible conjecture is our estimation of the intercept is not optimal and a refined approach is expected. Finally, it would be interesting to explore the mathematical properties of PRSVM and perform generalization analysis to theoretically explain its effectiveness.

References

[1] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," *Neural networks*, vol. 21, no. 2-3, pp. 427–436, 2008.

[2] V. Soler, J. Cerquides, J. Sabria, J. Roig, and M. Prim, "Imbalanced datasets classification by fuzzy rule extraction and genetic algorithms," in *Sixth IEEE International Conference on Data Mining-Workshops (ICDMW'06)*. IEEE, 2006, pp. 330–336.

[3] F. Ren, P. Cao, W. Li, D. Zhao, and O. Zaiane, "Ensemble based adaptive over-sampling method for imbalanced data learning in computer aided detection of microaneurysm," *Computerized Medical Imaging and Graphics*, vol. 55, pp. 54–67, 2017.

[4] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[5] T. Fawcett, ""in vivo" spam filtering: a challenge problem for kdd," *ACM SIGKDD Explorations Newsletter*, vol. 5, no. 2, pp. 140–148, 2003.

[6] M. Tavallaee, N. Stakhanova, and A. A. Ghorbani, "Toward credible evaluation of anomaly-based intrusion-detection methods," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 5, pp. 516–524, 2010.

[7] G. M. Weiss, "Mining with rarity: a unifying framework," *ACM Sigkdd Explorations Newsletter*, vol. 6, no. 1, pp. 7–19, 2004.

[8] F. Provost and T. Fawcett, "Robust classification for imprecise environments," *Machine learning*, vol. 42, no. 3, pp. 203–231, 2001.

[9] W. W. Ng, G. Zeng, J. Zhang, D. S. Yeung, and W. Pedrycz, "Dual autoencoders features for imbalance classification problem," *Pattern Recognition*, vol. 60, pp. 875–889, 2016.

[10] S. Kotsiantis, D. Kanellopoulos, and P. Pintelas, "Handling imbalanced datasets: A review," *GESTS International Transactions on Computer Science and Engineering*, vol. 30, no. 1, pp. 25–36, 2006.

[11] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 4, pp. 463–484, 2011.

[12] F. J. Castellanos, J. J. Valero-Mas, J. Calvo-Zaragoza, and J. R. Rico-Juan, "Oversampling imbalanced data in the string space," *Pattern Recognition Letters*, vol. 103, pp. 32–38, 2018.

[13] X. Wang, J. Xu, T. Zeng, and L. Jing, "Local distribution-based adaptive minority oversampling for imbalanced data classification," *Neurocomputing*, vol. 422, pp. 200–213, 2021.

[14] W.-C. Lin, C.-F. Tsai, Y.-H. Hu, and J.-S. Jhang, "Clustering-based undersampling in class-imbalanced data," *Information Sciences*, vol. 409, pp. 17–26, 2017.

[15] M. Koziarski, "Radial-based undersampling for imbalanced data classification," *Pattern Recognition*, vol. 102, p. 107262, 2020.

[16] G. Menardi and N. Torelli, "Training and assessing classification rules with imbalanced data," *Data mining and knowledge discovery*, vol. 28, no. 1, pp. 92–122, 2014.

[17] K. Veropoulos, C. Campbell, and N. Cristianini, "Controlling the sensitivity of support vector machines," in *Proceedings of the international joint conference on AI*, 1999, pp. 55–60.

[18] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *European conference on machine learning*. Springer, 2004, pp. 39–50.

[19] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in *Proceedings of the ninth ACM international conference on Multimedia*, 2001, pp. 107–118.

[20] M. Grobelnik, "Feature selection for unbalanced class distribution and naive bayes," in *ICML '99: Proceedings of the Sixteenth International Conference on Machine Learning*. Citeseer, 1999, pp. 258–267.

[21] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[22] I. Steinwart and A. Christmann, *Support vector machines*. Springer Science & Business Media, 2008.

[23] B. Schölkopf, A. J. Smola, F. Bach *et al.*, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[24] Y. Feng, Y. Yang, X. Huang, S. Mehrkanoon, and J. A. Suykens, "Robust support vector machines for classification with nonconvex and smooth losses," *Neural computation*, vol. 28, no. 6, pp. 1217–1247, 2016.

[25] J. Huang and C. X. Ling, "Using auc and accuracy in evaluating learning algorithms," *IEEE Transactions on knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, 2005.

[26] B. A. Johnson, R. Tateishi, and N. T. Hoan, "A hybrid pansharpening approach and multiscale object-based image analysis for mapping diseased pine and oak trees," *International journal of remote sensing*, vol. 34, no. 20, pp. 6969–6982, 2013.