








Oversampling Method Based Covariance Matrix Estimation in High-Dimensional Imbalanced Classification

Ireimis Leguen-de-Varona¹(✉) , Julio Madera¹ , Hector Gonzalez² ,
Lise Tubex³ , and Tim Verdonck³ 

¹ Universidad de Camagüey “Ignacio Agramonte Loynaz”, Camaguey, Cuba
{ireimis.leguen,julio.madera}@reduc.edu.cu

² Universidad de las Ciencias Informaticas (UCI), Havana, La Habana, Cuba
hglez@uci.cu

³ University of Antwerp, Antwerp, Belgium
{lise.tubex,tim.verdonck}@uantwerpen.be

Abstract. Class imbalance is a common problem in (binary) classification problems. It appears in many application domains, such as text classification, fraud detection, churn prediction and medical diagnosis. A widely used approach to cope with this problem at the data level is the Synthetic Minority Oversampling Technique (SMOTE) which uses the K-Nearest Neighbors (KNN) algorithm to generate new, artificial instances in the minority class. It is however known that SMOTE is not ideal for high-dimensional data. Therefore, we propose an alternative oversampling strategy for imbalanced classification problems in high dimensions. Our approach is based on the sparse inverse covariance matrix estimated through the Ledoit-Wolf method for high-dimensional data. The results show that our proposal has a competitive performance with respect to popular competitors.

Keywords: Imbalanced classes · Oversampling · High-dimensional data · Sparse Covariance Matrix Estimation · Ledoit-Wolf estimator

1 Introduction

Oversampling is a popular solution to the problem of imbalanced data, where the number of instances in one class, i.e. the minority class, is significantly lower than the number of instances in the other class. In oversampling strategies, the minority class is artificially increased by creating new synthetic data based on the existing sample [21, 23].

A very popular and widely used oversampling method is SMOTE (Synthetic Minority Oversampling Technique) [2, 5]. Various alternatives have been proposed in literature to achieve a good balance such as: SMOTE-Borderline [6], SMOTE-RSB* [18], ADASYN [7], ROS [22] and SMOTE-COV [13].

The sample covariance matrix is a commonly used estimator of the true covariance matrix, and is calculated directly from the data by taking the average of the outer product of the centered data matrix. However, the sample covariance matrix can be unreliable when the number of variables is large compared to the number of observations, as it can be noisy and have unstable eigenvalues.

Covariance matrix estimation methods with sparsity and shrinkage estimation like the Ledoit-Wolf estimator improved the accuracy and stability of the estimated covariance matrix [10–12]. These methods typically involve some form of regularization or shrinkage, which involves adding a bias to the estimator to reduce its variance [3, 14]. In addition to the Ledoit-Wolf shrinkage estimator, there are several other methods that have been proposed for estimating the covariance matrix in high-dimensional settings. Some of these methods include Graphical Lasso with ℓ_1 regularization, Sparse PCA, Random matrix theory and Bayesian methods [12].

In high-dimensional imbalanced data, oversampling can be particularly challenging to generate meaningful synthetic samples. One approach to overcome this challenge is to simulate synthetic data by using the sparse covariance matrix while oversampling and improve the classifiers. It is known that the behavior of SMOTE in high-dimensional data is not always ideal. (i) Oversampling can lead to an increase in the number of redundant or irrelevant features, which can reduce the performance of the classifier. This is because synthetic samples generated by SMOTE are based on existing features, and can therefore inherit the same irrelevant or noisy features [9]. (ii) SMOTE can lead to overfitting. This is because SMOTE generates synthetic samples by interpolating between existing samples, which can lead to over-representation of certain regions of the feature space. In high-dimensional data this can be biased, since the number of combinations of features grows exponentially [9]. (iii) SMOTE can be computationally expensive in high-dimensional data, as the number of possible combinations of features grows exponentially. This can make it difficult to generate a sufficient number of synthetic samples to balance the class distribution in the minority class [9]. (iv) SMOTE uses the classical Euclidean distance metric to compute the neighbors. In the high-dimensional case, it may follow that a lot of instances or all of them have the same distances. This can lead to an ineffective interpolation [1]. (v) SMOTE can experience over-generalization. Class overlap can be increased because the method ignores the majority class, allowing the creation of synthetic samples over the majority class [8, 15, 16].

Some modifications of SMOTE to balance data sets by applying feature selection or reduction before or after generating synthetic instances to obtain good results in high-dimensional classification problems have been proposed recently, see for example SDDSMOTE [17], FW-SMOTE [19] and SMOTE-SF [20].

The main contributions of this paper are: (i) We introduce a novel strategy of the resampling based on the Ledoit-Wolf covariance matrix and shrinkage selection in high-dimensional imbalanced classification. (ii) We propose an empirical evaluation of the SMOTE algorithms for imbalanced classification in small and high dimensions synthetics data sets.

2 Covariance Matrix Estimation in High-Dimensional Imbalanced Classification

Let a set of data, independent and identically distributed (i.i.d.) $X = \{X_1, \dots, X_N\}$ with $X_i \in \mathbb{R}^p$, be N samples drawn from a p -dimensional Gaussian distribution $\mathcal{N}(\mu, \Sigma)$. The task to estimate the inverse covariance matrix $\Theta = \Sigma^{-1}$ (also known as the covariance precision), solves the following loglikelihood regularized optimization problem

$$\Theta^* = \underset{\Theta \succeq 0}{\operatorname{argmin}} \{-\log \det(\Theta) + \operatorname{tr}(S\Theta) + g(\Theta)\}. \quad (1)$$

where $g(\Theta)$ is the convex and normally non-differentiable regularization function and $S, \Theta \in \mathbb{R}^{p \times p}$ are the estimated sample covariance matrix and inverse covariance matrix. The expression to compute the sample mean μ and the sample covariance matrix S are:

$$\mu = \frac{1}{N} \sum_{i=1}^N X_i, \quad S = \frac{1}{N} \sum_{i=1}^N (X_i - \mu)(X_i - \mu)^T. \quad (2)$$

In high-dimensional conditions ($p \approx N$ or $p \gg N$) the sample covariance matrix S will be singular [4], and the likelihood estimator of the covariance matrix has many weaknesses such as inaccuracy. Several studies [2, 3, 8–11], research the problem of the high dimensions using sparsity and shrinkage methods for estimating Θ considering that the number of the parameters increase quadratically with respect to the number of variables p .

In the Ledoit-Wolf class of estimators a linear combinations of the identity matrix \mathbb{I}_p and the sample covariance matrix S is considered, so that the optimization problem of the shrinkage estimation and selection became

$$\begin{aligned} & \min_{\rho_1, \rho_2} \mathbb{E} \left[\|\hat{\Theta} - \Theta\|_F^2 \right] \\ & \text{s.t. } \hat{\Theta} = \rho_1 \mathbb{I}_p + \rho_2 S. \end{aligned} \quad (3)$$

The solution to Eq. 3 can also be written as a convex linear combination

$$\hat{\Theta}^* = \lambda^* \mu_p \mathbb{I}_p + (1 - \lambda^*) S, \quad \lambda^* = \frac{\beta^2}{\gamma^2}, \quad (4)$$

where $\beta^2 = \mathbb{E} [\|S - \Theta_T\|_F^2]$ and $\gamma^2 = \|S - \mu_p \mathbb{I}_p\|_F^2$. With the asymptotic analysis we have special interest in the problem with a large number of attributes $p/N \rightarrow c \in (0, \infty]$, that is p and N have similar behaviour in infinity. The case of $p \gg N$ was not considered in this study. Finally, the covariance matrix estimator $\hat{\Theta}^*$ and the optimal shrinkage parameter λ^* can be computed in the following steps described in Algorithm 1. The synthetic data that must be generated for the minority class can be simulated using the multivariate normal distribution with $\mathcal{N}(\mu, \Sigma^*)$ in the proportion established in the configuration.

Algorithm 1. Ledoit-Wolf covariance matrix estimation**Input:** X **Output:** Θ^*, λ^* Compute μ and S (eq. 2)

$$\hat{\gamma}^2 = \|S - \mu\mathbb{I}_p\|_F^2$$

$$\hat{\beta}^2 = \min\{\hat{\gamma}^2, \frac{1}{N} \sum_{i=1}^N \|S - X_i^T X_i\|_F^2\}$$

Compute λ^* and Θ^* with eq. 3

3 Numerical Data Simulation and Empirical Study

In this empirical evaluation, we compare several classifiers (MLP, SVM, KNN) in different conditions of imbalance and number of features (Twelves synthetic datasets with $p = \{50, 200, 500, 1000, 3000, 5000\}$ and $IR = \{0.03, 0.05\}$). Also, some variants of the SMOTE oversampling strategy were combined with the classifiers to improve the accuracy. Table 1 shows in details the main characteristics of the classifiers and the resampling methods. In case of the oversampling methods studied, two resampling strategies were evaluated while the classifiers tuned several parameters to choose the best model parameter. Each pipeline was executed over five splits and three iterations of the classifier. In this primary research, it was decided to conduct an empirical study with data sets generated using the multivariate normal distribution $\mathcal{N}(0, \Sigma)$ for the binary classification problem. Two levels of the imbalance were considered, namely 3% and 5%. Also, six different feature sizes ($p = 50, 200, 500, 1000, 3000, 5000$) were used to build the numerical simulation of the moderate (first three p values) and high-dimensional (the last three) case. In all of the databases, the number of samples was $N = 1000$ ($p/N \rightarrow c = 5$ for the more high-dimensional cases) and 60% of the features were considered informative. Figure 1 shows the two principal component of a simulated database.

Table 1. The parameters considered for fine tuning of the classifiers.

	Method	Fine Tuning
Oversampling strategies	SMOTE	Resampling 50:50
	ADASYN	Resampling 50:50
	ROS	Resampling 50:50
	Cov_HD	
Classifiers	KNN	n_neighbors
	Random Forest	max_depth, n_estimators, min_samples_split, min_samples_leaf
	MLP	hidden_layer_sizes, activation, solver, learning_rate, alpha, learning_rate
	SVM	kernel, gamma, C

In our future work, five highly imbalanced and high-dimensional data sets from the GEMLeR collection with continuous attributes and a binary class will be considered for the empirical evaluation.

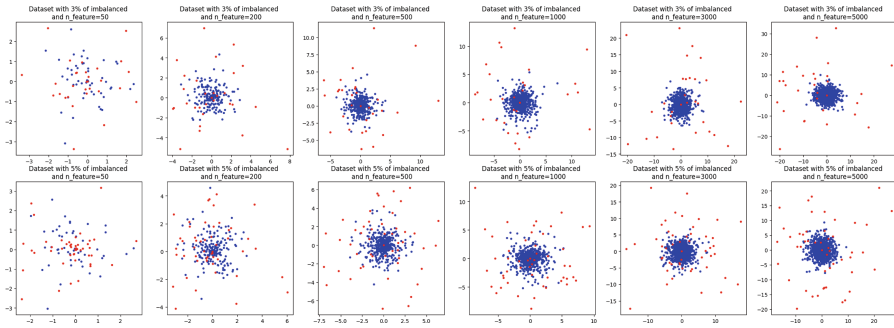


Fig. 1. A simulated database in high-dimensional imbalanced classification.

3.1 Empirical Evaluation

The AUC (Area Under the Curve) is commonly used in machine learning for evaluating the performance of binary classification models. The AUC represents the area under the Receiver Operating Characteristic (ROC) curve, which is a plot of the true positive rate (TPR) against the false positive rate (FPR) at various classification thresholds. The TPR is the fraction of positive samples that are correctly classified as positive, while the FPR is the fraction of negative samples that are incorrectly classified as positive. Similar to precedent works, the AUC is also useful for imbalanced classification problems. The AUC metric is attractive because it is insensitive to the threshold used to classify instances, and provides a single number that summarizes the overall performance of the model.

In Table 2, we show the AUC metric for each classifier and oversampling strategy with all possible combinations of data sets that we propose. The best values are indicated in bold.

The AUC metric in the combination of KNN + COV_HD shows the most difference with respect to the rest of the combinations. The Friedman test with Holm correction posthoc, for each classifier in combination with the oversampling strategies, shows only significant differences in the combination KNN + COV_HD with respect to the rest of the strategies.

Another encouraging result in the empirical evaluation is that, although there are no significant differences, in the case of SVM and MLP the Friedman rankings give the proposed approach the first place. A differentiated analysis of the Friedman test where we focus on the three bases we considered high-dimensional for our study, is shown in Table 3.

Table 2. The AUC metric in the empirical evaluation. The best values are indicated in bold.

Classifiers	Algorithms	50 features	100 features	500 features	1000 features	3000 features	5000 features
KNN 3% imbalance	COV_HD	0.895201	0.914929	0.873684	0.93615	0.936687	0.933655
	ADASYN	0.810062	0.871649	0.565015	0.60443	0.5	0.501577
	ROS	0.639319	0.743857	0.723529	0.7742	0.539938	0.752563
	SMOTE	0.873375	0.844378	0.585139	0.587025	0.501548	0.506309
KNN 5% imbalance	COV_HD	0.847678	0.962677	0.89195	0.908879	0.943653	0.932986
	ADASYN	0.5	0.5	0.5	0.5	0.5	0.5
	ROS	0.734211	0.673678	0.868885	0.766381	0.587616	0.780436
	SMOTE	0.5	0.5	0.5	0.5	0.5	0.5
Classifiers	Algorithms	50 features	100 features	500 features	1000 features	3000 features	5000 features
MLP 3% imbalance	COV_HD	0.866873	0.89315	0.868421	0.887193	0.822291	0.730087
	ADASYN	0.765325	0.921258	0.855418	0.8414	0.809288	0.761632
	ROS	0.754489	0.916977	0.829721	0.880305	0.669969	0.669558
	SMOTE	0.871207	0.88347	0.729412	0.800819	0.780186	0.707216
MLP 5% imbalance	COV_HD	0.574303	0.69099	0.627245	0.634773	0.612384	0.604617
	ADASYN	0.66161	0.670514	0.60031	0.526806	0.604334	0.495532
	ROS	0.544892	0.622673	0.65356	0.694713	0.375851	0.632725
	SMOTE	0.673375	0.587118	0.721053	0.601452	0.613003	0.520104
Classifiers	Algorithms	50 features	100 features	500 features	1000 features	3000 features	5000 features
SVM 3% imbalance	COV_HD	0.873976	0.946667	0.80901	0.925051	0.787658	0.83101
	ADASYN	0.796351	0.918283	0.872673	0.886465	0.801065	0.844747
	ROS	0.866158	0.508333	0.852197	0.89707	0.5	0.5
	SMOTE	0.863179	0.912727	0.803611	0.882525	0.785489	0.853636
SVM 5% imbalance	COV_HD	0.600521	0.758586	0.621742	0.645657	0.642219	0.5
	ADASYN	0.630119	0.661717	0.503351	0.5	0.649106	0.671818
	ROS	0.766754	0.5	0.633842	0.53626	0.5	0.514747
	SMOTE	0.627513	0.657778	0.5	0.526061	0.649106	0.671717

Table 3. The Friedman test analysis in high-dimensional data sets.

	Classifier	Ranking	Statistic	P-value
high-dimensional	KNN	COV_HD, ADASYN, SMOTE, ROS	11.45	0.009
	SVM	COV_HD, ADASYN, ROS, SMOTE	3.6	0.308
	MLP	COV_HD, ADASYN, SMOTE, ROS	1.4	0.706

4 Concluding Remarks and Further Work

In this paper we have introduced a new classification approach for high-dimensional imbalanced problems based on sparse inverse covariance estimation using the Ledoit-Wolf method. The empirical evaluation demonstrated the effectiveness of oversampling data through sparser covariance estimation compared to other state-of-the-art methods. COV_HD showed similar or comparable results considering the results of the AUC metric and the Friedman test. In addition, we found significant differences for specific classifiers. In general, we can conclude that the strategy of resampling based on sparser covariance matrix is a competitive method for high-dimensional imbalanced classification problems.

Future work is planned in several directions to expand this contribution. On the one hand, the most exhaustive evaluation of the sparser covariance matrix for high dimensions should be considered while introducing the new block-wise

covariance learning with very high efficiency. On the other hand, a real problem in high and ultra-high dimensions with imbalanced classes from several domains can be considered.

Acknowledgments. We would like to thank VLIR (Vlaamse Inter Universitaire Raad, Flemish Interuniversity Council, Belgium) for supporting this work under the project Cuban ICT NETWORK program: “Strengthening the ICT role in Cuban Universities for the development of the society”; specifically to Project 1: “Strengthening the research on ICT and its knowledge transference to the Cuban society (RESICT)” and also to the Cuban national project “Plataforma para el análisis de grandes volúmenes de datos y su aplicación a sectores estratégicos”.

References

1. Blagus, R., Lusa, L.: SMOTE for high-dimensional class-imbalanced data. *BMC Bioinform.* **14**, 106 (2013). <https://doi.org/10.1186/1471-2105-14-106>
2. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002)
3. Chen, Y., Wiesel, A., Hero, A.O.: Shrinkage estimation of high dimensional covariance matrices. In: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2937–2940. IEEE (2009)
4. Clemmensen, L., Hastie, T., Witten, D., Ersbøll, B.: Sparse discriminant analysis. *Technometrics* **53**(4), 406–413 (2011)
5. Fernández, A., García, S., Herrera, F., Chawla, N.V.: SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. *J. Artif. Intell. Res.* **61**, 863–905 (2018)
6. Han, H., Wang, W.-Y., Mao, B.-H.: Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In: Huang, D.-S., Zhang, X.-P., Huang, G.-B. (eds.) *ICIC 2005*. LNCS, vol. 3644, pp. 878–887. Springer, Heidelberg (2005). https://doi.org/10.1007/11538059_91
7. He, H., Bai, Y., García, E.A., Li, S.: ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1322–1328. IEEE (2008)
8. He, H., García, E.A.: Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **21**(9), 1263–1284 (2009)
9. Hsieh, C.-J., Sustik, M.A., Dhillon, I.S., Ravikumar, P.K., Poldrack, R.: BIG & QUIC: sparse inverse covariance estimation for a million variables. In: *Advances in Neural Information Processing Systems*, vol. 26 (2013)
10. Ledoit, O., Wolf, M.: Honey, i shrunk the sample covariance matrix. *UPF Economics and Business Working Paper* (691) (2003)
11. Ledoit, O., Wolf, M.: A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Anal.* **88**(2), 365–411 (2004)
12. Ledoit, O., Wolf, M.: The power of (non-) linear shrinking: a review and guide to covariance matrix estimation. *J. Financ. Economet.* **20**(1), 187–218 (2022)
13. Leguen-deVarona, I., Madera, J., Martínez-López, Y., Hernández-Nieto, J.C.: SMOTE-Cov: a new oversampling method based on the covariance matrix. In: Vasant, P., Litvinchev, I., Marmolejo-Saucedo, J.A., Rodríguez-Aguilar, R., Martínez-Ríos, F. (eds.) *Data Analysis and Optimization for Engineering and Computing*

- Problems. EICC, pp. 207–215. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-48149-0_15
14. Lotfi, R., Shahsavani, D., Arashi, M.: Classification in high dimension using the Ledoit-Wolf shrinkage method. *Mathematics* **10**(21), 4069 (2022)
 15. López, V., Fernández, A., García, S., Palade, V., Herrera, F.: An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. *Inf. Sci.* **250**, 113–141 (2013)
 16. Nekooimehr, I., Lai-Yuen, S.K.: Adaptive semi-supervised weighted oversampling (A-SUWO) for imbalanced datasets. *Expert Syst. Appl.* **46**, 405–416 (2016)
 17. Li, M., Wan, Q., Deng, X., Yang, H.: Synthetic minority oversampling technique based on sample density distribution for enhanced classification on imbalanced microarray data. In: ICCDA (2022)
 18. Ramentol, E., Caballero, Y., Bello, R., Herrera, F.: SMOTE-RSB*: a hybrid pre-processing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. *Knowl. Inf. Syst.* **33**, 245–265 (2012). <https://doi.org/10.1007/s10115-011-0465-6>
 19. Fernandez, A., Maldonado, S., Vairetti, C., Herrera, F.: FW-SMOTE: a feature-weighted oversampling approach for imbalanced classification. *Pattern Recogn.* **124**, 108511 (2022)
 20. López, J., Maldonado, S., Vairetti, C.: An alternative SMOTE oversampling strategy for high-dimensional datasets. *Appl. Soft Comput. J.* **76**, 380–389 (2019)
 21. Sharma, S., Gosain, A., Jain, S.: A review of the oversampling techniques in class imbalance problem. In: Khanna, A., Gupta, D., Bhattacharyya, S., Hassanien, A.E., Anand, S., Jaiswal, A. (eds.) *International Conference on Innovative Computing and Communications*. AISC, vol. 1387, pp. 459–472. Springer, Singapore (2022). https://doi.org/10.1007/978-981-16-2594-7_38
 22. Saadatfar, H., Mayabadi, S.: Two density-based sampling approaches for imbalanced and overlapping data. *Knowl.-Based Syst.* **241**, 108217 (2022)
 23. Wei, G., Weimeng, M., Song, Y., Dou, J.: An improved and random synthetic minority oversampling technique for imbalanced data. *Knowl.-Based Syst.* **248**, 108839 (2022)