Gaussian-based oversampling approach for imbalanced and overlapped class adapting the minimum covariance determinant

Seung Jee Yang [1,4] Kyung Joon Cha [2,4] Hoeil Chung [3,4]

[1]Department of Applied Statistics, Hanyang University [2]Department of Mathematics, Hanyang University [3]Department of Chemistry, Hanyang University [4]Institute for Convergence of Basic Science, Hanyang University

Introduction

- Many oversampling methods address this issue by generating interpolated instances, but they may lead to overfitting
- Recent studies suggest that class overlap, rather than class imbalance, negatively affects classifier performance
- In this study, we propose a Gaussian-based oversampling adapting minimum covariance determinant (GOMCD) to deal with the class imbalance and overlap simultaneously
- We define the degree of class overlap to generate additional instances in order to improve minority class classification
- Evaluation on simulation and benchmark datasets shows that GOMCD performs well in handling imbalanced data with class overlap

Goals

- Approximate the distribution of the minority class to generate artificial instances following the approximated distribution
- Mitigate the influence of outliers in the process of estimating the distribution using minimum covariance determinant (MCD) estimators
- Generate artificial instances in overlapping area rather than in less overlapping areas

Gaussian mixture model (GMM) with MCD

Probability density function in GMM with MCD is expressed as:

$$p(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi}) = \sum_{g=1}^{G} \pi_g N(\boldsymbol{x}|\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)$$

$$\text{s.t. } (\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g) = \begin{cases} (\boldsymbol{\mu}_{g,GMM},\boldsymbol{\Sigma}_{g,GMM}), & n_g \le 2m \\ (\boldsymbol{\mu}_{g,MCD},\boldsymbol{\Sigma}_{g,MCD}), & n_g > 2m \end{cases}$$

where $\boldsymbol{x} \in \mathbb{R}^m$

Degree of class overlap

For each component of GMM, the degree of class overlap comprise ratio between Mahalanobis-type distance of the minority class and that of the majority class as

$$o_g = \frac{\sum_{i=1}^{n^+} W(d_{i,g}^2) d_{i,g}^2}{\sum_{j=1}^{n^-} W(d_{j,g}^2) d_{j,g}^2} \sim F_{m_g^+,\, m_g^-}$$

where $d_g^2 = (\boldsymbol{x}-\boldsymbol{\mu}_g)^T \boldsymbol{\Sigma}_g (\boldsymbol{x}-\boldsymbol{\mu}_g)$, $W(d^2) = I(P(\chi_m^2 > d^2) \ge \alpha)$, and $m_g = m \times \sum_{i=1}^{n} W\left(d_{i,g}^2\right)$.

Because o_g depends on the degrees of freedom, we use the cumulative probability p_g indirectly instead of o_g as

$$p_g = P\left(F_{m_g^+,\, m_g^-} < o_g\right).$$

Then, cumulative probabilities p_g are standardized, as p_g^*, so that the sum is 1.

$$p_g^* = \frac{p_g}{\sum_{g=1}^{G} p_g}$$

Generation artificial instance

The artificial instance $\boldsymbol{x}_{art}$ follows the distribution as:

$$p(\boldsymbol{x}_{art}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi}) = \sum_{g=1}^{G} \left(\beta p_g^* + (1-\beta)\pi_g\right) \times N(\boldsymbol{x}_{art}|\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g)$$

$$\text{s.t. } (\boldsymbol{\mu}_g,\boldsymbol{\Sigma}_g) = \begin{cases} (\boldsymbol{\mu}_{g,GMM},\boldsymbol{\Sigma}_{g,GMM}), & n_g \le 2m \\ (\boldsymbol{\mu}_{g,MCD},\boldsymbol{\Sigma}_{g,MCD}), & n_g > 2m \end{cases}$$

- The correction parameter β is applied to sampling weights
- The role of β is to determine the proportion of sampling that considers the class overlap among the total oversampling sizes

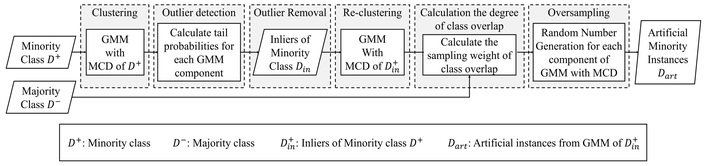

Figure 1. Flowchart of GOMCD

Simulation Study

We varied the imbalance ratio (IR) and adjusted the mean vectors to control the degree of class overlap, generating 20 simulation datasets with 1000 observations in the two-dimensional space.

[Mean vector, Variance]	IR and (Sample size of the Majority, Minority clas)			
	2, (667, 333)	5, (833, 167)	10, (909, 91)	20, (952, 48)
[(-1, -1), 0.4], [(2, 2), 0.4]	A1	A2	A3	A4
[(-1.5, -1.5), 0.4], [(2, 2), 0.4]	B1	B2	B3	B4
[(-1.5, -1.5), 0.4], [(2.5, 2.5), 0.4]	C1	C2	C3	C4
Three components				
[Mean vector, Variance]	IR and (Sample size of the Majority, Minority clas)			
	2, (1334, 666)	5, (1667, 333)	10, (1818, 182)	20, (1904, 96)
[(-1, -1), 0.4], [(1, 1), 0.4], [(-2.7, 2.7), 0.2]	D1	D2	D3	D4
[(-1.5, -1.5), 0.4], [(1.5, 1.5), 0.4], [(-2.7, 2.7), 0.2]	E1	E2	E3	E4

Table 1. Different mean vector and variance settings for generating synthetic datasets

Classifier	Metrics	$\beta = 0.00$	$\beta = 0.25$	$\beta = 0.50$	$\beta = 0.75$	$\beta = 1.00$
kNN	Recall	**2.210**	2.558	2.890	3.295	4.047
	Precision	3.730	3.455	3.315	2.590	**1.910**
	F1-score	3.240	3.112	3.095	2.805	**2.747**
	G-mean	**2.590**	2.760	2.918	3.088	3.645
RF	Recall	**2.260**	2.495	2.905	3.407	3.932
	Precision	3.805	3.415	3.215	2.642	**1.922**
	F1-score	3.325	3.105	3.112	2.812	**2.645**
	G-mean	**2.640**	2.692	2.920	3.180	3.568
SVM	Recall	**2.060**	2.358	2.925	3.485	4.172
	Precision	3.765	3.405	3.208	2.612	**2.010**
	F1-score	3.132	2.900	2.990	**2.898**	3.080
	G-mean	**2.412**	2.450	2.927	3.252	3.957

Table 2. Average rankings of performance metric with varying β of GOMCD

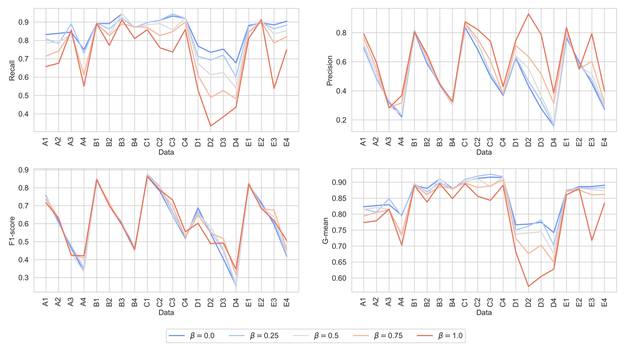

Figure 2. Influence of varying β on classification performance using SVM

Artificial samples generated by different methods

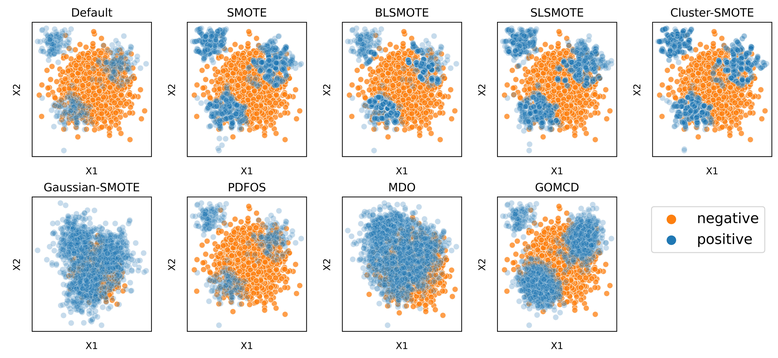

Figure 3. Simulation dataset E2 and artificial samples

Comparison oversampling methods: SMOTE[1], BLSMOTE[2], SLSMOTE[3], Cluster-SMOTE[4], Gaussian-SMOTE[5], PDFOS[6], MDO[7]

Results

We used 24 datasets that exhibit a wide range of IRs, ranging from 2.9 to 72.69, from the KEEL repository. 11 out of 24 datasets have an IR exceeding 10. The correction parameter β in GOMCD was set to 0.25.

Classifier	Metrics	GOMCD	Default	SMOTE	BLSMOTE	SLSMOTE	Cluster -SMOTE	Gaussian -SMOTE	PDFOS	MDO
kNN	Recall	**4.075**	6.675	4.083	5.260	4.348	4.396	5.775	4.906	5.481
	Precision	5.115	4.406	5.637	5.283	5.994	5.577	4.275	4.521	**4.192**
	F1-score	4.604	5.298	5.202	5.256	5.583	5.196	4.721	4.646	**4.494**
	G-mean	**4.238**	5.896	4.673	5.379	5.167	4.821	5.208	4.773	4.846
RF	Recall	4.256	6.090	4.852	5.687	4.715	5.217	4.762	**4.138**	5.283
	Precision	**4.613**	4.719	4.956	4.958	5.129	4.844	5.179	5.340	5.262
	F1-score	**4.340**	5.310	4.808	5.360	4.833	4.925	5.077	5.096	5.250
	G-mean	**4.192**	5.529	4.892	5.477	4.850	5.069	5.062	4.581	5.348
SVM	Recall	**3.775**	5.850	5.308	5.935	4.952	5.804	4.594	3.865	4.917
	Precision	5.252	**3.990**	4.971	4.896	5.623	4.771	5.512	5.390	4.596
	F1-score	4.583	4.619	5.008	5.390	5.448	5.206	5.192	4.988	**4.567**
	G-mean	**4.133**	5.004	5.087	5.681	5.392	5.431	5.123	4.473	4.675

Table 3. Average rankings of performance metric on KEEL datasets with comparison methods

Conclusions

- We defined the degree of class overlap based on the Mahalanobis-type distance which can help to address the class overlap by generating more minority class instances as the degree of overlap increases
- The analysis of 24 KEEL datasets shows that GOMCD achieved a higher recall and G-mean, regardless of IR increasing

Acknowledgement

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (NRF-2020R1A6A1A06046728).

References

[1] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of artificial intelligence research 16 (2002) 321–357.

[2] H. Han, W.-Y. Wang, B.-H. Mao, Borderline-smote: a new over-sampling method in imbalanced data sets learning, in: International conference on intelligent computing, Springer, 2005, pp. 878–887.

[3] C. Bunkhumpornpat, K. Sinapiromsaran, C. Lursinsap, Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem, in: Advances in Knowledge Discovery and Data Mining: 13th Pacific-Asia Conference, PAKDD 2009 Bangkok, Thailand, April 27-30, 2009 Proceedings 13, Springer, 2009, pp. 475–482.

[4] D. A. Cieslak, N. V. Chawla, A. Striegel, Combating imbalance in network intrusion datasets., in: GrC, 2006, pp. 732–737.

[5] H. Lee, J. Kim, S. Kim, Gaussian-based smote algorithm for solving skewed class distributions, International Journal of Fuzzy Logic and Intelligent Systems 17 (4) (2017) 229–234.

[6] M. Gao, X. Hong, S. Chen, C. J. Harris, E. Khalaf, Pdfos: Pdf estimation based over-sampling for imbalanced two-class problems, Neurocomputing 138 (2014) 248–259.

[7] L. Abdi, S. Hashemi, To combat multi-class imbalanced problems by means of over-sampling techniques, IEEE transactions on Knowledge and Data Engineering 28 (1) (2015) 238–251.