Multi-oversampling with Evidence Fusion for Imbalanced Data Classification

Hongpeng Tian[1], Zuowei Zhang[2](✉), Zhunga Liu[2], and Jingwei Zuo[3]

[1] School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou, China

[2] School of Automation, Northwestern Polytechnical University, Xi'an, China
{zhangzuowei,liuzhunga}@nwpu.edu.cn

[3] Technology Innovation Institute, Abu Dhabi, UAE
Jingwei.Zuo@tii.ae

Abstract. Oversampling methods concentrate on creating a balanced dataset by generating samples, widely utilized in classifying imbalanced data. However, current oversampling methods overlook the uncertainty in the samples produced, potentially shifting the data's distribution and adversely affecting the classification outcomes. To address this problem, we introduce a multi-oversampling with evidence fusion (MOEF) method for imbalanced data classification based on Dempster-Shafer theory. We first design a multi-oversampling strategy to produce various balanced datasets, characterizing the uncertainty of generated samples. Then, we develop a discounting fusion rule based on the inconsistency of data distribution post-oversampling, thereby mitigating the adverse effects of data distribution alterations on classification. Extensive testing on various imbalanced datasets indicates that the proposed MOEF method exhibits more satisfactory performance than other related methods.

Keywords: Oversampling · Evidence fusion · Imbalanced data · Dempster-Shafer theory

1 Introduction

Imbalanced data classification is a pervasive challenge in machine learning, occurring when datasets exhibit an unequal distribution of samples between classes [1]. This can significantly impair the performance of basic classifiers, which often prioritize overall accuracy, leading to a pronounced bias towards the majority class and neglect of minority class samples. Correctly identifying minority class samples is crucial in many applications [2], such as disease diagnosis, fraud detection, and anomaly identification, where the misclassification of such samples can have severe consequences.

Recently, a large number of methods for classifying imbalanced data have been developed. Among them, sampling methods [3], one of the most popular and effective methods, focus on preprocessing the input data to balance the

Y. Bi et al. (Eds.): BELIEF 2024, LNAI 14909, pp. 68–77, 2024.
https://doi.org/10.1007/978-3-031-67977-3_8

classes, leading to more equitable model learning and improved performance on minority class prediction compared to other methods. They can be broadly classified into undersampling and oversampling methods. Undersampling methods [4,5], aiming to counteract class imbalance by downsizing the majority class. Random undersampling (RUS) [4], for instance, alleviates class imbalance by randomly eliminating samples from the majority class until a desired balance is achieved. However, RUS may lead to information loss, as valuable samples from the majority class could be discarded. Unlike undersampling, oversampling methods produce artificial data aimed at augmenting minority sample numbers, thereby preserving crucial details about minority samples and enhancing the conceptual portrayal of the minority group. Random oversampling (ROS) [6], for instance, addresses imbalance by duplicating random samples from the minority class, thereby increasing their representation in the dataset. Chawla *et al.* [7] introduce a synthetic minority oversampling technique (SMOTE), synthesizing minority class samples by interpolating between existing samples and their nearest neighbors. This method effectively enhances minority class representation while preserving data distribution characteristics. Han *et al.* [8] present a Borderline-SMOTE, an extension of SMOTE that focuses on oversampling minority samples near the decision boundary. Chao *et al.* [9] propose a novel data augmentation method called H-SMOTE, combining the notion of neighbors and Manhattan distance to produce new samples.

The aforementioned oversampling methods have proven effective in classifying imbalanced data. However, each of these methods yields unique synthetic specimens for the minority class but fails to accurately represent the uncertainty in the produced samples. Inevitably, there is some deviation between the produced sample and the actual sample. Furthermore, this generational divergence could alter the original dataset's distribution, potentially leading to poor classification outcomes.

Dempster-Shafer theory [10,11], also known as the theory of belief functions or evidence reasoning, has been appealing for reasoning uncertain information and widely used in data classification [12–15]. To overcome the above limitations of existing oversampling methods, we propose a multi-oversampling with evidence fusion (MOEF) method for imbalanced data classification with Dempster-Shafer theory in this paper. The contributions of MOEF can be summarized in three aspects. 1) We design a multi-oversampling strategy to generate multiple versions of synthetic samples, characterizing the uncertainty of generated samples. 2) We develop an evidence fusion rule according to the inconsistency of data distribution post-oversampling, weakening the negative impact of changes in data distribution on classification. 3) We apply MOEF to several real imbalanced datasets to demonstrate its superiority over other related methods.

The rest of this paper is arranged as follows. The proposed method is presented in detail in Sect. 2. Then, it is tested in Sect. 3 and compared with several other typical methods, followed by conclusions.

2 Multi-oversampling with Evidence Fusion for Imbalanced Data Classification

This part introduces the proposed MOEF method in detail. Assume that a test set $Y = \{\mathbf{y}_1, ..., \mathbf{y}_M\}$ is classified under the discernment framework $\Omega = \{\omega_{min}, \omega_{maj}\}$, based on a training set $X = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ across H distinct attribute spaces. X_{min} and X_{maj} denote the minority and majority class, respectively.

2.1 Multi-oversampling for the Minority Class

In this subsection, we introduce a multi-oversampling technology for the minority class. In this way, various balanced datasets can be acquired, thereby training a basic classifier.

A training sample $\mathbf{x}_i$ is randomly selected from the minority class X_{min}. For $\mathbf{x}_i$, we search neighbors from X_{min}, denoted as $\mathbf{x}_k$ $(k = 1, ..., K)$. Then, we can generate synthetic samples between $\mathbf{y}_i$ and neighbors $\mathbf{x}_k$ $(k = 1, ..., K)$ like SMOTE. Then, other training samples from the minority class also generate synthetic samples in this way, and this process is terminated until the number of generated samples is equal to the number $(i.e., |X_{maj}| - |X_{min}|)$ of synthetic samples that need to be expanded. However, the generated samples are random and there exists uncertainty in this process. It cannot characterize the uncertainty if we just generate a specific synthetic sample between $\mathbf{x}_i$ and a neighbor $\mathbf{x}_k$. Thus, we introduce a multi-oversampling technology for $\mathbf{x}_i$ to generate multiple synthetic samples, thereby characterizing the uncertainty of synthetic samples. The generated sample $\mathbf{x}_i^t$ $(t = 1, ..., T)$ is given by:

$$\mathbf{x}_i^t = \mathbf{x}_i + \alpha_t(\mathbf{x}_k - \mathbf{x}_i) \tag{1}$$

where α_t is a random number, such that $\alpha_t \in [0, 1]$. By generating T different values of α_t, $\mathbf{x}_i$ can generate T synthetic samples in this way. Each sample used for oversampling can generate T synthetic samples, which are combined with training samples to balance the training set. By doing this, we obtain a total of T balanced training sets, denoted as X^t $(t = 1, ..., T)$.

A basic classifier is trained using T balanced training sets to classify test samples. With a given test sample $\mathbf{y}_j$, T classification outcomes P_j^t $(t = 1, ..., T)$ are achievable. Decision-making for $\mathbf{y}_j$ can be based on the Dempster-Shafer (DS) rule. Nonetheless, oversampling might alter the distribution of the minority class. The outcomes yielded by these classifiers vary in dependability and could lead to discrepancies, resulting in implausible results according to the DS rule in this case. Consequently, the subsequent part will outline methods for efficiently assessing the reliability of various classification outcomes and their combinations.

2.2 Evidence Fusion with Discounting Factors

This part assesses the reliability of various classifiers based on the inconsistency in the distribution of data post-oversampling. Subsequently, we implement evidence fusion with different discounting factors using the DS rule [10].

The unavoidable discrepancy between generated synthetic and actual samples leads to a variance in the distribution of the initial minority class compared to the class post-oversampling, potentially adversely impacting the classification outcomes. The maximum mean discrepancy (MMD) [16] is utilized to measure the variance in the distribution between the initial minority class and the class post-oversampling, indicated as:

$$MMD(X_{min}, X_{min}^t) = \left\| \frac{1}{|X_{min}|} \sum_{\mathbf{x}_i \in X_{min}} \phi(\mathbf{x}_i) - \frac{1}{|X_{min}^t|} \sum_{\mathbf{x}_j \in X_{min}^t} \phi(\mathbf{x}_j) \right\| \quad (2)$$

where X_{min}^t symbolizes the minority class oversampling at the t-th level. The symbol $|.|$ represents the number of elements. $\phi(.)$ denotes a mapping function used to align X_{min} and X_{min}^t into a singular space, thus determining their variance.

For the training set X^t after oversampling, the lower value of inconsistency of the minority class, the more reliability of the classification result P_j^t corresponding to X^t. Thus, the reliability δ^t of P_j^t is denoted as:

$$\delta^t = e^{-MMD(X_{\min}, X_{\min}^t)} \Big/ \sum_{t=1}^{T} e^{-MMD(X_{\min}, X_{\min}^t)} \quad (3)$$

The discounted masses of belief are obtained by:

$$\begin{cases} m_j^t(A) = \delta^t p_j^t(A), A \subset \Omega; \\ m_j^t(\Omega) = 1 - \delta^t + \delta^t p_j^t(\Omega). \end{cases} \quad (4)$$

Through this process, information that has been devalued can be submitted to the total unknown class (Ω), thereby reducing the degree of discord among pieces of evidence. This permits the application of the DS rule for fusing discounted results. The combined masses of belief $m_j(A)$ for the sample $\mathbf{y}_j$ can be transformed into pignistic probability for the ultimate decision. Defining the pignistic probability is as follows:

$$BetP(\omega_c) = \sum_{A \in 2^\Omega, \omega_c \in A} \frac{1}{|A|} m_j(A) \quad (5)$$

where $|A|$ represents the number of elements in A. Subsequently, the sample $\mathbf{y}_j$ can be assigned to the class with the highest probability.

3 Experiment Applications

This part evaluates the efficacy of the proposed MOEF method against a range of other commonly used methods. A pair of prevalent indexes [17], named F-measure (FM) and G-mean (GM), commonly applied in classifying imbalanced data, are utilized to assess various methods.

3.1 Benchmark Datasets

To assess the effectiveness of diverse methods in classifying imbalanced data, we employ twelve realistic imbalanced datasets extracted from two reputable sources: the KEEL dataset repository (available at http://www.keel.es/) and the UCI Machine Learning Repository (available at http://archive.ics.uci.edu/ml). A summary of the pivotal characteristics of these datasets utilized in the experimental phase is presented in Table 1, detailing the total number of attributes (#Attr), samples (#Size), the majority class samples (#Maj), the minority class samples (#Min), and the imbalance ratio (#IR). To guarantee a sturdy and comprehensive analysis, a rigorous five-fold stratified cross-validation scheme is applied to each dataset, thereby ensuring the reliability and generalizability of the evaluation results.

Table 1. Basic information of the keel datasets

Data	#Attr.	#Size.	#Min.	#Maj.	#IR.
Immunotherapy	7	90	19	71	3.74
Climate	18	540	46	494	10.74
Ecoli3	7	336	35	301	8.60
Ecoli4	7	336	20	316	15.80
Page-blocks0	10	5472	559	4913	8.79
Glass	19	2308	329	1979	6.02
Shuttlec2vsc4	9	129	6	123	20.50
Statlog	13	270	120	150	1.25
Wdbc	7	90	19	71	3.74
Yeast1vs7	7	457	30	427	14.23
Yeast1	8	1484	429	1055	2.46
Vehicle2	18	846	218	628	2.88

3.2 Comparison Methods

The performance of the proposed method is systematically evaluated in comparison with other related methods. Specifically, RUS [4] alleviates class imbalance by randomly removing majority samples, while ROS [6] duplicates minority samples. SMOTE [7] generates synthetic minority samples along linear interpolations between a sample and its nearest neighbors. Borderline-SMOTE [8] focuses on oversampling minority samples near the decision boundary. H-SMOTE [9] combines nearest neighbors and Manhattan distance to create synthetic samples directed toward the minority class center.

3.3 Performance Evaluation

This study utilizes twelve imbalanced datasets to investigate MOEF's efficacy by contrasting it with alternative comparison methods in real-world datasets. Table 2 and Table 3 detail the FM and GM values for various classification methods. In the last row of these tables, we report the number of wins/ties/losses (W/T/L) for each method compared to the one with the highest rank. It's evident that in the majority of datasets, MOEF outperforms other comparison methods. This is due to MOEF producing various iterations of synthetic samples, thereby characterizing the uncertainty in the oversampling process. Furthermore, MOEF assesses the inconsistency in data distribution post-oversampling, thereby mitigating the adverse effects of data distribution alterations on classification. Consequently, MOEF is capable of achieving more robust and satisfactory performance compared to other comparison methods.

Table 2. FM of imbalanced datasets by different methods (IN %)

Datasets	RUS	ROS	SMOTE	Borderline-SMOTE	H-SMOTE	MOEF
Immunotherapy	49.22	47.08	52.69	45.32	37.98	**54.95**
Climate	54.44	42.49	62.19	60.91	**62.85**	62.29
Ecoli3	58.39	50.37	59.38	56.89	59.74	**62.97**
Ecoli4	73.55	62.33	76.55	73.09	27.45	**7 8.00**
Page-blocks0	60.27	58.70	59.62	46.03	**68.57**	59.54
Glass	98.48	89.76	98.48	78.92	98.33	**98.48**
Shuttlec2vsc4	81.33	71.82	**93.33**	**93.33**	**93.33**	**93.33**
Statlog	81.52	82.56	81.70	80.26	81.51	**82.95**
Wdbc	51.48	45.33	53.48	44.46	43.33	**54.71**
Yeast1vs7	40.03	28.98	36.87	34.03	29.26	**40.33**
Yeast1	57.58	57.37	57.57	55.57	47.07	**57.81**
Vehicle2	75.69	72.19	75.58	73.50	69.86	**76.01**
W/T/L	0/0/12	0/0/12	0/1/11	0/1/11	2/1/9	**9/1/2**

3.4 Influence of *T*

The parameter T plays a pivotal role in the proposed MOEF method and could substantially influence the efficiency of MOEF. T denotes the count of different forms of generated synthetic samples. This experiment utilizes a range of imbalanced datasets to examine the efficacy of MOEF across different T values,

Table 3. GM of imbalanced datasets by different methods (IN %)

Datasets	RUS	ROS	SMOTE	Borderline-SMOTE	H-SMOTE	MOEF
Immunotherapy	67.47	67.13	72.19	64.83	50.55	**75.05**
Climate	85.27	80.15	86.91	85.85	68.41	**86.92**
Ecoli3	82.46	86.02	85.05	86.60	79.79	**87.07**
Ecoli4	95.20	92.91	95.50	94.91	72.96	**95.66**
Page-blocks0	85.41	87.03	87.27	83.19	77.79	**87.47**
Glass	**99.11**	97.46	**99.11**	95.42	99.09	**99.11**
Shuttlec2vsc4	92.85	87.35	94.14	**94.14**	**94.14**	**94.14**
Statlog	82.98	83.81	82.77	81.55	82.96	**84.22**
Wdbc	67.67	63.62	68.44	58.97	52.58	**70.20**
Yeast1vs7	75.73	75.69	74.26	71.29	46.89	**78.10**
Yeast1	66.97	66.51	66.79	62.24	59.29	**67.05**
Vehicle2	86.33	83.71	86.31	84.28	79.42	**86.71**
W/T/L	0/1/11	0/0/12	0/1/11	0/1/11	0/1/11	**10/2/0**

aiding in the application of T. The illustrations displayed in Fig. 1 symbolize the classification outcomes of MOEF, each with varying parameter values. In this context, the x-coordinate signifies T values between 2 and 8, while the y-coordinate indicates the values of FM and GM. The findings clearly show that MOEF remains stable regardless of T's value, with minimal fluctuations noted as T increases. Furthermore, the value of T should not be too small, as it may fail to characterize the uncertainty of generated samples. However, if the T value is set too large, it will bring a lot of computational burden. Therefore, our suggestion is to set $T \in [3, 5]$ as the default value in applications.

4 Complexity Analysis and Runtime Comparison

MOEF's computational complexity primarily hinges on calculating the distances between between samples of minority class samples during the process of over-sampling. In the minority class $\mathcal{X}_{min}$, the sample $\mathbf{x}_i$ calculates the distance from $\mathbf{x}_i$ to every minority sample in $\mathcal{X}_{min}$, with the computational complexity denoted as $\mathcal{O}(|\mathcal{X}_{min}|)$, where $|\mathcal{X}_{min}|$ signifies the sample count in $\mathcal{X}_{min}$. It's presumed that $N'(N' \leq |\mathcal{X}_{\mathrm{min}}|)$ samples require neighbors searching upon the conclusion of the over-sampling procedure. Consequently, MOEF's overall computational complexity is $\mathcal{O}(N' \, |\mathcal{X}_{min}|)$.

Table 4 displays the execution time in seconds for MOEF and various other methods. It's evident that MOEF's execution duration isn't the briefest, as it requires computing numerous distances among samples to acquire neighboring samples. Within practical scenarios, MOEF proves more apt for situations demanding high precision, in contrast to situations where efficient computing isn't essential.

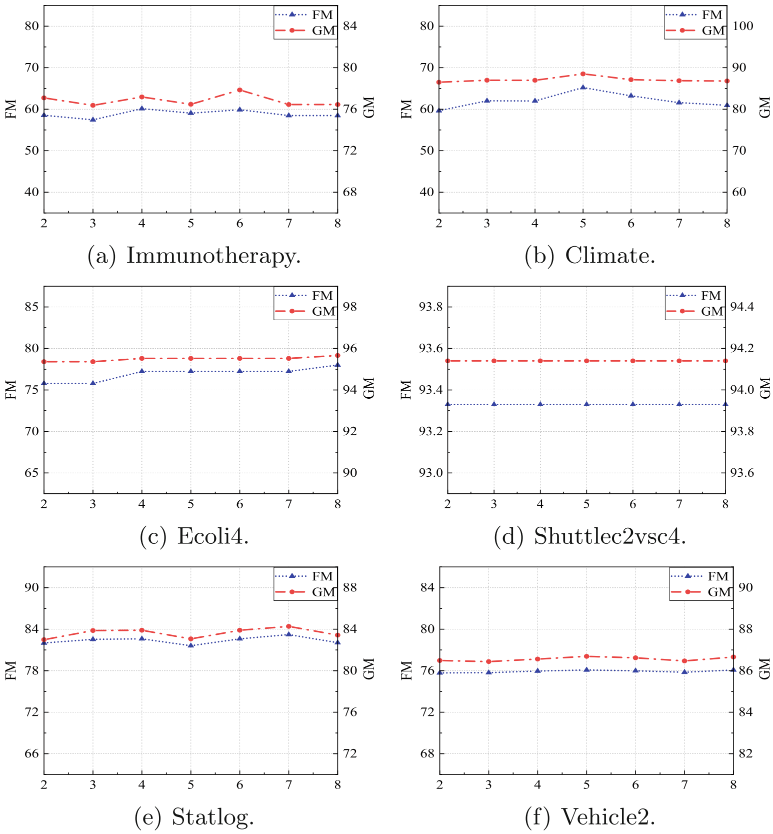

Fig. 1. Classification results of MOEF with various values of T.

Table 4. Execution time of different methods (In seconds)

Datasets	RUS	ROS	SMOTE	Borderline-SMOTE	H-SMOTE	MOEF
Immunotherapy	0.17	1.28	0.26	0.11	0.23	0.53
Climate	0.15	0.78	0.38	0.31	0.30	0.81
Ecoli3	0.15	0.76	0.29	0.19	0.20	0.48
Ecoli4	0.15	0.77	0.30	0.18	0.20	0.51
Page-blocks0	0.18	1.52	4.99	6.39	115.96	16.05
Glass	0.17	0.92	1.40	2.31	8.88	3.41
Shuttlec2vsc4	0.15	0.78	0.24	0.12	0.19	0.29
Statlog	0.15	0.76	0.21	0.10	0.19	0.32
Wdbc	0.15	0.78	0.21	0.09	0.19	0.25
Yeast1vs7	0.15	0.76	0.35	0.25	0.26	0.67
Yeast1	0.15	0.81	0.49	0.57	2.49	1.67
Vehicle2	0.16	0.78	0.40	0.30	0.61	0.86

5 Conclusion

This paper proposes a multi-oversampling with evidence fusion (MOEF) method for imbalanced data classification based on Dempster-Shafer theory. MOEF implements multi-oversampling to characterize the uncertainty of generated samples in the oversampling process. Moreover, MOEF quantifies the degree of inconsistency in data distribution post-oversampling, weakening the negative impact of data distribution alterations on classification. The experiments on synthetic and several real imbalanced datasets have verified the effectiveness of MOEF compared to typical methods. Moreover, we also investigate the influence of the parameter on the classification performance of MOEF and provide guidance on parameter settings. In the future, we will extend the application scope of MOEF to other real-world tasks.

Acknowledgment. This work was supported by the Fundamental Research Funds for the Central Universities under Grant G2023KY05102.

References

1. Xu, Y., Yu, Z., Chen, C.L.P.: Classifier ensemble based on multiview optimization for high-dimensional imbalanced data classification. IEEE Trans. Neural Netw. Learn. Syst. **35**(1), 870–883 (2024)
2. Bai, L., Ju, T., Wang, H., Lei, M., Pan, X.: Two-step ensemble under-sampling algorithm for massive imbalanced data classification. Inf. Sci. 120351 (2024)
3. Tian, H., Zhang, Z., Martin, A., Liu, Z.: Reliability-based imbalanced data classification with Dempster-Shafer theory. In: Le Hegarat-Mascle, S., Bloch, I., Aldea, E. (eds.) BELIEF 2022. LNCS, vol. 13506, pp. 77–86. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-17801-6_8
4. Zhang, Y., Liu, G., Luan, W.: An approach to class imbalance problem based on stacking and inverse random under sampling methods. In: 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), pp. 1–6. IEEE (2018)
5. Lin, W.C., Tsai, C.F., Hu, Y.H., et al.: Clustering-based undersampling in class-imbalanced data. Inf. Sci. **409**, 17–26 (2017)
6. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. **21**(9), 1263–1284 (2009)
7. Chawla, N.V., Bowyer, K.W., Hall, L.O., et al.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)
8. 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
9. Chao, X., Zhang, L.: Few-shot imbalanced classification based on data augmentation. Multimed. Syst. **29**(5), 2843–2851 (2023)
10. Dempster, A.P.: Upper and Lower probabilities induced by a multivalued mapping. Ann. Statist. **83**, 325–339 (1967)
11. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)

12. Zhang, Z., Tian, H., Yan, L.Z., Martin, A., Zhou, K.: Learning a credal classifier with optimized and adaptive multiestimation for missing data imputation. IEEE Trans. Syst. Man Cybern. Syst. **52**(7), 1–13 (2021)
13. Zhang, Z., Liu, Z., Ning, L., Martin, A., Xiong, J.: Representation of imprecision in deep neural networks for image classification. IEEE Trans. Neural Netw. Learn. Syst. (2023). https://doi.org/10.1109/TNNLS.2023.3329712
14. Niu, J., Liu, Z.G., Lu, Y., Wen, Z.: Evidential combination of classifiers for imbalanced data. IEEE Trans. Syst. Man Cybern. **52**(12), 7642–7653 (2022)
15. Zhang, Z., Ye, S., Zhang, Y., Ding, W., Wang, H.: Belief combination of classifiers for incomplete data. IEEE-CAA J. Automatica Sin. **9**(4), 652–667 (2022)
16. Jia, X., Zhao, M., Di, Y., Yang, Q., Lee, J.: Assessment of data suitability for machine prognosis using maximum mean discrepancy. IEEE Trans. Ind. Electron. **65**(7), 5872–5881 (2017)
17. Wong, T.T.: Linear approximation of f-measure for the performance evaluation of classification algorithms on imbalanced data sets. IEEE Trans. Knowl. Data Eng. **34**(2), 753–763 (2022)