https://doi.org/10.46335/IJIES.2024.9.1.15 **e-ISSN: 2456-3463**
Vol. 9, No. 1, 2024, PP. 84-89

International Journal of Innovations in Engineering and Science, www.ijies.net

A Systematic Approach of Feature Selection for Imbalanced Data: A Review

Mr. Vijay Santosh Tawar[1], Mr. Machchhindra Jibhau Garde[2]

[1]Assistant Professor – Department of Electronics & Telecommunication Engineering,
SSVPS' B. S. Deore College of Engineering, Dhule, India, 424002
[2]Assistant Professor – Department of Electronics Engineering
SSVPS' B. S. Deore College of Engineering, Dhule, India, 424002

Email of Corresponding Author: ***kvtawar@gmail.com***

Received on*: 17 March, 2024* ***Revised on****: 20 April, 2024* ***Published on****: 22 April, 2024*

Abstract – *There are several feature selection methods proposed for imbalanced data. These methods can be categorized into filtering methods, wrapper methods, and built-in methods. The filtering method ranks objects based on statistical properties and selects the highest ranked object. The wrapper method uses a special classifier to evaluate the performance of different feature subsets and selects the best subset. The built-in method involves feature selection for the classifier training process. The purpose of methods used for imbalanced data is to solve problems associated with imbalanced class distribution in classification problems. The goal of this method is to improve the performance of existing classification algorithms when working with imbalanced data. In this paper, we review feature selection methods based on a variety of techniques, including multi-objective ant colony optimization, specifically designed for imbalanced data.*

Keywords- Random over-sampling (ROS), Adaptive synthetic sampling (ADASYN), Genetic algorithm-based under-sampling (GAUS).

I .INTRODUCTION

The need for feature selection in imbalanced data is determined by several factors. Feature selection helps eliminate over fitting of imbalanced data by reducing the number of irrelevant or redundant features that can cause the model to learn from noise or be biased toward most classes [1]. The classifier may consider minority class instances as outliers or noise, resulting in biased predictions. Feature selection can help mitigate this bias by optimizing the contrast between classes and focusing on features rather than training examples. Feature selection can improve the predictive performance of a model by selecting relevant features that have a stronger relationship with the target variable [5]. This can improve classification and model interpretation performance. Feature selection reduces data dimensionality, reducing computation time and storage requirements. This is especially important when working with multidimensional imbalanced data [3]. Imbalanced data can negatively impact the performance of feature selection algorithms designed for balanced data.

Using multi-objective ant colony optimization, feature selection can be effectively implemented while avoiding the negative consequences of imbalanced data [2]. Because the data distribution of high-dimensional imbalanced data is complex, feature selection is important to identify discriminatory features and improve classification performance. Imbalanced data refers to a situation where the number of instances of

https://doi.org/10.46335/IJIES.2024.9.1.15 e-ISSN: 2456-3463
Vol. 9, No. 1, 2024, PP. 84-89

one class is significantly more or fewer than the number of instances of another class. Traditional classification algorithms assume that the data is well distributed, but imbalanced data can bias training algorithms. As you know, research on imbalanced classification has been attracting attention recently. To solve problems related to imbalanced class distribution and improve the performance of classifiers, a method to classify imbalanced data is needed [8].

Existing algorithms can be applied to imbalanced data using a variety of methods, including data-level methods, algorithm-level methods, and ensemble learning methods. Data layer methods use sampling techniques to balance imbalanced data, such as undersampling and oversampling [11]. Algorithm-level methods create new algorithms or modify existing ones to reduce the drawbacks of imbalanced data, such as cost-aware learning, single-class learning, and feature methods. Ensemble learning methods combine ensemble learning with data- or algorithm-level methods to improve classification performance [14]. High-dimensional imbalanced data presents additional challenges due to its dimensionality and imbalance. Feature selection is an important aspect of classification for imbalanced data, as it helps identify relevant features and reduce dimensionality [6]. Multi-objective ant colony optimization and genetic algorithms are used in the proposed classification method for imbalanced data. Figure 1 and figure 2 below depicts illustration for balanced data and imbalanced data –

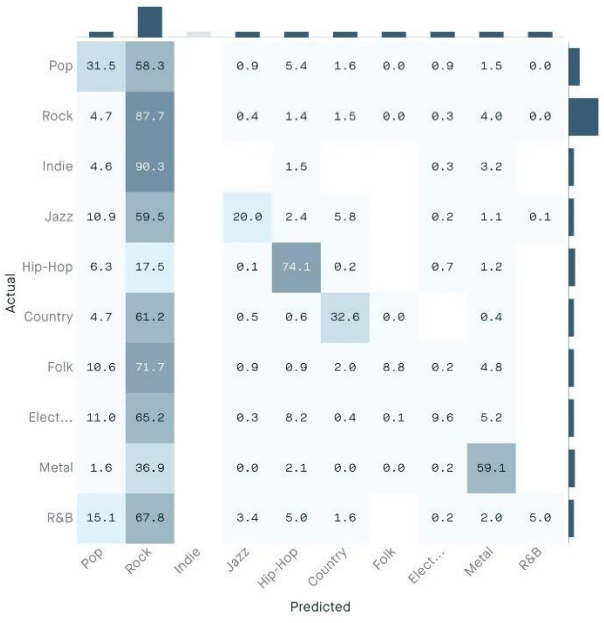

Fig. 1 - Depiction of Imbalanced Data

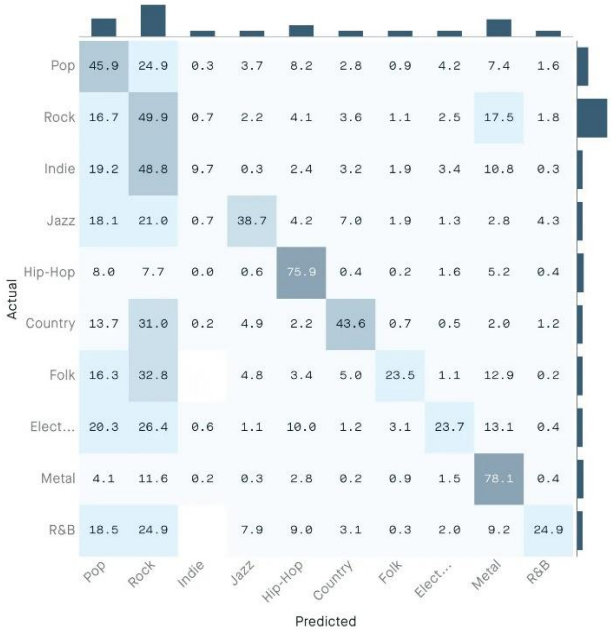

Fig. 2: Depiction of Balanced Data

II. LITERATURE REVIEW

Following are the related works carried out and methods used -

Under sampling technology: Under sampling balances the data by selecting instances of the majority class and combining them with instances of the minority class. Random Under sampling (RUS) is a widely used method based on evolutionary algorithms. Yu et al. proposed an under sampling strategy based on ant colony optimization [9]. Krawczy et al. We build an algorithm based on boosting using under sampled genetic algorithms as building blocks. Triguero et al. use genetic algorithm under sampling with map pruning architecture to balance large, unbalanced data. Oversampling method. Common oversampling technologies include Random Oversampling (ROS) and Synthetic Minority Sampling (SMOTE). SMOTE selects one minority class instance and creates new minority class instances based on the difference between itself and its selected neighbor classes [5]. Adaptive Synthetic Sampling (ADASYN) is an efficient approach based on SMOTE [7]. Ramenthol et al. use fuzzy approximate sets to measure the degree to which newly created instances differ from the original instances.

Algorithm-level methods: Cost-sensitive approaches do not change the data distribution, but generate a cost

https://doi.org/10.46335/IJIES.2024.9.1.15 e-ISSN: 2456-3463
Vol. 9, No. 1, 2024, PP. 84-89

matrix to assign the higher cost of misclassifying minority class occurrences [13]. Single-class learning methods also do not change the data distribution, but rather obtain similarity values between instances based on their attributes and classify each instance according to predefined similarity criteria.

Feature selection method: Yin et al. come up with a feature selection method based on decomposition for binary imbalance problems [8]. Moayedikia et al. use feature selection based on harmonic search to remove the influence of high-dimensional imbalanced data [12]. Du et al. To improve classifier performance, we implement feature selection for multi-class imbalanced data using genetic methods. Fernandez et al. associate sampling with feature selection, training samples for chromosomes, and feature encoding [4]. Du et al. combine the area under the ROC curve and the difference between the selected and original sample numbers as the optimization objective.

Ensemble learning methods: Bagging, boosting, and Adaboost are well-known ensemble learning methods. Sun et al. propose a bag-based multi-classifier method. Guo et al. use feature selection based on Adaboost and binary particle swarm optimization to solve multi-class imbalanced classification problems. Liu et al. developed a self-adaptive ensemble classification algorithm that simultaneously uses ensemble learning, data-level approaches, and algorithm-level methods to explicitly solve the multi-class imbalanced classification problem [4]. Ensemble learning techniques can increase the flexibility and reliability of the original algorithm, but are time-consuming, especially for Boosting and Adaboost [6].

Structure and methodology of GU-MOACOFS: GU-MOACOFS is a platform for resampling data using GAUS (Genetic Algorithm Undersampling). Bagging was chosen for its high throughput, ease of deployment, and minimal time investment compared to alternative iterative methods. V-statistic is proposed to improve sampling efficiency by measuring the distribution of undersampled data using a genetic algorithm. SU is used to discard noise and useless features to reduce time consumption when the data dimensionality is high. To implement feature selection, multi-criteria optimization of ant colonies is used [6]. GU-MOACOFS first uses sampling and then feature selection using distributional information in individual subsets of the samples. The V statistic is used to measure the complexity of unbalanced data, including measures of overlap of individual feature values, class separability, geometry, topology, and manifold density. GAUS is an independent method that uses evolutionary algorithms to resample data to solve problems with standard algorithms, which can be highly correlated with classifiers and can be time-consuming [12]. In this study, we use the single-point intersection suite of GAUS.

Multi-objective Ant Colony Optimization Feature Selection approach for stratification of Imbalanced Data:

Feature selection and optimization: Feature selection is a common preprocessing strategy that can improve model interpretation, reduce data storage, and improve classification performance. Traditional feature selection methods may cause more errors due to data imbalance. Multi-objective ant colony optimization is used to implement feature selection that is best suited to solve a subset of the problem. One of the proposed methods may run into problems if the data dimensionality is quite large [6]. To solve this problem, we use the metric SU to measure entropy-based feature correlation. SU successfully selects features from large datasets. This method can be used to initially measure feature correlation to remove noise and redundant features.

Complexity analysis of GU-MOACOFS: The complexity of GU-MOACOFS is determined by the number of data subsets, the dimensionality of the original data, the number of GAUS iterations, the GAUS population size, the number of MOACO iterations, and the MOACO population size. The sampling time of Bootstrap is O(n), the overall complexity of GAUS crossover and mutation operations is O(N_g), and the GAUS evaluation time for each V-statistic solution is O(C). The most time-consuming part of MOACO is the ants finding a solution, which is O(ite_mXN_mXC2).

GU-MOACOFS pseudocode: The code begins by initializing the parameters and setting the maximum iteration value (iter). If the dimensionality of the input data exceeds a certain value, the correlation of each feature is extracted, sorted in descending order, and the previous N features are selected to generate new input data. The code then uses bootstrapping to create n groups of data subsets. The algorithm uses an undersampling implementation to generate an appropriate subset of samples. The number of selected objects is fixed, 100 for all high-dimensional datasets [12].

https://doi.org/10.46335/IJIES.2024.9.1.15 e-ISSN: 2456-3463
Vol. 9, No. 1, 2024, PP. 84-89

Datasets and measurements: The study uses 14 data sets, including 10 low-dimensional data sets and 4 high-dimensional data sets for comparative study. The number of selected objects is fixed and equal to 100 for all high-dimensional datasets [9]. Classification success rate is used to evaluate the results, AUC ranges from 0.5 to 1, with higher AUC indicating a better classifier. Geometric mean (Gmean) is used to determine the performance of the classifier [10].

Genetic algorithm undersampling evaluation: In the review, observation of the performance is done for the proposed genetic algorithm based undersampling (GAUS). The algorithms compared were ROS, RUS, and SMOTE. GAUS settings include maximum number of iterations, population size, crossover rate, mutation rate, and mutation rate. Cross-testing was performed in five rounds of 20 [8]. Feature selection was performed in SU before running each method on four multivariate data sets. GAUS, SMOTE, and RUS achieved 7, 4, and 1 best results on 10 low-dimensional datasets. GAUS performed best on eight datasets, including ROS from GLASS4 and RUS from YEAST2VS8. Random sampling methods (ROS and RUS) have been shown to perform worse than criterion-based sampling methods (GAUS and SMOTE) [8]. GAUS performed better than SMOTE in most cases because it took into account object classification ability and sample size. In four high-dimensional data sets, GAUS gave all four top results, while ROS and RUS had precision and recall of zero. GAUS performed better on GLI85 and COLON, while RUS performed better on CNS. Studies have concluded that random sampling methods are not suitable for high-dimensional imbalanced datasets and that GAUS has better end-to-end classification performance [9].

III .COLLATION OF EXPERIMENTS

- One of the proposed systems, Genetic Algorithm Undersampling (GAUS), was compared with other algorithms including Random Oversampling (ROS), Random Undersampling (RUS), and Synthetic Minority Oversampling technology (SMOTE) [8].
- The parameters of GAUS were set as maximum number of iterations (ite_g) = 200, population size (N_g) = 80, crossover rate = 0.7, mutation rate = 0.3, and mutation rate = 0.25.
- The performance of the four compared algorithms (GAUS, ROS, RUS and SMOTE) was evaluated using F1, Gmean, AUC and Acc scores.
- To ensure a fair comparison between algorithms, Symmetric Uncertainty (SU) feature selection was used before running each algorithm on four multivariate data sets [12].
- We computed the number of best results obtained by the four algorithms for 10 low-dimensional data sets. GAUS performed better compared to other algorithms in terms of F1, Gmean, AUC and Acc scores.
- GAUS performed better than other algorithms on GLI85 and COLON datasets, while ROS and RUS achieved better results on DLBCL and CNS datasets, respectively.
- GAUS is based on V-statistics and uses feature selection to make full use of the information in training samples to improve the performance of the classifier, resulting in improved classification performance compared to other algorithms [11].

IV.PERFORMANCE CHARACTERISTICS AND COMPARISON OF CLASSIFICATION ALGORITHMS

Bias towards the majority class: Imbalanced data is the data that one of the class having sufficiently more instances than the other class. This can lead to majority class bias, where the algorithm tends to classify most instances as belonging to the majority class, which can lead to poor minority class prediction performance [3].

Reduced accuracy: Traditional classification algorithms assume a good distribution of data. However, imbalanced data violates this assumption and can lead to reduced accuracy in the classification results [6].

Inadequate depiction of minority class: Data imbalance can cause minority classes to be underrepresented during training, making it difficult for algorithms to learn patterns for minority classes and make accurate predictions [9].

Overfitting to the majority class: Imbalanced data can cause classification algorithms to over fit to the majority class, as it has more instances for learning. This can result in poor generalization and performance on unseen data [12].

Ineffective feature selection: Imbalanced data can also affect the effectiveness of feature selection algorithms. Traditional random sampling methods may not be suitable for high-dimensional imbalanced datasets, and alternative methods like GAUS (genetic algorithm based under-sampling) that consider feature selection and

https://doi.org/10.46335/IJIES.2024.9.1.15 **e-ISSN: 2456-3463**
Vol. 9, No. 1, 2024, PP. 84-89

International Journal of Innovations in Engineering and Science, www.ijies.net

sampling policies simultaneously may be more effective [9].

Need for specialized algorithms: Imbalanced classification problems require specialized algorithms that are designed to handle the challenges posed by imbalanced data. Traditional classification algorithms may not perform well on imbalanced datasets, and algorithms like IS+FS-MOEA (multi-objective evolutionary algorithm) and SYMON (feature selection algorithm) have been developed specifically for imbalanced classification problems [9].

Generic comparison of data and algorithmic level methods in imbalanced data classification:

Table 1 - Comparison of Data and Algorithm Level Methods

Sr.	Stratification of Imbalanced Data based on Data level Methods	Stratification of Imbalanced Data based on Algorithm Level Methods
1	Data-level methods change the data distribution through undersampling or oversampling.	Data distribution remains unaffected in these methods.
2	Random Undersampling (RUS) is one of the widely used undersampling methods.	Cost sensitive methods are a category of algorithm-level methods that use a cost matrix to assign a higher cost to misclassifying instances of a minority class.
3	Oversampling techniques involve duplicating or creating new instances of minority classes to balance the data.	The one-class learning method obtains similarity values between instances based on the characteristics of the instances, and classifies each instance using a predefined similarity threshold.
4	Cost-sensitive methods generate a cost matrix to assign a higher cost to misclassification of minority class instances compared to majority class instances.	Feature selection methods can shift the focus to features by optimizing the contrast between classes rather than training examples.
5	Undersampling techniques involve balancing the data by selecting some majority class instances and combining them with all minority class instances	Algorithm level methods that set up cost matrix are falling under cost sensitive methods

V.CONCLUSION

Overall, imbalanced data can negatively impact the performance of classification algorithms by introducing bias, reducing accuracy and inadequate representation of the minority class. Specialized algorithms and techniques, such as feature selection and sampling policies, are required to address the challenges posed by imbalanced data. Additionally, feature selection techniques, including feature selection and feature extraction, can be used to understand and optimize contrasts between classes and improve the performance of classifiers when classifying imbalanced data.

ACKNOWLEDGMENT

We would like to extend our sincere thanks to Prof. Vijay D. Chaudhari from GF's GCOE, Jalgaon for their kind cooperation and valuable guidance regarding publishing this research work.

REFERENCES

[1] *Khorshidi H A, Aickelin U., "Constructing classifiers for imbalanced data using diversity optimization", Information Sciences, 2021, 565: 1 -16*

[2] *Abid A. and Zou J., "Concrete auto encoders: Differentiable feature selection and reconstruction", Proceeding of 36th Int. Conf. Mach. Learn., vol. 97, PMLR, Long Beach, CA, 2019, pp. 444–453.*

[3] *Wang H., Li X., Jung C. and Wu C., "A machine learning method for selection of genetic variants to increase prediction accuracy of type 2 diabetes mellitus using sequencing data", Stat. Anal. Data Min. 13 (2020), 261–281.*

[4] *J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective" ACM Computer Survey., vol. 50, no. 6, p. 94, 2018.*

[5] *Q. Wang, Y. Qian, Q. Guo and J. Liang, "Local neighborhood rough set" in Knowl.-Based Syst., vol. 153, pp. 53-64, 2018.*

[6] *Liu Y., Wang Y., Ren X., Zhou H., and Diao X., A classification method based on feature selection for imbalanced data, IEEE Access 7 (2019), 81794–81807.*

[7] *S. Lin, T. Wang, W. Ding, J. Xu and Y. Lin, "Feature selection using Fisher score and multi-label neighborhood rough sets for multi-label classification", Info. Sci., vol.578, pp. 887-912, 2021.*

[8] *L. Sun, X. Zhang, Y. Qian, J. Xu and S. Zhang, "Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification", Inf. Sci., vol. 502, pp. 18-41, 2019.*

[9] *Haoyue Liu, Meng Zhou and Qing Liu, "An Embedded Feature Selection Method for Imbalanced Data Classification", IEEE / CAA J. Autom. Sinica, vol. 6, no. 3, pp. 703-715, May 2019.*

[10] *LI Y, CHAI Y, HU Y., "Review of imbalanced data classification methods", Control and Decision, 2019, 34(4): 673 -688*

[11] *Z. Cai and W. Zhu, "Feature selection for multi-label classification using neighborhood preservation", IEEE / CAA J. Autom. Sinica, vol. 5, no. 1, pp. 320–330, Jan. 2018.*

[12] *Khaldy MAI, Kambhampati C., "Resampling imbalanced class and the effectiveness of feature selection methods for heart failure dataset", Int. Rob Auto J., Feb. 2018*

[13] *Zhou P, Hu G, Li P, "Online feature selection for high-dimensional class-imbalanced data", Knowledge-Based Systems, 2017, 136: 187 -199*

[14] *Sharma R. K. and Chandra B., "Exploring autoencoders for unsupervised feature selection," Int. Joint Conf. Neural Netw. (IJCNN), IEEE, 2015, pp. 1–6*

[15] *D. Ramyachitra and P. Manikandan, "Imbalanced dataset classification and solutions: a review," Inter. J. Computing and Business Research (IJCBR), vol. 5, no. 4, Jul. 2014.*