Optimizing Classification Performance on Imbalanced Multi-Class Data using MDO

Dr Shivprasad[#1], K Sai Kiran[*2]

[#]*Dept of Computer Science, RYMEC Ballari, Visvesvaraya Technological University*

Abstract— **Class imbalance is a common challenge in multi-class classification tasks, where certain classes have significantly fewer instances than others, leading to biased model learning. Biased models when used in machine learning algorithms do not produce accurate results. In this paper, we propose an approach to address the imbalanced multi-class classification problem using the Mahalanobsis Distance based Over-sampling technique (MDO) technique. MDO is a powerful optimization method that combines the decision-making processes of multiple classifiers, effectively leveraging their strengths to improve overall performance. To deals with the emerging issues arising from multi-class skewed distributions, this paper presents use of over-sampling technique on the classes which have skewed data. In this technique the synthetic samples generated which have the same Mahalanobsis distance from their corresponding class mean. To evaluate the effectiveness of our approach, we conduct extensive experiments on diverse imbalanced multi-class datasets, comparing the performance of the MDO-based classifier against several baseline methods. The experimental results demonstrate substantial improvements in classification accuracy, precision, recall, and F1-score achieved by our MDO-based approach. Furthermore, the proposed method exhibits remarkable robustness across various imbalanced scenarios, outperforming traditional techniques by a significant margin. The findings of this study contribute to the field of imbalanced multi-class classification by providing a principled and effective solution using the MDO technique. Our approach not only enhances classification performance but also helps to address the inherent bias caused by class imbalance. The proposed method holds great promise for real-world applications in domains where imbalanced multi-class data is prevalent.**

Keywords— **Synthetic values, Optimizing dataset, Machine learning algorithms, Imbalanced data, Performance optimization**

I. Introduction

Data mining involves extracting valuable information from large sets of raw data through advanced mathematical algorithms, surpassing basic analysis to reveal patterns and predict future events. This process is crucial in diverse fields like business intelligence, finance, healthcare, and marketing. Data preprocessing is an essential data mining technique that transforms raw data into a usable format, resolving issues of data incompleteness, inconsistency, and errors, thus ensuring high-quality data for analysis. Class imbalance is a significant challenge in data preprocessing, occurring when certain classes have more instances than others, leading to biased learning algorithms. Addressing this, under-sampling and over-sampling techniques are used to balance the data distribution, improving classifier performance. Classification, another vital data mining function, involves categorizing items into specific classes, facilitating data-driven decision-making in applications like sentiment analysis, fraud detection, and customer segmentation. Data quality is critical for successful data mining and is achieved through effective preprocessing, ensuring data accuracy, completeness, consistency, timeliness, believability, and interoperability. The objectives of this study are threefold. Firstly, to reduce dataset dimensions using Principal Component Analysis (PCA) for improved performance and efficiency. Secondly, to overcome over-generalization issues and enhance classification accuracy in machine learning algorithms. Lastly, to address multi-class imbalance using the Mahalanobis Distance based over-sampling (MDO) technique, creating a more balanced and representative dataset. In conclusion, data mining and preprocessing play crucial roles in extracting insights from raw data and supporting data-driven decision-making. Overcoming class imbalance and ensuring data quality are essential steps in enhancing classification performance. The study's objectives contribute to the advancement of data mining and machine learning techniques

II. LITERATURE SURVEY

The class imbalance problem has been addressed through various approaches. With over-sampling techniques being widely adopted and often independently chosen by classifiers.

Chawla and Bowyer (2002) introduced SMOTE (Synthetic Minority Over-sampling Technique), generating synthetic samples based on similarities between minority class samples. However, SMOTE may lead to decision boundary overlapping in multi-class problems.

Bunkhumpornpat and Lursinsap (2009) proposed Safe-level SMOTE, a variant of SMOTE that considers the number of positive instances in K-nearest neighbors to differentiate between noise and safe levels.

Fan and Tang (2011) introduced MYSN, combining the large margin principle with SMOTE to address the over-generalization problem by lower bounding the sample-margin and hypothesis margin of instances.

Puntumapon and Waiyamai (2012) presented TRIM, which filters irrelevant majority data iteratively, focusing on precise minority class regions and generating synthetic samples via SMOTE for two-class problems.

He and Bai (2008) devised ADASYN (Adaptive Synthetic Sampling), which adaptively creates synthetic samples based on the distribution of the data. However, ADASYN may suffer from the over-generalization problem.

Fernandez and Lopez (2013) explored pair-wise and ad hoc procedures for multi-class imbalanced problems, utilizing class decomposition schemes like one versus one (OAO) and one versus all (OAA) to apply two-class methods. OAO may become computationally expensive with a high number of classes, and OAA can exacerbate imbalance issues.

Fernandez-Navarro and Hervas-Martinez (2011) proposed a dynamic over-sampling method that employs a memetic algorithm to optimize radial basis functions neural networks for imbalanced datasets with more than two classes.

Lin and Tang (2013) introduced DyS, a dynamic sampling method for multilayer perceptron’s, which selects informative data for training the network at each iteration.

Chen and He (2011) developed Mu-SERA, an ensemble learning method inspired by SERA, which generates multiple hypotheses based on data chunks and weighs them to classify testing data using Mahalanobis distance evaluation.

In conclusion, researchers have devised various over-sampling techniques and dynamic sampling methods to tackle the class imbalance problem, each with its advantages and limitations. These approaches contribute to the advancement of handling imbalanced datasets in diverse machine learning applications.

III. PROBLEM STATEMENT

The problem at hand pertains to imbalanced datasets, where there is an unequal number of instances for different classes, impacting the accuracy of classification algorithms. Current solutions to address this issue operate either at the data level through preprocessing techniques or at the algorithmic level, involving one-class learning, ensemble learning algorithms, and cost-sensitive methods. However, these existing approaches face challenges when dealing with multi-class imbalance. Over-sampling techniques, which generate synthetic samples based on similarities between data points or near boundaries of minority samples, may lead to overlapping class regions and potential over-generalization problems in multi-class scenarios. Moreover, the performance of current over-sampling techniques is not fully evaluated for multi-class problems, and they may experience inefficiencies in scenarios with high-dimensional input data due to lengthy computations when selecting appropriate random numbers for different feature values. Hence, the objective is to develop a more effective and efficient solution to handle the class imbalance problem in multi-class datasets.

IV. SYSTEM ARCHITECTURE

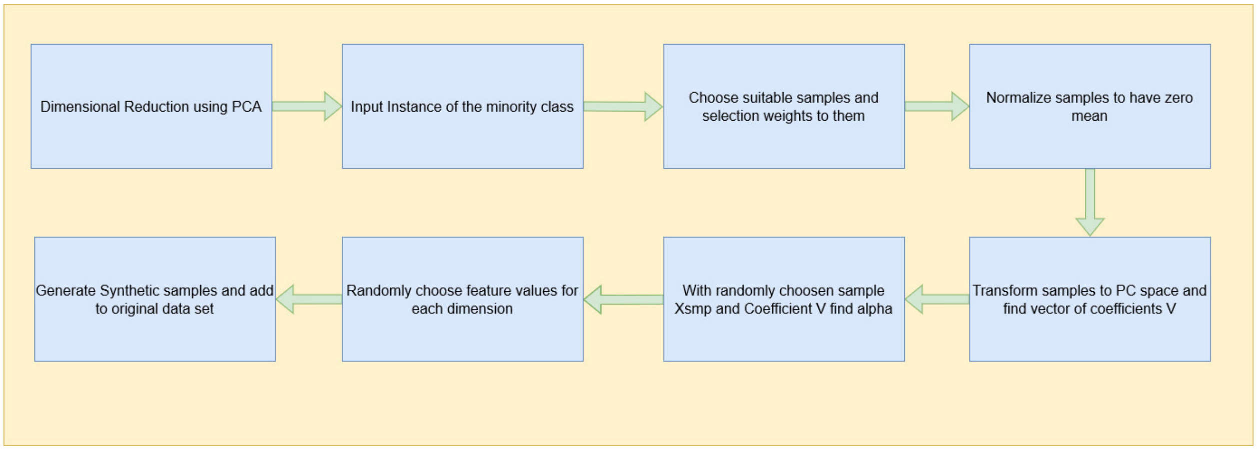
Dimensional Reduction using PCA
Input Instance of the minority class
Choose suitable samples and selection weights to them
Normalize samples to have zero mean
Transform samples to PC space and find vector of coefficients V
With randomly choosen sample Xsmp and Coefficient V find alpha
Randomly choose feature values for each dimension
Generate Synthetic samples and add to original data set

FIG: SYSTEM BLOCK DIAGRAM

The diagram depicted above illustrates the architectural blueprint employed for the implementation of the MDO technique, meticulously designed to grapple with multi-class imbalanced predicaments. In the primary phase, the dataset's dimensions undergo reduction, meticulously orchestrated through the integration of Principal Component Analysis (PCA), all in pursuit of bolstering the learning algorithm's efficacy. This culminates in a curated subset of columns, which subsequently interface with the MDO oversampling methodology.

Within the precincts of the "choose sample" function, the calculated Euclidean distances for each sample set the stage. For instances sharing a common class label, their K-nearest neighbors are meticulously identified. Guided by this framework, selection weights for each sample crystallize, extending the process back to the MDO technique. The returned parameters are then harnessed to endow each sample with a zero-mean configuration, subsequently undergoing transformation into Principal Components (PC) through eigenvalue decomposition. The over-sampling rate surfaces as a product of these computations.

The crux of the over-sampling endeavor lies within the over-sampling function. Here, the architecture harnesses the prowess of the ellipse equation to engender novel synthetic samples, each meticulously curated to adhere to the original sample space. This harmonization effort, blending the synthetic and original samples, is engineered with a singular objective: to rectify class imbalances and foster equilibrium within the data distribution.

In essence, the MDO Over-sampling Function serves as a virtuoso performance, seamlessly blending the artistry of statistical calculations with the finesse of mathematical constructs. Its symphony of data augmentation, harmonization, and equilibrium propagation reshapes the contours of class imbalance, ushering forth a new era of data-driven excellence.

VI CONCLUSION

Hence, the innovative approach presented in the form of MDO (Mahalanobis distance-based over-sampling technique) takes center stage as a potent enhancer of learning efficacy within the intricate realm of multi-class imbalanced datasets. A transformative strategy is set in motion, harnessing the prowess of PCA (Principal Component Analysis) to streamline dimensions, culminating in an orchestration that harmonizes class boundaries and redefines learning dynamics.

MDO emerges as a vigilant guardian against the perils of class overlapping, effectively curating a landscape where classes retain their distinct identities. Notably, MDO charts a unique trajectory by not only elevating the generalization prowess of classifiers but also tempering the risks of over-fitting and over-generalization – an attribute that distinguishes it from its contemporaries.

In essence, the MDO methodology unveils an innovative narrative, where Mahalanobis distance and over-sampling converge to nurture an environment conducive to amplified learning accuracy and refined model behavior. Through this dynamic interplay, MDO forges a novel path towards equilibrium, embodying a pioneering stride in the ongoing quest for optimal data-driven solutions.

REFERENCES

[1] S. M. Metev and V. P. Veiko, *Laser Assisted Microtechnology*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.

[2] J. Breckling, Ed., *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.

[3] S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, “A novel ultrathin elevated channel low-temperature poly-Si TFT,” *IEEE Electron Device Lett.*, vol. 20, pp. 569–571, Nov. 1999.

[4] M. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, “High resolution fiber distributed measurements with coherent OFDR,” in *Proc. ECOC’00*, 2000, paper 11.3.4, p. 109.

[5] R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, “High-speed digital-to-RF converter,” U.S. Patent 5 668 842, Sept. 16, 1997.

[6] (2002) The IEEE website. [Online]. Available: http://www.ieee.org/

[7] M. Shell. (2002) IEEEtran homepage on CTAN. [Online]. Available: http://www.ctan.org/tex-archive/macros/latex/contrib/supported/IEEEtran/

[8] *FLEXChip Signal Processor (MC68175/D)*, Motorola, 1996.

[9] “PDCA12-70 data sheet,” Opto Speed SA, Mezzovico, Switzerland.

[10] A. Karnik, “Performance of TCP congestion control with rate feedback: TCP/ABR and rate adaptive TCP/IP,” M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan. 1999.

[11] J. Padhye, V. Firoiu, and D. Towsley, “A stochastic model of TCP Reno congestion avoidance and control,” Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.

[12] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification*, IEEE Std. 802.11, 1997.