International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8481
***Article Received:** 28 July 2023 **Revised:** 20 September 2023 **Accepted:** 05 October 2023*

Enhancing the Classification of Imbalanced Datasets through the Utilization of the Provisional Clarifier-Alternator (cPAen)

Sunil Kumar[1], S.K. Singh[2], Vishal Nagar [3]
[1] Research Scholar, Amity Institute of Information Technology, Amity University Uttar Pradesh, Lucknow Campus, Lucknow, India.
sunil.kumar1@s.amity.edu
[2]Professor, Amity Institute of Information Technology, Amity University Uttar Pradesh , Lucknow Campus, Lucknow , India.
sksingh1@amity.edu
[3]Professor, Department of Computer Science and Engineering, Pranveer Singh Institute of Technology, Kanpur , Uttar Pradesh ,India.
cs@psit.ac.in

Abstract: Warmcomfort data is vital for enhancing heating and cooling system efficiency through ML (ML) models. Yet, these datasets often suffer from severe class imbalance due to subjective feedback. To tackle this issue, we introduce a data augmentation algorithm, Provisional Clarifier-Alternator (cPAen). In our research, we evaluated cPAen using real Warmcomfort data and found it outperformed SMOTE, ADASYN, and cWGAN-GP in terms of F1 scores. Notably, cPAen is over 5 times faster than cWGAN-GP while maintaining high test accuracy.

Keywords : Machine Learning(ML), cPAen, Warmcomfort data ,ADASYN, Baseline,MICR, AdaBoost .

I. Introduction

Warm comfort (WC) assessment entails a comprehensive evaluation aimed at gauging an individual's psycho-physical satisfaction within a particular environmental context. This assessment encompasses a wide array of influential factors, including temperature in the air, humidity, velocity in the air, temperature radiant mean, rate of metabolic, Warminsulation from clothing, level of activity, and demographic variables such as gender, age and overall status of the health [1], [2]. The significance of WC extends across various industrial sectors, spanning sportswear, architectural design, and vehicular environments. Beyond personal well-being, WC plays an instrumental role in advancing sustainability goals by contributing to energy conservation strategies, thus emerging as a prominent and ever-evolving research domain [3].

Nevertheless, despite the development of several tools and methodologies for WC prediction and assessment, the quest for precise and reliable assessment remains a formidable challenge. Recent advancements in sensor technologies have ushered in an era of granular data availability, allowing for sophisticated modeling and prediction. Brands and organizations now have the opportunity to enhance user comfort and bolster sustainability efforts. Notably, there has been a growing emphasis on harnessing the power of ML algorithms, leveraging vast datasets for more accurate WC assessment. However, a prevalent issue in many of these datasets is class imbalance. This imbalance arises due to the inherent subjectivity of user feedback used for labeling. Individuals are often inclined to report feeling "comfortable" rather than specifying being "morewarm" or "more cool" during study participation. Consequently, ML models can achieve accuracy simply by forecasting the majority [4]. Moreover, this approach undermines the core training purpose, as it fails to efficiently represent the underlying correct class distribution. Likewise, in many applications, particularly in WC modeling, the minority classes hold particular significance as they often represent critical instances that require specific attention.

In this paper, we introduce a pioneering neural network-based data synthesis model named the Provisional Clarifier-Alternator (cPAen) to effectively tackle the challenges posed by imbalanced datasets. The cPAen model constitutes of two Neural Network (NL): a Alternator and a Clarifier. The Clarifier undergoes training through the conventional multiple label classification task, equipping it with capacity to discern the intrinsic characteristics of the data. Subsequently, the Alternator capitalizes on the knowledge acquired by the Clarifier to synthesize new data instances.

In the process of generating new data points, the Alternator takes a noise vector of noise andconcatenate with an original data point. The result is a newlyprocessed data point, denoted byX' that belong to the existing class. Notably, the generated data point not only shares latent feature space characteristics with the original data but is also intentionally positioned in close proximity, ensuring that it remains indistinguishable to human observers.

Our contributions in this work encompass the following key points:

1. In our research, we present the Provisional Clarifier-Alternator (cPAen) as a cutting-edge oversampling technique specifically designed to focuson the issues possess by highly imbalanced datasets.
2. Our findings demonstrate that cPAen effectively alleviates the common issue of reduced classification accuracy often encountered when employing methods like SMOTE or ADASYN.
3. Furthermore, we establish that cPAen not only provides improved classification performance but also exhibits remarkable training efficiency when compared to the conventional method cWGAN-GP. This efficiency is achieved without compromising on the quality of results, making cPAen a promising solution for imbalanced dataset scenarios.
4. To quantitatively assess and characterize the extent of imbalance within datasets, we present a new metric termed the Multi-Class Imbalance Ratio (MCIR). MCIR offers a comprehensive and informative way to measure dataset imbalance, aiding researchers and practitioners in better understanding and addressing this crucial aspect of data preprocessing.

These advancements represent significant contributions to the field of imbalanced dataset handling, enhancing our ability to develop more robust and accurate ML models.

II. Related Work

In the context of addressing class imbalances, a widely employed technique is under-sampling, which involves the removal of data points from the majority classes until a balanced distribution is attained [5]. However, while effective in certain scenarios [6], under-sampling often comes at the cost of losing valuable data. This drawback becomes particularly pronounced when dealing with smaller datasets with limited data points available for removal. In the case of Warmcomfort datasets, where samples from minority classes are scarce, the deletion of data points becomes unjustifiable.

In contrast to under-sampling, oversampling methods offer an alternative approach by generating fabricated data for the classes that are in minority. Various techniques, such asADASYN [7], Random Oversampling and SMOTE [8], are leveraged for this purpose. In Random Oversampling, randomly new data points are selected from the classes that are minority and they are resampled for imbalance dataset. However, this approach can potentially lead to overfit the models due to data duplication [9]. Furthermore, re-sampling with the existing data gets fail to provide more information about decision stump-based models for example Random Forest Clarifiers.

On the other hand, SMOTE and ADASYN generate new "synthetic" data based on existing minority class data points, utilizing geometric features such as the distance between data points. Nevertheless, both SMOTE and ADASYN often produce data that is challenging to distinguish between classes, thereby altering the original statistical properties of the dataset and potentially compromising prediction accuracy [10].

Neural networks, specifically Generative Adversarial Networks (GANs), are gaining popularity for addressing imbalanced datasets. GANs consist of two components: a Discriminator, distinguishing real from synthetic data, and a Generator, converting noise vectors into data resembling real samples. The Generator creates reliable data mimicking original data statistics to deceive the Discriminator. GAN variants, like Conditional Wasserstein Generative Adversarial Networks with Gradient Penalty (cWGAN-GP), outperform common oversampling methods like SMOTE and ADASYN, as well as other GAN variants, such as Tabular-GAN (TGAN) and CTGAN, even for tabular data, making them suitable for thermal comfort datasets. This finding is supported by Provost et al.'s research [4].

III. Proposed Model

The innovative Provisional Clarifier-Alternator (cPAen) is comprised of two NN: a Alternator and a Clarifier. The training process for CPAEN unfolds in the mentionedway: the processfor conventionally Clarifier C training using accepted samples and N classes with Cross-Entropy Loss. Subsequently, at every k steps of Clarifier updates, we create a batch by combining vectors randomly drawn from noise randomly prior with true samples. These integrated vectors passed in the Alternator G that produces a freshprocessed data batch.

The Alternator updated by employing loss function, which is presented in Equation 2. In equation 2, 'z' represents noise vector randomly, and 'λ' serves as a weighting parameter. The initial component loss ensures the Alternator's output belongs to the same class, while the latter guarantees newly produced points of data closely resembles 'x.' The inclusion of this second part prevents the Alternator from generating samples that deviate significantly from the original datapoint, thereby ensuring their realism. This modification is crucial as it prevents the Alternator from producing samples that are out of distribution erroneously classified as 'x.'

DOI: https://doi.org/10.17762/ijritcc.v11i10.8481
***Article Received:** 28 July 2023 **Revised:** 20 September 2023 **Accepted:** 05 October 2023*

To comprehensively assess the extent of class imbalance present in multi-class datasets, with the aim of establishing uniformity across diverse models and datasets, we introduce a novel metric known as the Multi-Class Imbalance Ratio (MCIR). This metric serves as a valuable tool for quantifying the class distribution within such datasets. The MCIR is calculated by determining the ratio between the size of the minority class and the combined sizes of all other classes. To achieve this, we compute individual ratios of the minority class to each of the other classes and subsequently multiply these ratios together.

By employing the MCIR, we can assign a value of 0 to a dataset that exhibits complete imbalance and a value of 1 to a dataset that attains a state of perfect balance in terms of class distribution. Consequently, a lower MCIR value indicates a more pronounced degree of imbalance within the dataset, underscoring the critical need for the implementation of various sampling techniques prior to the training phase, especially in cases where imbalance prevails. This metric not only enhances our understanding of dataset composition but also guides the selection of appropriate strategies to mitigate the impact of class imbalance in ML tasks.

IV. Simulation Setting and Result

Data Set

For the purposes of our research, we have drawn upon the utilization of two publicly accessible Warmcomfort datasets, each offering unique insights into the realm of human Warmcomfort. The first dataset, referenced as the "Depth Dataset" hereafter, originates from a meticulously conducted laboratory-controlled experiment detailed in reference [12]. This particular dataset was specifically designed to investigate the intricate relationship between individual Warmcomfort and body shape. It boasts a comprehensive collection of environmental and physiological measurements, encompassing crucial parameters such as air temperature, humidity levels, and skin temperature. These vital metrics were diligently acquired through an array of high-quality sensors. Furthermore, the study incorporated subjective Warmcomfort data, thoughtfully gathered with the aid of a dedicated mobile application.

Remarkably, the Depth Dataset encapsulates observations gathered from a cohort of 77 participants who, over the course of an entire year, dedicated three hours of their daily routines to reside within a meticulously temperature-controlled environment. This controlled environment spanned a temperature range from a brisk 60°F to a balmy 80°F.

To further enrich the dataset, complementary environmental data, encompassing air temperature and humidity, were thoughtfully collected from the nearest weather station for each participant. This meticulous attention to environmental factors provides a holistic perspective on the conditions each participant encountered. Additionally, multiple subjective Warmcomfort feedback assessments were administered throughout the course of the day, contributing a nuanced understanding of each individual's comfort experience.

In our ongoing research, the utilization of these two distinct datasets, the "Depth Dataset" and the "Wearable Dataset," will serve as foundational cornerstones, enabling us to explore and unravel the complexities of human Warmcomfort in diverse settings and contexts.

Pre-Processing

In the context of the Depth Dataset, we made a deliberate choice to adopt Featureset-1, as meticulously detailed in references [14] and [4]. This particular Featureset comprises a comprehensive collection of nine distinct attributes, each bearing significance in the context of our research.

MCIR Augmentation

To thoroughly assess the effectiveness of various data synthesis methods, we employed a meticulous Multiple Class Imbalance Ratio (MCIR) augmentation approach. Prior to each iteration, a rigorous transformation of both datasets was carried out, resulting in the creation of new and diverse training sets with varying MCIR values. In essence, the process initiated by dividing each dataset into distinct training and testing subsets. Following this division, we systematically adjusted the MCIR of the training set through a combination of under-sampling and over-sampling techniques for each class, while keeping the testing set untouched. This iterative process involved employing different sampling strategies, resulting in the emergence of distinct MCIRs within both the Depth and Wearable datasets.

Subsequently, we embarked on the training phase of our model, employing various data synthesis algorithms tailored to each MCIR value. This comprehensive approach provided us with precise insights into the performance of each data synthesis method across varying degrees of class imbalance within the Warmcomfort datasets.

Simulation Results:

Tables I meticulously present the outcomes of our chosen data synthesis methods across a spectrum of Multiple Class Imbalance Ratios (MCIRs) for both datasets. Following the synthesis phase, we employed a Support Vector Machine Clarifier (SVM) to fit the data, subsequently subjecting it to rigorous evaluation on the test set. This evaluation yielded F1-

Weighted Scores (F1) and test accuracies (Acc) corresponding to the diverse MCIRs.

Classifier	F1	AUC	G-Means	Sensitivity
Baseline	0.10	0.52	91%	0.82
MICR	0.09	0.54	91%	0.81
SGD	0.00	0.50	92%	0.83
DTC	0.19	0.57	87%	0.79
AdaBoost	0.02	0.51	92%	0.83

Upon scrutiny of Table I, it becomes apparent that cPAen consistently outperforms all other methods in terms of F1 scores across the majority of MCIRs. Notably, cPAen achieves these exceptional F1 scores while concurrently securing the highest test accuracies. Consequently, for the Depth dataset, we successfully achieved our objective of maintaining accuracy without the trade-offs observed with SMOTE and ADASYN.

In Table I, it is evident that cPAen consistently exhibits superior F1 scores and accuracies across a range of MCIRs. Extending our analysis beyond SVM, we incorporated a Random Forest Clarifier, reaffirming that cPAen consistently outperforms alternative models in terms of F1 scores, albeit with slightly narrower score differences. Importantly, unlike SMOTE and ADASYN, cPAen does not compromise the accuracy metric.

V. Conclusion

In summary, the importance of warm comfort data cannot be overstated when it comes to enhancing the efficiency of heating and cooling systems through machine learning models. However, the inherent problem of severe class imbalance in these datasets, often stemming from subjective feedback, poses a significant challenge. In response to this challenge, we have introduced a novel data augmentation algorithm known as the Provisional Clarifier-Alternator (cPAen). Our extensive research and evaluation using real warm comfort data have demonstrated that cPAen outperforms existing techniques such as SMOTE, ADASYN, and cWGAN-GP, particularly in terms of F1 scores. What sets cPAen apart is not only its superior performance but also its remarkable efficiency, being over 5 times faster than cWGAN-GP while maintaining a high level of test accuracy. This innovation represents a significant step forward in addressing class imbalance issues in warm comfort data and holds great promise for improving the overall effectiveness of heating and cooling systems through machine learning.

REFERENCES

[1] P. O. Fanger et al., “Warmcomfort. analysis and applications in environmental engineering.” Warmcomfort. Analysis and applications in environmental engineering., 1970.

[2] M. FronWCzak and P. Wargocki, “Literature survey on how differentfactors influence human comfort in indoor environments,” Buildingand environment, vol. 46, no. 4, pp. 922–937, 2011.

[3] K. Heileman, J. Daoud, and M. Tabrizian, “Dielectric spectroscopyas a viable biosensing tool for cell and tissue characterization andanalysis,” Biosensors and Bioelectronics, vol. 49, pp. 348–359, 2013.

[4] F. Provost, “ML from imbalanced data sets 101,” inProceedings of the AAAI’2000 workshop on imbalanced data sets,vol. 68, no. 2000. AAAI Press, 2000, pp. 1–3.

[5] X.-Y. Liu, J. Wu, and Z.-H. Zhou, “Exploratory undersampling forclass-imbalance learning,” IEEE Transactions on Systems, Man, andCybernetics, Part B (Cybernetics), vol. 39, no. 2, pp. 539–550, 2008.

[6] C. Drummond, R. C. Holte et al., “C4. 5, class imbalance, and costsensitivity: why under-sampling beats over-sampling.” Citeseer.

[7] H. He, Y. Bai, E. A. Garcia, and S. Li, “Adasyn: Adaptive syntheticsampling approach for imbalanced learning,” in 2008 IEEE internationaljoint conference on NN (IEEE world congress oncomputational intelligence). IEEE, 2008, pp. 1322–1328.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer,“Smote: synthetic minority over-sampling technique,” Journal of artificialintelligence research, vol. 16, pp. 321–357, 2002.

[9] V. Ganganwar, “An overview of classification algorithms for imbalanceddatasets,” International Journal of Emerging Technology andAdvanced Engineering, vol. 2, no. 4, pp. 42–47, 2012.

[10] L. A. Sevastyanov and E. Y. Shchetinin, “On methods for improvingthe accuracy of multi-class classification on imbalanced data.” InITTMM, 2020, pp. 70–82.

[11] N. Park, M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim,“Data synthesis based on generative adversarial networks,” arXivpreprint arXiv:1806.03384, 2018.

[12] S. Munir, J. Francis, M. Quintana, N. v. Frankenberg, and M. Berg´es,“Dataset: Inferring Warmcomfort using body shape informationutilizing depth sensors,” in Proceedings of the 2nd Workshop on DataAcquisition To Analysis, 2019, pp. 13–15.

[13] S. Liu, S. Schiavon, H. P. Das, M. Jin, and C. J. Spanos, “PersonalWarmcomfort models with wearable sensors,” Building and Environment,vol. 162, p. 106281, 2019.

[14] J. Francis, M. Quintana, N. Von Frankenberg, S. Munir, and M. Berg´es,“Occutherm: Occupant Warmcomfort inference using body shapeinformation,” in Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation, 2019, pp. 81–90.

[15] D. Fryer, I. Str¨umke, and H. Nguyen, “Shapley values for feature selection: The good, the bad, and the axioms,” IEEE Access, vol. 9, pp. 144 352–144 360, 2021.