Conditional Synthetic Data Generation for Personal Thermal Comfort Models

Hari Prasanna Das and Costas J. Spanos
{hpdas,spanos}@berkeley.edu
Department of Electrical Engineering and Computer Sciences, UC Berkeley

ABSTRACT

Personal thermal comfort models aim to predict an individual's thermal comfort response, instead of the average response of a large group. Recently, machine learning algorithms have proven to be having enormous potential as a candidate for personal thermal comfort models. But, often within the normal settings of a building, personal thermal comfort data obtained via experiments are heavily class-imbalanced. There are a disproportionately high number of data samples for the "Prefer No Change"class, as compared with the "Prefer Warmer"and "Prefer Cooler"classes. Machine learning algorithms trained on such class-imbalanced data perform sub-optimally when deployed in the real world. To develop robust machine learning-based applications using the above class-imbalanced data, as well as for privacy-preserving data sharing, we propose to implement a state-of-the-art conditional synthetic data generator to generate synthetic data corresponding to the low-frequency classes. Via experiments, we show that the synthetic data generated has a distribution that mimics the real data distribution. The proposed method can be extended for use by other smart building datasets/use-cases.

CCS CONCEPTS

• **Computing methodologies → Neural networks**.

KEYWORDS

Thermal Comfort, Conditional Synthetic Data Generation, Class Imbalance

ACM Reference Format:
Hari Prasanna Das and Costas J. Spanos. 2022. Conditional Synthetic Data Generation for Personal Thermal Comfort Models. In *DLP '22: Workshop on Deep Learning Practice and Theory for High-Dimensional Sparse and Imbalanced Data, August 15, 2022, Washington DC, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Humans spend more than 90% of their day indoors, where their well-being, performance and energy consumption are demonstrably linked to thermal comfort. But, study shows that only 40% of commercial building occupants are satisfied with their thermal environment [10]. There has been significant amount of research done to develop models to accurately predict thermal comfort metrics for occupants in a building. Contrary to conventional group-based thermal comfort models, personal thermal comfort models [13] focus on developing thermal comfort predictors at a building occupant level. They have proved efficient in human-centric cyber-physical systems to efficiently regulate the building control systems, as well as to understand the correlation between human factors affecting comfort. The general process is to conduct experiments with human subjects and collect their physiological signals along with other environmental parameters, and thermal sensations and preference. Then prediction models are trained to predict the thermal preference that governs the thermal comfort management actuators/ controllers. Recently, machine learning models have been introduced to successfully predict thermal comfort.

In real life, often the thermal comfort data obtained is highly class-imbalanced. For instance, in the experiment in Liu et al. [14], on an average for each subject, around 65% of the data belonged to the "Prefer No Change"class, and the rest equally divided between the "Prefer Warmer"and "Prefer Cooler"classes. Machine learning algorithms require high amounts of varied data for efficient performance. Under such class-imbalance, machine learning algorithms perform sub-optimally. In case of buildings, having access to significant amounts of real data for the low-frequency classes, with human subjects is hard and expensive. To balance the classes, recent works have proposed undersampling the high-frequency class to match the count with low-frequency classes, or oversampling the low-frequency classes to match with the high-frequency class. In the former method, there is loss of information, which is undesirable, and in the latter case, there is possibility of overfitting. Another challenge that is faced comes from the concern of privacy. Often, sharing of thermal comfort data that are associated with users in a building face the challenge of privacy issues.

To deal with the above challenges, we propose to generate conditional synthetic data for personal thermal comfort models. We propose to use the conditional generative models proposed in Das et al. [5] to generate synthetic data for the "Prefer No Change", "Prefer Warmer"and "Prefer Cooler"classes. The inputs to the generative model are thermal comfort features including physiological signals, temperature, humidity, clothing, activity levels, external parameters etc. The model is capable of extracting the feature representations corresponding to the individual classes, and also to generate new synthetic data keeping the conditional feature representation intact and changing the local noise. Our results show that the proposed model is able to generate synthetic data that mimic the real data.

DLP '22, August 15, 2022, Washington DC, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

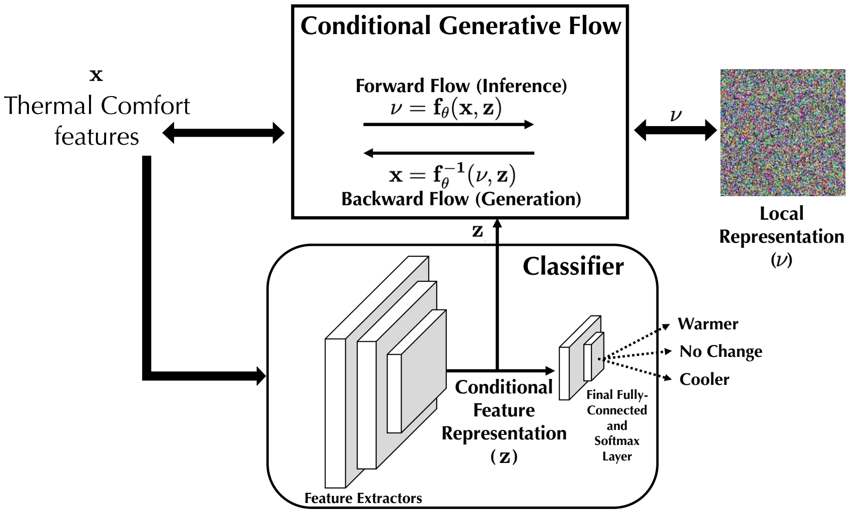
x
Thermal Comfort features
Conditional Generative Flow
Forward Flow (Inference)
$\nu = \mathbf{f}_\theta(\mathbf{x},\mathbf{z})$
$\mathbf{x} = \mathbf{f}_\theta^{-1}(\nu,\mathbf{z})$
Backward Flow (Generation)
ν
Local Representation (ν)
z
Classifier
Warmer
No Change
Cooler
Conditional Feature Representation (z)
Final Fully-Connected and Softmax Layer
Feature Extractors

Figure 1: Illustration of the proposed conditional synthetic generation. (Best viewed in color)

2 RELATED WORKS

Synthetic data generation has been proposed to expand the diversity and amount of the existing training data in many different fields, often to improve the robustness of machine learning models. A few examples are as following. In healthcare, Ghorbani et al. [8] propose a generative adversarial network (GAN [9, 19])-based synthetic data generator to improve the diversity and the amount of skin lesion images. Kohlberger et al. [12] synthesize pathology images for cancer with realistic out-of-focus characteristics to evaluate general pathology images for focus quality issues. Han et al. [11] propose synthetic generation to produce high-resolution artificial radiographs. For privacy-preserving data sharing, Xu et al. [17] propose a method to model tabular data to enable their synthetic generation. In computer vision Das et al. [4] propose synthetic data generation across multiple domains. In smart buildings, Quintana et al. [16] used a conditional tabular GAN based model for thermal comfort synthetic data generation. We use a state-of-the-art conditional synthetic data generation model that has shown improved results over all baselines to generate thermal comfort synthetic data.

3 METHODS

3.1 Thermal Preference Classifier

Our model is based on the method proposed in [5]. Suppose we have N samples $\mathbf{X}$ with labels Y, with 3 possible thermal preference classes, Warmer/No Change/Cooler. We first train a classifier C (consisting of a feature extractor network denoted by $g(\cdot)$, and a final fully-connected and softmax layer, denoted by $h(\cdot)$, i.e. $C(x) = h(g(x))$) to classify the input sample (which in our case are thermal comfort features) and associated labels as Warmer/No Change/Cooler. Mathematically, this step solves the following minimization with backpropagation:

$$\min_C \mathcal{L}_C(\mathbf{X}, Y) = -\mathbb{E}_{(x,y)\sim(\mathbf{X},Y)} \sum_{l=1}^{2} \left[\mathbb{I}_{[l=y]} \log C(x))\right] \qquad (1)$$

By virtue of the training process, the classifier learns to discard local information and preserve the features necessary for classification (conditional information) towards the downstream layers. Once the classifier is trained, we freeze its parameters, and use it to extract the conditional (Warmer/No Change/Cooler) feature representation $z = g(x)$ (as a vector without spatial characteristics) at the output of the feature extractor network for input image x. The dimension of z is chosen such that $\dim(z) << \dim(x)$.

3.2 Conditional Generative Flow

During the training phase for the flow model, the conditional feature representation z is fed to the conditional generative flow. The flow model is trained using maximum-likelihood, transforming x to its local representation ν, i.e.

$$f_\theta(x, z) = \nu \sim \mathcal{N}(0, I) \qquad (2)$$

with ν having the same dimension as x by the inherent design of flow models. We use the method introduced by Das et al. [5], Ma et al. [15] to incorporate the conditional input z in flow model. Coupling layers in affine flow models have scale $(s(\cdot))$ and shift $(b(\cdot))$ networks [2, 6], which are fed with inputs after splitting, and their outputs are concatenated before passing on to the next layer. We incorporate the conditional information z in the scale and shift networks. Mathematically, (with x as the input, D as input dimension, d as the split size,and y as output of the layer),

$$\begin{aligned} x_{1:d}, x_{d+1:D} &= \text{split}(x) \\ y_{1:d} &= x_{1:d} \\ y_{d+1:D} &= s(x_{1:d}, z) \odot x_{d+1:D} + b(x_{1:d}, z) \\ y &= \text{concat}(y_{1:d}, y_{d+1:D}) \end{aligned}$$

Table 1: Thermal Preference classification performance with classifiers trained on real and synthetic data. The first number among the pair in each box is performance with a classifier trained on real data, while the second number is with a classifier trained on synthetic data generated by our proposed model.

		Subject ID				
		1	2	3	4	5
Classification Metrics	Cohen's Kappa	28.77%/27%	24.59%/23.12%	19.23%/17.91%	33.65%/31.78%	18.37%/15.49%
	Accuracy	84.3%/79.56%	79.22%/75.76%	63.47%/59.03%	77.19%/77.01%	63.22%/61.42%
	AUC	0.81/0.79	0.8/0.77	0.67/0.62	0.78/0.77	0.76/0.74

Since flow models are bijective mappings, the exact x can be reconstructed by the inverse flow with z and v as inputs. During the generation phase, for an input sample x, we compute the conditional feature representation z. Keeping the conditional feature representation the same, we sample a new local representation $\tilde{v}$, and generate a conditional synthetic sample $\tilde{x}$, i.e.

$$\tilde{v} \in \mathcal{N}(0, I), \quad \tilde{x} = f_\theta^{-1}(\tilde{v}, z) \tag{3}$$

Here, $\tilde{x}$ has the same conditional (Warmer/No Change/Cooler) features as x , but has a different local representation. An illustration of the proposed model is provided in Fig. 1.

4 EXPERIMENTS AND RESULTS

In [14], authors conducted an experiment to collect physiological signals (e.g., skin temperature at various parts of the body, heart rate) of 14 subjects (6 female and 8 male adults) and environmental parameters (e.g., air temperature, relative humidity) for 2–4 weeks (at least 20 h per day). The subjects also took an online survey, where they reported their thermal sensation (on a scale of -3 to +3) and thermal preference (Warmer, Cooler, No Change) among other parameters.

For this work, we generated synthetic data for the 3 thermal preference classes (Warmer, No Change, Cooler) for 5 of the subjects. We designed fully-connected neural networks for the feature extractor, classifier, and conditional generator blocks. A test set is held out from the real dataset to be used for quantitative testing. We then compare the classification performance (COVID/Non-COVID) on this test set for a classifier trained on real data vs a classifier trained on the generated synthetic data. Since the datasets are imbalanced, we report the cohens kappa, accuracy and AUC score (together referred to as classification metrics).

The classification results for a classifier trained on the real data vs a classifier trained on purely conditional synthetic data, and tested on a hold-out set of real data, is given in Table 1. The classifier trained with synthetic data from our proposed model has the close classification performance to that of the classifier trained on real data. This shows the capability of our method to generate synthetic samples with a distribution that closely matches the real conditional data distribution.

5 CONCLUSION AND FUTURE WORK

We presented preliminary results for thermal comfort synthetic data generation using a state-of-the-art conditional synthetic data generation model. The results show that the generative model is capable of generating synthetic data that are close in distribution with the real data. There are numerous future work to the preliminary work that we have presented. The network of the models can be improved (with e.g. ResNets) for better results. Various scenarios can be explored such as mixing and interpolation in the latent space to generate unseen data. A similar methodology can be extended for synthetic data generation in several more smart building use cases [1, 3, 7, 18].

REFERENCES

[1] Bingqing Chen, Priya Donti, Kyri Baker, J Zico Kolter, and Mario Berges. 2021. Enforcing Policy Feasibility Constraints through Differentiable Projection for Energy Optimization. *arXiv preprint arXiv:2105.08881* (2021).

[2] Hari Prasanna Das, Pieter Abbeel, and Costas J Spanos. 2019. Dimensionality reduction flows. *arXiv preprint arXiv:1908.01686* (2019), 1–10.

[3] Hari Prasanna Das, Ioannis C Konstantakopoulos, Aummul Baneen Manasawala, Tanya Veeravalli, Huihan Liu, and Costas J Spanos. 2019. A novel graphical lasso based approach towards segmentation analysis in energy game-theoretic frameworks. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. IEEE, 1702–1709.

[4] Hari Prasanna Das, Ryan Tran, Japjot Singh, Yu-Wen Lin, and Costas J. Spanos. 2021. CDCGen: Cross-Domain Conditional Generation via Normalizing Flows and Adversarial Training. arXiv:2108.11368 [cs.CV]

[5] Hari Prasanna Das, Ryan Tran, Japjot Singh, Xiangyu Yue, Geoff Tison, Alberto Sangiovanni-Vincentelli, and Costas J Spanos. 2021. Conditional Synthetic Data Generation for Robust Machine Learning Applications with Limited Pandemic Data. *arXiv preprint arXiv:2109.06486* (2021).

[6] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. 2017. Density estimation using Real NVP. arXiv:1605.08803 [cs.LG]

[7] Priya L Donti and J Zico Kolter. 2021. Machine Learning for Sustainable Energy Systems. *Annual Review of Environment and Resources* 46 (2021).

[8] Amirata Ghorbani, Vivek Natarajan, David Coz, and Yuan Liu. 2019. DermGAN: Synthetic Generation of Clinical Skin Images with Pathology. arXiv:1911.08716 [cs.CV]

[9] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. arXiv:1406.2661 [stat.ML]

[10] Lindsay T Graham, Thomas Parkinson, and Stefano Schiavon. 2021. Lessons learned from 20 years of CBE's occupant surveys. *Buildings and Cities* 2, 1 (2021).

[11] Tianyu Han, Sven Nebelung, Christoph Haarburger, Nicolas Horst, Sebastian Reinartz, Dorit Merhof, Fabian Kiessling, Volkmar Schulz, and Daniel Truhn. 2019. Breaking Medical Data Sharing Boundaries by Employing Artificial Radiographs. *bioRxiv* (2019). https://doi.org/10.1101/841619 arXiv:https://www.biorxiv.org/content/early/2019/11/14/841619.full.pdf

[12] Timo Kohlberger, Yun Liu, Melissa Moran, Po-Hsuan Cameron Chen, Trissia Brown, Jason D Hipp, Craig H Mermel, and Martin C Stumpe. 2019. Whole-slide image focus quality: Automatic assessment and impact on ai cancer detection. *Journal of pathology informatics* 10 (2019).

[13] Shichao Liu. 2018. Personal thermal comfort models based on physiological parameters measured by wearable sensors. *10th Windsor Conference: Rethinking Comfort* (2018).

[14] Shichao Liu, Stefano Schiavon, Hari Prasanna Das, Ming Jin, and Costas J Spanos. 2019. Personal thermal comfort models with wearable sensors. *Building and Environment* 162 (2019), 106281.

[15] Xuezhe Ma, Xiang Kong, Shanghang Zhang, and Eduard H Hovy. 2021. Decoupling Global and Local Representations via Invertible Generative Flows. In *International Conference on Learning Representations*.

[16] Matias Quintana, Stefano Schiavon, Kwok Wai Tham, and Clayton Miller. 2020. Balancing thermal comfort datasets: We GAN, but should we?. In *Proceedings of the 7th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. 120–129.

[17] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. Modeling tabular data using conditional gan. *Advances in Neural Information Processing Systems* 32 (2019).

[18] Han Zou, Hari Prasanna Das, Jianfei Yang, Yuxun Zhou, and Costas Spanos. 2019. Machine Learning empowered Occupancy Sensing for Smart Buildings. *Climate Change + AI Workshop, International Conference on Machine Learning (ICML)* (2019).

[19] Han Zou, Yuxun Zhou, Jianfei Yang, Huihan Liu, Hari Prasanna Das, and Costas J Spanos. 2019. Consensus Adversarial Domain Adaptation. In *AAAI Conference on Artificial Intelligence 2019.*