Poster Abstract: Towards Class-Balancing Human Comfort Datasets with GANs

Matias Quintana
National University of Singapore
School of Design and Environment
matias@u.nus.edu

Clayton Miller
National University of Singapore
School of Design and Environment
clayton@nus.edu.sg

ABSTRACT

Human comfort datasets are widely used in smart buildings. From thermal comfort prediction to personalized indoor environments, labelled subjective responses from participants in an experiment are required to feed different machine learning models. However, many of these datasets are small in samples per participants, number of participants, or suffer from a class-imbalance of its subjective responses. In this work we explore the use of Generative Adversarial Networks to generate synthetic samples to be used in combination with real ones for data-driven applications in the built environment.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**: *Modeling and simulation*.

KEYWORDS

Generative Methods, Human Comfort, Data Augmentation

ACM Reference Format:

Matias Quintana and Clayton Miller. 2019. Poster Abstract: Towards Class-Balancing Human Comfort Datasets with GANs. In *The 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys '19)*, November 13–14, 2019, New York, NY, USA. ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/3360322.3361016>

1 INTRODUCTION

The intersection of thermal comfort and building control are exploiting the use of data-driven and grey box techniques for energy savings and human comfort in the built environment [4]. Different approaches require environmental measurements as well as the users' subjective response as ground truth label in order to model the occupant's preference [7]. However, the data used is often limited by administrative overhead, cohort size, and users' behaviour during the experiments (i.e. survey fatigue), producing a class-balance issue that is often ignored or remove [1]. In this paper, we proposed the use of Generative Adversarial Networks (GANs) as a pre-processing step to tackle the class-imbalance nature of subjective human responses datasets. We show that a combined dataset of real and synthetic samples (less predominant classes)

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
BuildSys '19, November 13–14, 2019, New York, NY, USA
© 2019 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-7005-9/19/11.
<https://doi.org/10.1145/3360322.3361016>

Table 1: Dataset sample distribution by Subjective Comfort Label in the original and train set

Comfort Index	Label	Dataset Count	Train Set Count
Uncomfortably Warm	+2	48	44
Slightly Uncomfortably Warm	+1	198	171
Comfortable	0	1152	818
Slightly Uncomfortably Cold	-1	452	303
Uncomfortably Cold	-2	217	172

Table 2: Generated datasets subjective comfort class distribution. Numbers in parenthesis indicate the real and synthetic samples, respectively.

Label	Train + Synth1 Count	Train + Synth2 Count	Train + Synth3 Count
+2	88 (44+44)	132 (44+88)	176 (44+132)
+1	342 (171+171)	513 (171+342)	684 (171+513)
0	818 (818+0)	818 (818+0)	818 (818+0)
-1	606 (303+303)	909 (303+606)	909 (303+606)
-2	344 (172+172)	516 (171+344)	688 (172+516)

sets the stage for a more diverse and rich dataset for modeling techniques.

1.1 Balancing Classes

GANs were first proposed in [2] and have been widely explored since then. This technique uses a Generator G to capture the underlying data distribution of a dataset and a Discriminator D to estimate the probability that a given sample comes from the original dataset rather than being created by G . This architecture has allowed researchers to tackle different image-based problems such as style transfer and image-to-image translation [8]. On the other hand, only a few tackle the issue of tabular data. Table-GAN [5] uses existing image-based GAN architecture to synthesize fake tables that are statistically similar to the original dataset, however the focus remains in anonymizing the data rather than creating complementary fake samples. Another existing approach is Tabular-GAN [6] which focuses on the marginal distribution of the features using a more complex network architecture for G .

2 SYNTHETIC SAMPLE GENERATION

For this work we selected a thermal comfort dataset from 77 participants (with a train and test split of 0.7 split across the participants) in a fully-controlled thermal chamber as part of [3]. The dataset

contains a total of 9 continuous and discrete features including environmental and physiological measurements, as well as the subjective comfort label, making it the tenth feature). Table 1 shows the distribution of samples by subjective comfort label and it clearly shows the clear predominance of the label “0” (Comfortable) against other classes like “+2” (Uncomfortably Warm), having 24 times more instances. Three GAN architectures were chosen: Tabular-GAN [6] and two vanilla GANs. One is a plain vanilla GAN and the other one, GANCorr, has an additional correlation loss in G that tries to minimize the correlation between the ground truth training and the generated batch. Code and implementation details can be found in <https://github.com/buds-lab/generative-methods-for-human-comfort>. In order to increase the non “0” classes, each GAN generated samples of the remaining classes in the order of one, two, and three times the initial train set size (see Table 2).

3 CONCLUSION AND DISCUSSION

The datasets listed in Table 2, alongside the synthetic sets themselves, were used as training sets for three commonly used thermal comfort prediction baselines: K -Nearest Neighbours (KNN), Support Vector Machines (SVM), and Naive-Bayes (NB). The hyperparameters were found performing cross-validation and the performance reported was obtained with the initial test set from 2. Figure 1 show the test set F1-micro score of the datasets used for training and their respective GANs used (Y-axis) and the different models (X-axis).

Figure 1: F1-micro score of different datasets of and their respective GAN. The row *Train* shows the performance of the original train set.

From Figure 1 we see purely synthetic datasets of the same size as the dataset achieve similar performance to real samples with NB and SVM. This indicates the generated samples captured the feature relationships of the original dataset. However, when the size of the generated samples surpasses the real samples, purely generated dataset have a significant drop in performance but when mixed with real samples, some maintain the same accuracy: Train+Synthetic2

and Train+Synthetic3 with TGAN and GANCorr for KNN and NB respectively. Even though a clear increase in accuracy is not shown yet, besides a slight increase in KNN with Train+Synth3 and Table-GAN, maintaining the baseline performance indicates the GAN was able to capture the features relationships. These preliminary results show promise in future explorations of augmenting human subjective labelled ground truth data. However, a more fair comparison against other methods such as random oversampling and SMOTE are needed. Future work will consider different ways of partition the data (per participant or in a holistic approach), address the sequence-type behaviour of human responses, and different initializations for G and D .

ACKNOWLEDGMENTS

This material is based upon work supported by the U.S. Department of Energy’s Office of Energy Efficiency and Renewable Energy (EERE) under the Building Technologies Office Award Number DE-EE0007682. **Disclaimer:** This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

REFERENCES

- [1] Damien Fay, Liam O’Toole, and Kenneth N. Brown. 2017. Gaussian Process models for ubiquitous user comfort preference sampling: global priors, active sampling and outlier rejection. *Pervasive and Mobile Computing* 39 (2017), 135–158. <https://doi.org/10.1016/j.pmcj.2016.08.012>
- [2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. *Corrosion* (2014), iii. <https://doi.org/10.1016/B978-0-408-00109-0.50001-8> arXiv:arXiv:1011.1669v3
- [3] Sirajum Munir, Le Tran, Jonathan Francis, Charles Shelton, Ripudaman Singh Arora, Craig Hesling, Matias Quintana, Anand Krishnan Prakash, Anthony Rowe, and Mario Berges. 2017. Demo: FORK: Fine grained Occupancy estimator using Kinect on ARM Embedded Platforms. In *BuildSys ’17 Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments*. <https://doi.org/10.1145/3137133.3141461>
- [4] June Young Park and Zoltan Nagy. 2018. Comprehensive analysis of the relationship between thermal comfort and building control research - A data-driven literature review. *Renewable and Sustainable Energy Reviews* 82, July (2018), 2664–2679. <https://doi.org/10.1016/j.rser.2017.09.102>
- [5] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. 2018. Data Synthesis based on Generative Adversarial Networks. 11, 10 (2018), 1071–1083. <https://doi.org/10.14778/3231751.3231757> arXiv:1806.03384
- [6] Lei Xu and Kalyan Veeramachaneni. 2018. Synthesizing Tabular Data using Generative Adversarial Networks. (2018). arXiv:1811.11264 <http://arxiv.org/abs/1811.11264>
- [7] Liang Zhang, Abraham Hang-yat Lam, and Dan Wang. 2014. Strategy-proof thermal comfort voting in buildings. (2014), 160–163. <https://doi.org/10.1145/2674061.2674074>
- [8] Jun Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *Proceedings of the IEEE International Conference on Computer Vision* 2017-Octob (2017), 2242–2251. <https://doi.org/10.1109/ICCV.2017.244> arXiv:arXiv:1703.10593v6