TECHNICAL ALGORITHMIC BIAS IN A MUSIC RECOMMENDER

Arthur Flexer,[1] Monika Dörfler,[2] Jan Schlüter,[1] Thomas Grill[1]
[1]Austrian Research Institute for Artificial Intelligence (OFAI), Vienna, Austria
[2]Numerical Harmonic Analysis Group, Faculty of Mathematics, University of Vienna, Austria
arthur.flexer|jan.schlueter|thomas.grill@ofai.at, monika.doerfler@univie.ac.at

1. INTRODUCTION

Our work brings the problem of technical algorithm bias to the attention of the music information retrieval (MIR) community. We illustrate this so far neglected problem for a real world music recommender, where due to a problem of measuring distances in high dimensional spaces, songs closer to the center of all data are recommended over and over again, while songs far from the center are not recommended at all. We show that these so-called hub songs do not carry a specific semantic meaning and that deleting them from the data base promotes other songs to hub songs being recommended disturbingly often as a consequence. We argue for the ethical responsibility of MIR researchers to assure that their algorithms are unbiased and fair. More detail concerning the experiments can be found in [2].

2. RELATED WORK

The term algorithmic bias is used to describe systematic and repeatable errors that create unfair outcomes in computer experiments, i.e., generating one result for certain users or certain data and a different result for others [4]. In MIR, the majority of respective research is concerned with unfairness due to bias in training data, e.g. music collections neglecting music outside European or US culture areas, which has been researched most prominently within the CompMusic project (`http://compmusic.upf.edu/`). Technical algorithmic bias on the other hand arises from specifically technical constraints, which may be due to hardware, software or even peripherals. In a very recent overview article [5] on ethical dimensions of MIR technology, algorithmic bias is being discussed for a hypothetical music recommender where "a large number of artists is never recommended [...] due to a lack of user data or other artifacts that are not completely understood". This paper aims at understanding such a failure of a specific music recommendation system as a problem of technical algorithmic bias due to hubness, which is a general problem of learning in high dimensions [6, 7].

Hubness was first noted as a problem in audio-based music recommendation [1], more specifically that certain hub songs were being recommended conspicuously often in nearest neighbor-based playlists, while other songs acting as anti-hubs were never recommended. Hubness is related to the phenomenon of concentration of distances, where all pairwise distances are approximately the same for dimensionality approaching infinity. One of the important factors deciding which data object acts as a hub is the distance to the data mean. Objects in close proximity of the sample mean of some data distribution are prone to become hubs in high dimensional spaces [6]. On the other hand, anti-hubs are typically far from centers.

3. DATA

For our analysis we use data from a real-world music discovery system (`http://fm4.orf.at/soundpark`) where artists can upload and present their music free of charge, with most recent uploads being displayed at the top of the website. To allow a more appealing access regardless of a song's publication date, a recommendation system using a content-based music similarity measure was implemented [3] as a visualization of a k-nearest neighbor graph showing the $k = 5$ most similar songs to the currently selected track. This similarity measure is based on timbre information computed via the following steps: divide raw audio data into short overlapping segments, apply a Hann window to each segment, compute power spectrum matrix via FFT, transform power spectrum to mel-scale using a filter bank of triangular filters, convert to decibel scale by taking the logarithm, apply discrete cosine transform to compress and smooth the mel power spectrum to 20 MFCCs, train a single Gaussian (G1) to model all of the segments represented as MFCCs for each of the songs, compute a distance matrix between all songs using the Kullback-Leibler (KL) divergence between respective G1 models. For our experiments we use a development data base of 16583 songs.

4. RESULTS

The major evaluation measure to characterize hubness is the k-occurrence O^k, i.e. the number of times a song occurs in the first k nearest neighbors of all other songs in the database [1]. The mean O^k across all songs in a database is equal to k. Any k-occurrence significantly bigger than k therefore indicates existence of a hub. We select $k = 5$ because our music recommender always shows the five most

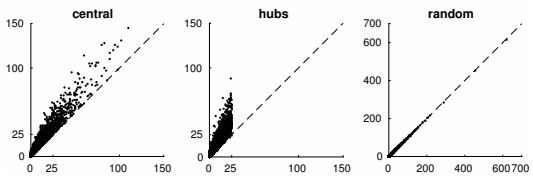

Figure 1. k-occurrences before (x-axis) and after (y-axis) deletion of songs.

similar songs. We define that any song with $O^k > 5k = 25$ is a hub, any song with $O^k = 0$ is an anti-hub, any song with $O^k > 0 \wedge O^k \leq 5k = 25$ is a so-called normal object. The number of hubs in our data base is 653 and the number of anti-hubs 5953, which means that more than a third of the data are never being recommended. The one largest hub appears in the recommendation lists of 620 other songs. Although the 653 hub songs constitute only about 4% of the data base, they dominate the recommendation lists by covering 40.11% of them.

What makes a song a hub song? Looking at the mean KL divergences ($\pm$ standard deviation) of songs to the center of the data, one can see that indeed hub songs are on average (51.63 ± 4.55) closer to the center than normal songs (68.60 ± 67.01), while anti-hubs are furthest away (76.22 ± 44.67). Listening to the largest hub song, a random normal and an anti-hub song (sound files accessible at `http://ofai.at/~arthur.flexer/ismir2018.html`), as well as many other randomly chosen songs, does not offer any clues as to what makes a song act as a hub song.

This casts doubt whether a song's property of being or not being a hub depends on the semantic content of that particular song at all. A waveform, being perceptually equivalent to the original sound, can in principle be reconstructed from a sufficiently redundant spectrogram. A meaningful approximate reconstruction can also still be obtained from MFCCs, which are a version of spectrogram coefficients in which information compression and averaging has been performed in frequency. But using a single Gaussian to model the spectral content of a song essentially averages over all MFCCs in time. The remaining average information does not allow to make meaningful statements about the original waveform used for computing MFCCs. The same average can be obtained through data with very different distributions around the mean values, just to give a trivial example. Therefore very different audio signals can be close to the center of all data and hence very different audio signals can become hub songs.

What if we delete central or hub songs from the data base? We removed the 653 most central songs (i.e. those with minimal KL divergence to the center of the data), or we removed the 653 hub songs (i.e. every song with $O^k > 25$), or, as a control, 653 random songs. In Figure 1 we plotted the results with the k-occurrences before deletion on the x-axis and the k-occurrences after deletion on the y-axis. Looking at the left and central plots giving the results for deletion of the most central or the hub songs, one can see that many k-occurrences increase after deletion, i.e. most points in the plot are above the dashed diagonal axis. Other songs now take over the role of deleted hub songs or deleted central songs. Looking at the right plot giving the results for deletion of 653 random songs, one can see that k-occurrences basically remain identical.

5. CONCLUSION

The intention of this paper was to bring the ethical responsibility to produce fair and unbiased MIR systems to the attention of the MIR community. This was done by presenting an example of technical algorithmic bias, where a music recommendation system, due to a problem of high-dimensional machine learning, favors a small group of songs in its recommendations. These so-called hub songs dominate the recommendation lists not because they have a specific sound or semantic meaning, but because the algorithmic bias in connection with the signal representation and modeling of the system favors songs close to the center of the data set, a requirement which these hub songs fulfill almost by accident. It is our hope that this paper will trigger a larger discussion about the technical biases built into our MIR algorithms. Ways to reduce hubness are described in [7] and at `https://github.com/OFAI/hub-toolbox-python3`.

Acknowledgements: This work was supported by the Austrian Science Fund (FWF P27082) and the Vienna Science and Technology Fund (WWTF MA14-018).

6. REFERENCES

[1] Aucouturier, J.-J., Pachet F.: Improving Timbre Similarity: How high is the sky?, *Journal of Negative Results in Speech and Audio Sciences*, 1(1), 2004.

[2] Flexer A., Dörfler M., Schlüter J, Grill T.: Hubness as a case of technical algorithmic bias in music recommendation, *Proceedings of 6th IEEE ICDM Workshop on High Dimensional Data Mining*, to appear, 2018.

[3] Flexer A., Stevens J.: Mutual proximity graphs for improved reachability in music recommendation, *Journal of New Music Research*, 47(1), pp. 17-28, 2018.

[4] Friedman B., Nissenbaum H.: Bias in Computer Systems, *ACM Trans. on Information Systems*, 14 (3): 330-347, 1996.

[5] Holzapfel A., Sturm B., Coeckelbergh M.: Ethical Dimensions of Music Information Retrieval Technology, Transactions of the International Society for Music Information Retrieval, in press, 2018.

[6] Radovanović M., Nanopoulos A., Ivanović M.: Hubs in space: Popular nearest neighbors in high-dimensional data, *Journal of Machine Learning Research*, 11:2487-2531, 2010.

[7] Schnitzer D., Flexer A., Schedl M., Widmer G.: Local and Global Scaling Reduce Hubs in Space, *Journal of Machine Learning Research*, 13:2871-2902, 2012.