SHAPING DATASETS: OPTIMAL DATA SELECTION FOR SPECIFIC TARGET DISTRIBUTIONS ACROSS DIMENSIONS

Vassilios Vonikakis[1], *Ramanathan Subramanian*[1,2], *Stefan Winkler*[1]

[1] Advanced Digital Sciences Center (ADSC), University of Illinois at Urbana-Champaign, Singapore
[2] Center for Visual Information Technology, Int'l Institute of Information Technology, Hyderabad, India
{bbonik, Subramanian.R, Stefan.Winkler}@adsc.com.sg

ABSTRACT

This paper presents a method for dataset manipulation based on Mixed Integer Linear Programming (MILP). The proposed optimization can narrow down a dataset to a particular size, while enforcing specific distributions across different dimensions. It essentially leverages the redundancies of an initial dataset in order to generate more *compact* versions of it, with a specific target distribution across each dimension. If the desired target distribution is uniform, then the effect is balancing: all values across all different dimensions are equally represented. Other types of target distributions can also be specified, depending on the nature of the problem. The proposed approach may be used in machine learning, for shaping training and testing datasets, or in crowdsourcing, for preparing datasets of a manageable size.

Index Terms— Mixed Integer Linear Programming (MILP), datasets, balancing, crowdsourcing.

1. INTRODUCTION

Data is abundant in our time. The widespread use of cameras, microphones, and other sensors in our everyday lives has made it easier than ever to collect all types of data. This has contributed significantly to advancement in many scientific domains, including especially image processing and computer vision. It is not an exaggeration to claim that nowadays, there is at least one dataset for (almost) every research problem.

Advances in big data analytics have contributed to the notion that 'bigger is better'. However, little attention is usually given to feature distributions in the dataset. Consequently, some datasets may be significantly skewed towards specific attributes. For example, a dataset for gender estimation with the majority of its images depicting people of a specific age group is non-representative of real life, and may not be appropriate for use as a training set. Moreover, the skewness may affect more than one dimension of interest, which limits re-usability of a model learned from a particular dataset.

Depending on the objective, different approaches may be used in order to deal with imbalanced datasets. Undersampling (reducing the over-represented classes) and oversampling (replicating the under-represented classes) are two typical approaches [1]. Assigning different importance weights to data points is also another technique that may result in more 'balanced' classifiers. For a more comprehensive review of existing balancing techniques please refer to [2].

Although existing techniques may alleviate imbalanced feature distributions, they do not explicitly provide a solution for narrowing down the size of a dataset while simultaneously enforcing specific target distributions (not necessarily uniform) for different features. Synthesizing a smaller subset via subsampling with a particular distribution for different dimensions is a challenging combinatorial problem that is of interest in many different areas.

In machine learning, creating datasets with different feature distributions can provide insights regarding the generalizability of a learning algorithm. Crowdsourcing, where smaller chunks of a larger dataset are annotated by workers, is another field where a subsampling technique could be of potential use. Apart from plain annotations, crowdsourcing is also used in an exploratory way for discovering user preferences and behaviors. For example, large scale crowdsourcing studies have analyzed how image attributes impact memorability [3], image appeal [4, 5], and visual summarization [6]. In these studies no particular attention is given to the distributions of each image attribute in the crowdsourced dataset. The danger here is that the values of some attributes may be over/under-represented in the dataset, contributing to skewed results regarding the workers' preferences and behaviors. Moreover, the crowdsourcing scenario necessitates narrowing down to a manageable size the number of items the workers have to interact with, especially in exploratory studies. This is due to the fact that large sets are difficult to process, and workers may not pay equal attention to all items.

To this end, a new dataset *shaping* technique is introduced, based on Mixed Integer Linear Programming (MILP).

This study is supported by the research grant for the Human-Centered Cyber-physical Systems Programme at the Advanced Digital Sciences Center from Singapore's Agency for Science, Technology and Research (A*STAR).

The proposed optimization can narrow down a dataset to a particular given size, while enforcing specific distributions across different dimensions. It essentially leverages the redundancies of an initial dataset, in order to generate more *compact* versions of it, with a specific target distribution across each dimension. If the target distribution is uniform, then the effect is balancing: all values across all different dimensions are equally represented. Other types of target distributions can also be used, depending on the nature of the problem.

The rest of the paper is organized as follows. Section 2 describes the proposed approach based on MILP optimization. Section 3 demonstrates the results of the proposed aproach in two publicly available datasets. Concluding remarks are given in Section 4.

2. OPTIMIZATION FOR DATASET CREATION

Let $S = \left\{\mathbf{q}_i \mid \mathbf{q}_i \in \mathbb{R}^M, \mathbf{q}_i \sim D_S^M\right\}_{i=1}^K$ be an initial set of observations of a random variable, forming the data matrix $\mathbf{Q} = [q_{ij}]_{K\times M}$. Assuming there is sufficient redundancy across all M dimensions, the objective is to select a subset of observations $s \subset S$ with $s = \left\{\hat{\mathbf{q}}_i \mid \hat{\mathbf{q}}_i \in S, \hat{\mathbf{q}}_i \sim D_s^M\right\}_{i=1}^N$, $N \ll K$, and $\hat{\mathbf{Q}} = [\hat{q}_{ij}]_{N\times M}$ denoting the reduced data matrix. Enforcing $D_s^M = U$ ensures that s will have a uniform distribution, resulting in a balancing effect. However, other target distributions may be preferred, depending on the problem.

Let matrix $\mathbf{D} \in \mathbb{R}^{H\times M}$ represent the target distribution D_s^M, such that each of its columns $\mathbf{D}_{*j}$ contains the Probability Mass Function (PMF) of D_s^M across the j^{th} dimension, quantized into H intervals (bins). Let $B = \{\mathbf{B}^m\}_{m=1}^M$ denote a set of M binary matrices, with $\mathbf{B}^m \in \mathbb{Z}_2^{H\times K}$, such that each binary element b_{ij}^m denotes whether or not the j^{th} item of S belongs to the i^{th} interval of the target PMF for dimension m. Finally, we introduce a binary vector $\mathbf{x} \in \mathbb{Z}_2^K$, whose coefficients x_i are decision variables determining whether or not the i^{th} item of S belongs to the subset s. The problem then can be formulated as the following minimization:

$$\min_{\mathbf{x}} \sum_{m=1}^{M} \|\mathbf{B}^m\mathbf{x} - N\mathbf{D}_{*m}\|_1 \text{ s.t. } \|\mathbf{x}\|_1 = N, \tag{1}$$

which essentially means, “select those N elements from S that minimize the L1 distance from the target PMF and thus approximate D_s^M”. The above minimization can be solved by using a set of auxiliary vectors $Z = \{\mathbf{z}_i\}_{i=1}^M$ with $\mathbf{z}_i \in \mathbb{R}_+^H$, in order to handle the absolute values of the L1 norm:

$$\left.\begin{aligned} \mathbf{B}^m\mathbf{x} - N\mathbf{D}_{*m} &\le \mathbf{z}_m \\ \mathbf{B}^m\mathbf{x} - N\mathbf{D}_{*m} &\ge -\mathbf{z}_m \end{aligned}\right\} \Rightarrow \left.\begin{aligned} \mathbf{B}^m\mathbf{x} - \mathbf{z}_m &\le N\mathbf{D}_{*m} \\ -\mathbf{B}^m\mathbf{x} - \mathbf{z}_m &\le -N\mathbf{D}_{*m} \end{aligned}\right\} \tag{2}$$

for each dimension m and minimizing over Z. The final optimization can be expressed as MILP as follows:

$$\text{Minimize } \mathbf{c}^{\mathsf{T}}\tilde{\mathbf{x}} \text{ s.t. } \mathbf{A}\tilde{\mathbf{x}} \le \mathbf{b}, \tag{3}$$

with $\mathbf{c} = \left[\mathbf{0}_K^{\mathsf{T}} \;\; \mathbf{1}_{HM}^{\mathsf{T}}\right]^{\mathsf{T}}$, $\tilde{\mathbf{x}} = \left[\mathbf{x}^{\mathsf{T}} \;\; \mathbf{z}_1^{\mathsf{T}} \cdots \mathbf{z}_M^{\mathsf{T}}\right]^{\mathsf{T}}$ and

$$\mathbf{A} = \left[\begin{array}{cc} \mathbf{1}_K^{\mathsf{T}} & \mathbf{0}_{HM}^{\mathsf{T}} \\ -\mathbf{1}_K^{\mathsf{T}} & \mathbf{0}_{HM}^{\mathsf{T}} \\ \hline \mathbf{B}^1 & \\ \vdots & -\mathbf{I}_{HM} \\ \mathbf{B}^M & \\ \hline -\mathbf{B}^1 & \\ \vdots & -\mathbf{I}_{HM} \\ -\mathbf{B}^M & \end{array}\right], \mathbf{b} = \left[\begin{array}{c} N \\ -N \\ \hline N\mathbf{D}_{*1} \\ \vdots \\ N\mathbf{D}_{*M} \\ \hline -N\mathbf{D}_{*1} \\ \vdots \\ -N\mathbf{D}_{*M} \end{array}\right]$$

$\mathbf{A} \in \mathbb{Z}^{(2+2HM)\times(K+HM)}$, $\mathbf{b} \in \mathbb{R}^{(2+2HM)}$ and $\mathbf{c} \in \mathbb{Z}_2^{K+HM}$, while $\tilde{\mathbf{x}}$ is also of size $K + HM$ and contains both the integer and real optimization variables. The first two rows of $\mathbf{A}$ and $\mathbf{b}$ address the equality constraint of the integer variables $\|\mathbf{x}\|_1 = N$ expressed as two inequality constraints $\sum_{i=1}^K x_i \le N$ and $-\sum_{i=1}^K x_i \le -N$. The lower two sections address the constraints for the real auxiliary variables, derived from the upper and lower parts of Eq. (2).

MILP problems are NP-hard combinatorial problems. However, modern branch-and-bound algorithms can solve many real world problems reliably and fast [8]. Such algorithms solve the LP relaxation problem to obtain fractional solutions and create two sub-branches by adding new constraints [9]. As an indication, our implementation[1] uses MATLAB’s *intlinprog* function and is able to solve the problem of Eq. (3) for a dataset of $K = 220,000$ observations, $M = 30$ dimensions, $H = 100$ quantization bins, and $N = 10000$ selections, in approximately 85 seconds on a typical PC with 16GB of memory. Smaller datasets, similar to the ones in Section 3, require less than a second. This demonstrates that the proposed approach can be easily used at least in small or medium sized datasets.

3. EXPERIMENTAL RESULTS

In order to demonstrate the utility of the proposed approach, we apply it to two different image datasets: the Gallagher dataset [7], which contains family photos, and the Helen dataset [10], which comprises images of faces used for facial landmark recognition and tracking. In both cases, we create subsets of the original datasets with different distributions across different dimensions using the above technique.

3.1. Gallagher Dataset

The Gallagher dataset [7] comprises 589 photos depicting family moments, and is a typical example of a personal photo collection. These images have many different characteristics,

[1] Our Matlab code is freely available at `https://sites.google.com/site/vonikakis/software-code/dataset_shaping`

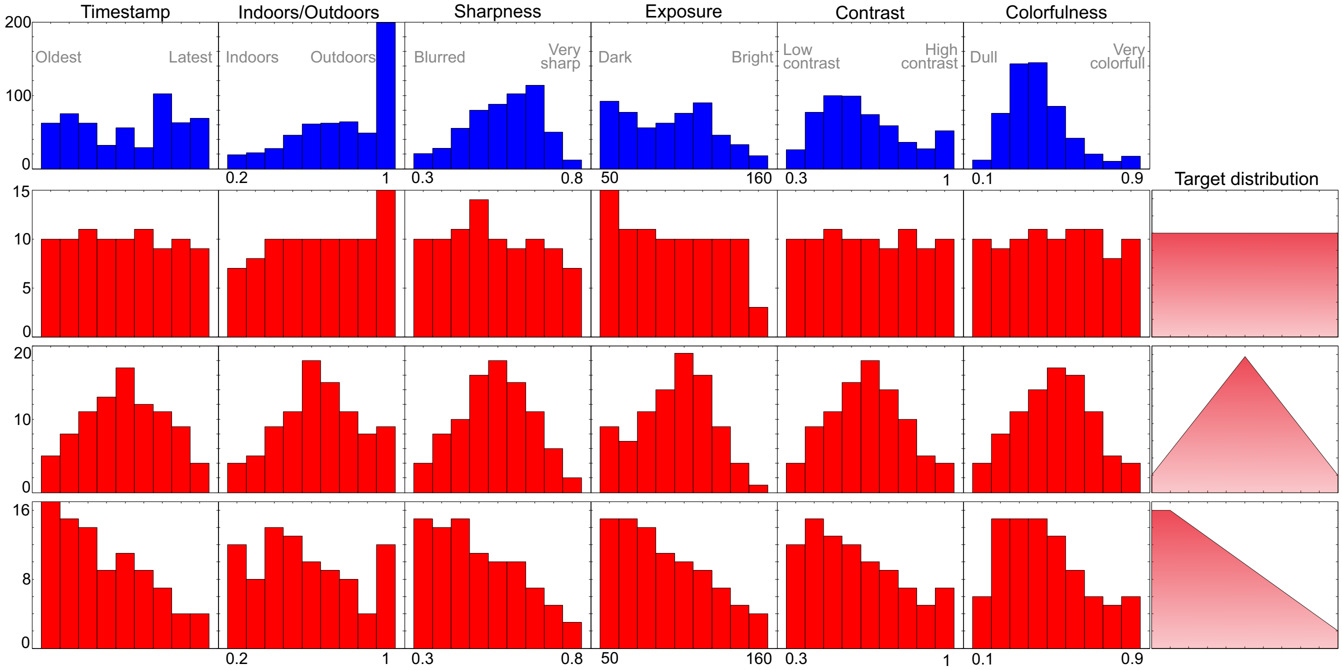

Fig. 1. Different ways of selecting 90 out of the *same* 589 photos from the Gallagher dataset [7], according to different target distributions. Top row: original distributions of the dataset. Second row: enforcing a uniform distribution. Third row: enforcing a triangular distribution. Bottom row: enforcing a linearly-descending distribution.

and are captured under a diverse set of conditions. We analyze the images according to 6 different attributes/dimensions.

1. **Capture Time:** We extract the EXIF timestamps from the captured photos.

2. **Scene Type:** Indoors/outdoors. We employ the *Relative Attributes* approach proposed in [11], using gist features and color histograms, and train the system to compute a real-valued rank specifying the indoor/outdoor-ness for each image.

3. **Image Sharpness:** Overall perceived sharpness of an image, computed similar to [12].

4. **Image Exposure:** Overall exposure of an image, computed by the mean value of the luminance component.

5. **Image Contrast:** Overall contrast of an image, computed by the variation coefficient (relative standard deviation) of the luminance component.

6. **Image Colorfulness:** Perceived vividness of colors of an image, as computed in [13].

All the above attributes are normalized to the interval $[0,1]$ using *min-max* normalization and then quantized into 9 discrete levels ($H = 9$). In some cases where the original distribution was severely skewed, a simple non-linear mapping (e.g. logarithmic) was applied to the data before quantization.

Fig. 1 depicts the initial distributions, as well as the results of the proposed method for selecting 90 out of the 589 images, using 3 different target distributions: uniform, triangular, and linearly-descending. One immediate observation is that the original dataset distributions (blue) are far from uniform and follow different shapes for all 6 dimensions. For all subsets created by our method however, it is evident that the resulting distributions across all 6 dimensions closely resemble the target. Naturally, the match is not perfect, because the assumption of 'sufficient variation across all dimensions' for the given original dataset (Gallagher) is not met. In other words, the original dataset is not *redundant* enough in order to include all possible combinations of values across the dimensions of interest. For example, this is the case for the 'indoors/outdoors' attribute, where the initial distribution is heavily skewed. It should be noted that the degree of skewness of the data is not necessarily a problem by itself for the proposed method. It may become a problem however, when the requested number of selected items N becomes comparable to the total number of available items K.

3.2. Helen Dataset

The Helen dataset [10] comprises 2330 images of faces in various head poses and expressions, and is mainly used for training algorithms for the detection of facial registration points. An important aspect of a successful facial point detector algorithm is to detect the correct location of facial points even in

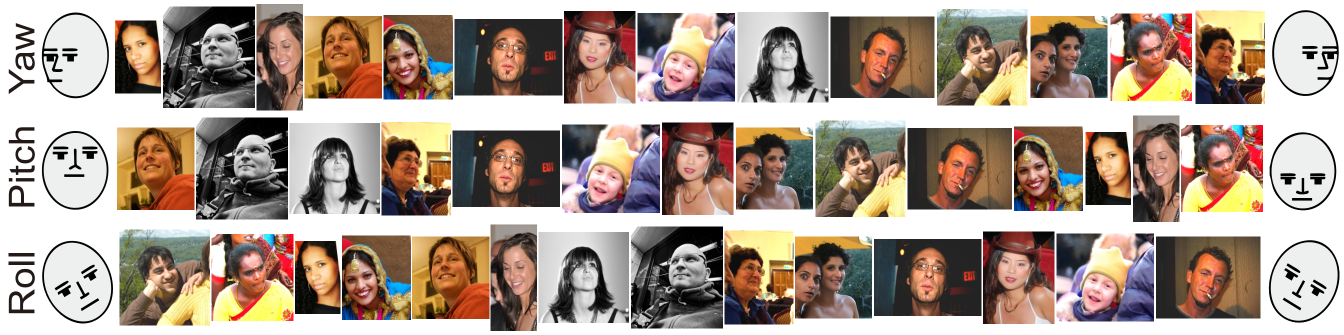

Fig. 2. Selecting 14 images from the Helen dataset [10] in a balanced way. Images are arranged according to Yaw (top row), Pitch (middle row), and Roll (bottom row).

extreme non-frontal head poses. To this end, the initial distribution of head poses in the training dataset is very important; a dataset with considerably more frontal faces will be skewed and will not have sufficient training examples for other head poses. As such, algorithms trained on such datasets may exhibit good performance for frontal faces but may suffer for other head poses.

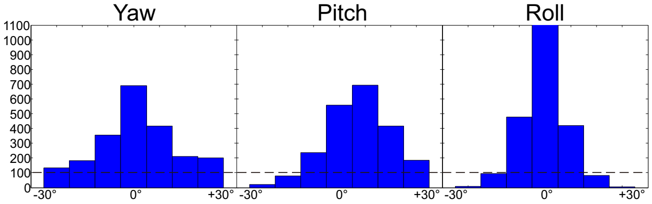

Fig. 3. Distributions of the original Helen dataset (total 2330 images) across the three head-pose dimensions.

Table 1. MSE between a perfectly uniform distribution and the one generated by the proposed approach, for different dataset sizes (as percentage of the complete Helen dataset).

Dataset Size	Yaw	Pitch	Roll	Mean
100% (original)	0.048	0.078	0.198	0.108
75%	0.022	0.07	0.155	0.082
50%	0.001	0.039	0.116	0.052
25%	0	0.018	0.069	0.029
5%	0	0	0.041	0.014
1%	0	0	0	0

In this context, we analyze the Helen dataset in terms of three attributes/dimensions which characterize the orientation of the head: *yaw*, *pitch* and *roll*. All three attributes were estimated by the Intraface library [14], truncated to the interval $[-30, 30]$ and quantized into $H = 7$ bins. Fig. 3 depicts the distribution of the whole dataset for the three head pose dimensions. It is clear that the dataset is not well balanced and strongly skewed towards frontal faces.

We apply the proposed approach, enforcing a uniform distribution for different resulting sizes of datasets (varying N). Table 1 shows the Mean Square Error (MSE) between a perfectly uniform distribution and the one generated by the proposed approach. Dataset sizes are expressed as a percentage of the original Helen dataset. It is evident that as the number of target images decreases, the resulting distribution increasingly resembles a uniform distribution. Given the greater redundancy for non-frontal yaw instances in the dataset, convergence of the distribution to the target (uniform) happens earlier for yaw than pitch and roll.

To further demonstrate the balancing effect of the proposed algorithm in a qualitative manner, 14 images were selected from the Helen dataset so as to achieve a balanced distribution for the three head pose dimensions using the proposed algorithm. Fig. 2 depicts these images arranged in the specified order for roll, pitch, and yaw. It is evident that within the 14 selected images, different values corresponding to the target attribute/dimension are equally represented.

4. CONCLUSIONS

This paper introduces a methodology for dataset subsampling and shaping based on Mixed Integer Linear Programming (MILP). The proposed approach can narrow down a dataset to a particular size, while enforcing specific target distributions across different dimensions. Experimental results demonstrated the ability of the algorithm to undersample datasets and successfully enforce various target distributions across different dimensions and quantization ranges.

As a straightforward application, our algorithm can be used for balancing originally imbalanced datasets (enforcing a uniform distribution). Possible uses include machine learning and user studies involving crowdsourcing, where smaller balanced datasets can be created, to eliminate the effect of data biases on user behavior. Our technique can limit the cost of such studies (smaller number of items to interact with) and indirectly increases the quality of the acquired results (due to lower fragmentation of workers' attention).

5. REFERENCES

[1] Nathalie Japkowicz and Shaju Stephen, "The class imbalance problem: A systematic study," *Intelligent Data Analysis*, vol. 6, no. 5, pp. 429–449, Oct. 2002.

[2] Nitesh V. Chawla, "Data mining for imbalanced datasets: An overview," in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Eds., pp. 853–867. Springer, Boston, MA, 2005.

[3] P. Isola, Jianxiong Xiao, D. Parikh, A. Torralba, and A. Oliva, "What makes a photograph memorable?," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 7, pp. 1469–1482, July 2014.

[4] Vassilios Vonikakis, Ramanathan Subramanian, and Stefan Winkler, "How do users make a people-centric slideshow?," in *Proc. 2nd ACM International Workshop on Crowdsourcing for Multimedia (CrowdMM)*, Barcelona, Spain, 2013, pp. 13–14.

[5] Vassilios Vonikakis, Ramanathan Subramanian, Jonas Arnfred, and Stefan Winkler, "Modeling image appeal based on crowd preferences for automated person-centric collage creation," in *Proc. 3rd ACM International Workshop on Crowdsourcing for Multimedia (CrowdMM)*, Orlando, FL, 2014, pp. 9–15.

[6] S. Rudinac, M. Larson, and A. Hanjalic, "Learning crowdsourced user preferences for visual summarization of image collections," *Multimedia, IEEE Transactions on*, vol. 15, no. 6, pp. 1231–1243, Oct. 2013.

[7] Andrew C. Gallagher and Tsuhan Chen, "Clothing cosegmentation for recognizing people," in *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, Anchorage, AK, June 2008, pp. 1–8.

[8] Alper Atamturk and Martin W.P. Savelsbergh, "Integer-programming software systems," *Annals of Operations Research*, vol. 140, no. 1, pp. 67–124, 2005.

[9] Jens Clausen, "Parallel branch and bound – principles and personal experiences," in *Parallel Computing in Optimization*, A. Migdalas, P. M. Pardalos, and S. Storøy, Eds., vol. 7 of *Applied Optimization*, pp. 239–267. Springer, 1997.

[10] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S. Huang, "Interactive facial feature localization," in *Proc. 12th European Conference on Computer Vision (ECCV)*, Florence, Italy, 2012, vol. 3, pp. 679–692.

[11] Devi Parikh and Kristen Grauman, "Relative attributes," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, Barcelona, Spain, Nov. 2011, pp. 503–510.

[12] Roni Ferzli and Lina J. Karam, "A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB)," *Image Processing, IEEE Transactions on*, vol. 18, no. 4, pp. 717–728, April 2009.

[13] David Hasler and Sabine E. Susstrunk, "Measuring colorfulness in natural images," in *Proc. SPIE*, 2003, vol. 5007, pp. 87–95.

[14] X. Xiong and F. de la Torre, "Supervised descent method and its applications to face alignment," in *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, Portland, OR, June 2013, pp. 532–539.