Crowdsorting Timed Comments about Music: Foundations for a New Crowdsourcing Task

Karthik Yadati, Pavala S.N. Chandrasekaran Ayyanathan, Martha Larson

Delft University of Technology, Netherlands
{n.k.yadati, p.s.n.chandrasekaranayyanathan, m.a.larson}@tudelft.nl

ABSTRACT

This paper provides an overview of the Crowdsorting Timed Comments about Music Task, a new task in the area of crowdsourcing for social media offered by the MediaEval 2014 Multimedia Benchmark. Data for this task is a set of Electronic Dance Music (EDM) tracks, collected from online music sharing platform Soundcloud. Given a set of noisy labels for segments of Electronic Dance Music (EDM) that were collected on Amazon Mechanical Turk, the task is to predict a single 'correct' label. The labels indicate whether or not a 'drop' occurs in the particular music segment. The larger aim of this task is to contribute to the development of hybrid human/conventional computation techniques to generate accurate labels for social multimedia content. For this reason, participants are also encouraged to predict labels by combining input from the crowd (i.e., human computation) with automatic computation (i.e., processing techniques applied to textual metadata and/or audio signal analysis).

1. INTRODUCTION

Multimedia content in the form of audio clips, videos and images is available aplenty on the internet and supervised machine learning algorithms which analyze such multimedia content require accurate labels. Crowdsourcing platforms such as Amazon Mechanical Turk (AMT) have created possibilities for labeling multimedia content by reaching out to a wider audience with different levels of expertise. Such platforms have simplified the process of obtaining labels for multimedia content from human annotators, which, previously, has been an expensive and time-consuming task. Labels gathered from these platforms are noisy since not all workers are dedicated to the task, the task is complex, or there is divergent perception among the workers on how to apply the labels. The quality and other characteristics of the labeled data directly affects how useful these labels are in applications. For example, if labels are used as input to machine learning algorithms, their quality will have a strong impact on performance. For this reason, it is imperative that efficient algorithms are developed that can generate reliable labels, given multiple noisy labels from the crowd. Generating a single, useful, 'correct' label from multiple noisy labels is in itself a challenging task requiring significant research.

Simplistic algorithms that aggregate labels, like majority voting, can refine the noisy labels to a certain extent [4]. The inherent limitation of simple aggregation algorithms, however, is that they require several labels per instance for acceptable quality, and for this reason incur high costs. A way to address the cost is to take the performance of individual workers into account. For example, Ipeirotis et al. [2] developed a quality management technique that involves assigning scalar values to workers by taking into account the quality of the workers' responses. Such a score can then be used as a weight on singular labels, to obtain a more accurate estimation of the aggregated label. In general, however, it remains difficult to outperform a majority-vote baseline, as demonstrated by the MediaEval 2013 Crowdsourcing Task, devoted to social images related to fashion [3].

Conventional computing approaches (for example, signal analysis), can be used to generate labels, and these can be combined with labels contributed by human annotators to achieve a better overall result. Such a combination is interesting in cases in which labels are to be used directly in an application. Investigation of such hybrid approaches that intelligently and effectively combine human input with conventional computation is a secondary area of focus for the Crowdsourcing task. A further area related to hybrid methods is the Active Learning paradigm [5], where the algorithm interactively queries for labels for specific data points which can then be obtained from crowdsourcing.

The remainder of this paper, presents an overview of the task and describe the dataset. We then explain the procedure to collect ground-truth labels and the evaluation metric used for the task.

2. TASK OVERVIEW

The basic objective of the task is to predict labels for all the 15-second music segments in the dataset. The music and the associated information has been retrieved from online music sharing platform Soundcloud[1]. The music segments are represented as triplets: track identifier, start-time and end-time of the 15-second segment. The labels reflect whether or not the segment contains a drop. A *drop* is a characteristic music event in Electronic Dance Music (EDM). Within the EDM community, drop is described as a moment of emotional release where people start to dance like crazy and can be more formally characterized as a building up of tension, which is followed by the re-introduction of the full bassline [1]. For each 15-second segment, the participants predict one of the three labels: *Label 1* segment contains a complete drop, *Label 2* segment contains a partial drop, and *Label 3* segment does not contain a drop.

[1] www.soundcloud.com

MediaEval 2014 Workshop, October 16-17, 2014, Barcelona, Spain

The participants have three sources of information which they can exploit in order to infer the correct label of a music segment: *a*) a set of 'basic human labels', which are labels collected from crowdworkers using an AMT microtask with basic quality control, *b*) the metadata associated with the music tracks (such as title, description, comments), *c*) the audio in the form of mp3 files.Participants were encouraged to use audio signal processing techniques to gain more insight into the music segments. They were also allowed to collect labels by designing their own microtasks (including the quality control mechanism) and running it on a crowdsourcing platform of their choice.

3. TASK DATASET

The dataset for the MediaEval 2014 Crowdsourcing Task consists of music tracks collected from online music sharing platform SoundCloud. The tracks were uploaded to SoundCloud with a Creative Commons Attribution license, which enables the dataset to be used for research purposes. The music tracks and the associated metadata were crawled using the SoundCloud API. An interesting feature of SoundCloud is that the user comments are associated with a timestamp and they refer to a particular time-point in the track. We exploited this feature so as to create a list of short 15-second segments, which might contain a *drop*. We collected all the timed comments which had the word 'drop' in them and extracted a 15-second segment centered at the timestamp of the comment. The dataset comprises 382 tracks belonging to various sub-genres of EDM (e.g., dubstep, electro) and their associated metadata in the form of XML files. The dataset also contains two sets of human generated labels. These labels are given to short 15-second segments from the music tracks based on the occurrence of a drop. The dataset contains a total of 591 15-second music segments. The first set of labels (referred to as 'basic human labels' or 'low fidelity ground truth') has been generated by AMT workers under the application of basic quality control. The second set of labels (referred to as the 'ground truth' or 'high-fidelity ground truth') contains more reliable labels that were created by trusted annotators. The task data was released in a single round and did not have a separate development set.

4. 'BASIC HUMAN LABELS'

Each 15-second music segment in the Crowdsourcing Task data is associated with three labels collected from three crowdworkers.. The crowdworkers listen to the segments in order to make the judgment if the segment should be labeled with *Label 1*, *2*, or *3*. Since we required the workers to be familiar with EDM, we conducted a recruitment task where the workers listen to two EDM tracks and identify the drop moments. Additional questions to judge their familiarity with EDM were asked. Based on their answers, they received a qualification that allowed them to carry out the labeling micro task. Crowdworkers must have answered all questions in the labeling microtask and the answers were required to be consistent. This simple quality control mechanism was designed so that the 'basic human labels' produced using this microtask would have noise levels characteristic of human annotations generated without a sophisticated mechanism for quality control. Out of the 591 assignments on AMT, there was no agreement between the workers for 61 assignments, partial agreement (2 out of 3 workers gave the same response) for 313 assignments and complete agreement (all 3 workers agreed on the same response) for 218 assignments.

5. GROUND TRUTH AND EVALUATION

The ground-truth labels were created by a panel of eight experts. Each music segment was labeled by three different experts and a single label was obtained through majority vote. These trusted annotations served as ground truth for evaluating the task. Out of the labels for 591 music segments, given by experts, there was no agreement between the experts for 33 segments, partial agreement (2 out of 3 experts gave the same response) for 380 segments and complete agreement (all 3 experts agreed on the same response) for 178 assignments. Since we are dealing with a multi-class classification problem (three classes), we chose the weighted F-measure as the official evaluation metric. It is the weighted average of the F-measure for individual classes and the weights are the number of true examples of that particular class.

6. OUTLOOK

The Crowdsourcing for Multimedia Task ran in MediaEval 2014 in its first year as a so-called 'Brave New Task'. The results and interest of participants in the task will inform the development of possible future tasks. In particular, we are interested in understanding how to collect 'basic human labels' in the way most useful for experimentation and how best to create high-fidelity ground truth against which predicted labels can be evaluated. We hope that the experiences this year will help us to develop better methods for studying hybrid human/conventional computation.

7. ACKNOWLEDGMENTS

This task is partly supported by funding from the European Commission's 7th Framework Programme under grant agreement N° 287704 (CUbRIK), N° 610594 (CrowdRec) and N° 601166 (PHENICX).

8. REFERENCES

[1] M. Butler. *Unlocking the Groove: Rhythm, Meter, and Musical Design in Electronic Dance Music*. Profiles in popular music. Indiana University Press, 2006.

[2] P. G. Ipeirotis, F. Provost, and J. Wang. Quality management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 64–67, 2010.

[3] B. Loni, J. Hare, M. Georgescu, M. Riegler, X. Zhu, M. Morchid, R. Dufour, and M. Larson. Getting by with a little help from the crowd: Practical approaches to social image labeling. In *CrowdMM 2014: ACM Multimedia Workshop on Crowdsourcing for Multimedia*, 2014.

[4] S. Nowak and S. Rüger. How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the international conference on Multimedia information retrieval*, MIR '10, pages 557–566, 2010.

[5] B. Settles. Active learning literature survey. Technical report, 2010.