Work Like a Bee - Taking Advantage of Diligent Crowdsourcing Workers

Michael Riegler
Simula
Research Laboratory AS
michael@simula.no

Preben N. Olsen
Simula
Research Laboratory AS
preben@simula.no

Pål Halvorsen
Simula
Research Laboratory AS
paalh@simula.no

ABSTRACT

This paper presents our approach for the Crowd Sourcing Task of the MediaEval 2014 Benchmark. The proposed solution is based on the assumption that the number of Human Intelligence Tasks (HITs) completed by a worker is representative of his diligence, making workers who completing high volumes of work more reliable than low-performing workers. Our approach gives a baseline evaluation indicating the usefulness of looking at the number of task completed by a worker.

1. INTRODUCTION

Crowdsourcing creates a lot of opportunities and is gaining momentum as an area of interest within the multimedia community. Moreover, current web-based services like Amazon Mechanical Turk, Mircrowoker, and Crowdflower have simplified the task of leveraging the power of human computation.

The biggest problem in crowdsourcing is still the reliability of the workers. The information we receive using crowdsourcing is unreliable because of workers who try to trick the system, spam or simply don't understand the task properly. The law of large numbers (LLN) describes how noise is averaged and its effects are removed with a large number of experiments, but increasing the number of experiments directly affects costs. This is why the crowdsourcing exercise for the *Crowdsorting Timed Comments Task* this year focuses on computing correct labels based on noisy crowdsourced, metadata or content information.

Related work in this area can for example be found in [4, 3]. These approaches try to calculate correctness of the workers or use the features of the media files like the global image feature, for a classification.

In contrast, the proposed solution presented here is based on the assumption that workers who complete a high number of tasks are high performers, either because they enjoy the task or that they understand the task well enough to do it efficiently. We believe that both of these circumstances lead to reliable results with respect to HITs. As a secondary approach, we also used labels collected from additional crowdsourcing workers which means that we asked new workers for HITs where the original workers could not come to an agreement.

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

2. APPROACHES

This section describes our two approaches. As mentioned, our main approach is to find the most diligent workers, while the second approach is based on the idea of collecting additional crowdsourcing votes. Quality control is a prerequisite of a well-designed crowdsourcing HIT and to increase the quality of votes for this work, the task organizers included a qualification HIT to make sure that workers understood the task at hand. As the main task was to classify *drops* in music tracks, the workers had to prove that they could classify a drop correctly. Only the workers who passed the qualification HIT were allowed to continue. Because of that pre-quality control, we did not perform any additional quality.

2.1 Diligent Workers

The idea of diligent workers is based on the work presented by Kazai *et al.* [2], which describes five different types of workers: (1) diligent, (2) competent, (3) sloppy, (4) incompetent, and (5) spammers. Diligent workers are identified by the number of completed HITs they produce for a particular task. They also state that most of the HITs are done by the same group of workers. The distribution of workers is a power law distribution and leads to around 54% of single HIT workers for a crowdsorucing tasks. An important insight from this work is that diligent workers can be identified by the number of HITs per task. After comparing the number completed HITs per worker, we chose a subset of diligent workers. The number of workers in this subset is chosen based on the overall distribution of performed HITs between all workers. Experiments on a development set showed that 30% of the best workers leads to a good result. This subset then represents diligent workers who can be trusted and their votes can be used in different ways, e.g., give a higher weight to their votes or only consider their votes.

2.2 Additional Crowdsourcing

For the HITs without a clear result through majority vote between the three provided workers or by weighted subset of the best performing workers, we used additional crowdsourcing. We developed an HTML and SQL-based platform that gave us the opportunity to perform the tests in our lab. The requirement for this additional test was that the participants had to try their best to find the right answer for the HIT.

3. EXPERIMENTAL SETUP

The provided dataset contains 591 songs, metadata, and

Table 1: Configuration of the four different methods evaluated.

Run	Description
R1	MJV with additional crowdsourcing
R2	Diligent workers vote only
R3	MJV with weighted diligent workers
R4	R3 with additional crowdsourcing

Table 2: *MediaEval 2014 Benchmark* results.

Run	WF1-score	True Labels	Predicted Labels
R1	0.7207	183, 63, 291	192, 68, 279
R2	0.6919	183, 63, 291	208, 95, 234
R3	0.6912	183, 63, 291	208, 87, 242
R4	0.6919	183, 63, 291	208, 95, 234

labels generated by human computation, but because some of the songs are duplicates only 537 of them we used in the evaluation. The task's main goal was to classify a drop in music within a limited timespan. A drop can be seen as an event that builds up to a change of the beat or melody in the song, i.e., a characteristic also found in electronic dance music, and is more than just a simple change. Workers could give three different labels to each song segment: (1) the segment contains a complete drop, (2) the segments only contains a partial drop, and (3) the segment contains no drop in music [1].

We assessed four different methods executed in four runs. The results are shown in Table 1 where a summarized overview and short descriptions of each method is provided. The first method (R1) considers the majority vote (MJV) between the three votes provided by the original dataset and additional for not clear answers. While in the second run (R2) we only consider the votes provided by our diligent workers subset. Our third method (R3) takes into account the majority votes, but adds a higher weight to votes provided by diligent workers. The fourth and last method (R4) used the results provided by R3, but with additional crowdsourcing for ambiguous answers (where MJV could not clearly lead to a label).

4. RESULTS

Table 2 describes our benchmark results, while Table 3 describes the results for the most frequent class baseline, in which case all labels get the most frequent class label in the dataset assigned. The performance is measured by the weighted harmonic mean of precision and recall (WF1-score). This is done to avoid unreliable results based on the imbalance of the classes.

We see from Table 2 that every method evaluated outperforms the most frequent class baseline by at least 30%. The best performing method is R1 with a WF1-score of 0.7207. Compared to R1, the three other methods have a performance drop of around 3%. These methods are not distinguishable with respect to the results they produces, which might be because each of the methods rely on the votes provided by the subset of diligent workers. We find it interesting that R3 and R4, which complements diligent workers with MJV and additional crowdsourcing, do not significantly increase performance compared to R2.

Moreover, the performance difference between R1 and R2 is low, which strongly indicates that the assumption of workers who complete the majority of crowdsourcing tasks also perform better is valid. This is a promising insight that can

Table 3: Most frequent class baseline for the given dataset.

Baseline	WF1-score	True Labels	Predicted Labels
MFC	0.3809	183, 63, 291	0, 0, 537

cut costs and yield more accurate crowdsourcing results. For example, by identifying diligent workers early in task execution one can annotate their votes and only consider them as in R2, or weight their votes differently as in R3. That said, we also want to point out that there is a chance our results are dataset specific and further investigations on multiple and larger datasets are needed.

At last, we want to highlight that additional crowdsourcing does not increase the accuracy when considering diligent workers. This indicates that the quality of work and worker motivation is more important than the number of workers used or votes gathered.

5. CONCLUSION

This paper presents two approaches for classifying drops in electronic dance music segments by utilizing human computation and crowdsourcing. The results and insights gained by evaluating four different methods indicate that the proposed approach, which assumes that diligent workers also provide better work quality, is promising. Our investigation also indicates that additional crowdsourcing does not improve results when considering diligent workers.

For assurance and increased certainty, we recognize the need for extending the work to include multiple and larger datasets. Additional future work includes pairing crowdsourcing results with computer generated content analysis and further classification of diligent workers.

6. ACKNOWLEDGMENT

This work has been funded by the NFR-funded FRINATEK project "Efficient Execution of Large Workloads on Elastic Heterogeneous Resources" (EONS) (project number 231687) and the iAD center for Research-based Innovation (project number 174867) funded by the Norwegian Research Council.

7. REFERENCES

[1] M. L. Karthik Yadati, Pavala S.N. Chandrasekaran Ayyanathan. Crowdsorting timed comments about music: Foundations for a new crowdsourcing task. In *MediaEval 2014 Workshop*, Barcelona, Spain, October 16-17 2014.

[2] G. Kazai, J. Kamps, and N. Milic-Frayling. Worker types and personality traits in crowdsourcing relevance labels. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1941–1944. ACM, 2011.

[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 '14, November 03 - 07 2014, Orlando, FL, USA*. ACM, 2014.

[4] M. Riegler, M. Lux, and C. Kofler. Frame the crowd: Global visual features labeling boosted with crowdsourcing information. In *MediaEval 2013 Workshop, Barcelona, Spain*, 2013.