Clustering and Evaluating Without Knowing How To: A Case Study of Fashion Datasets

Daniil Likhobaba, Daniil Fedulov, Dmitry Ustalov

Toloka
11158 Belgrade, Serbia
likhobaba-dp@toloka.ai, mr.fedulow@gmail.com, dustalov@toloka.ai

Abstract

Crowdsourcing allows running simple human intelligence tasks on a large crowd of workers, enabling solving problems for which it is difficult to formulate an algorithm or train a machine learning model in reasonable time. One of such problems is data clustering by an under-specified criterion that is simple for humans, but difficult for machines. In this demonstration paper, we build a crowdsourced system for image clustering and release its code under a free license at https://github.com/Toloka/crowdclustering. Our experiments on two different image datasets, dresses from Zalando's FEIDEGGER and shoes from the Toloka Shoes Dataset, confirm that one can yield meaningful clusters with no machine learning algorithms purely with crowdsourcing.

Introduction

Clustering is the task of grouping objects in such a way that objects in the same group (called a *cluster*) are more similar to each other than to those in other groups (Rokach and Maimon 2005). This is important process in machine learning and arises in many applications, such as text (Jain and Bhattacharjee 1992) and image (Coleman and Andrews 1979) segmentation, data mining (Judd, McKinley, and Jain 1998) and pattern recognition (Hamerly and Elkan 2002). In most cases, clustering is unsupervised task (Grira, Crucianu, and Boujemaa 2004) and it requires knowing the distances between objects (Jain, Murty, and Flynn 1999). However, the distances are often unknown, or clustering rules cannot be clearly defined (Ben Ayed, Ben Halima, and Alimi 2014). Crowdsourcing may help to cope with these problems as such tasks often are trivial for humans (Yuen, King, and Leung 2011). It is known that people can apply their life experience to solve creative tasks (Kittur 2010), such as toxicity detection (Aroyo et al. 2019), relative rankings (Luon, Aperjis, and Huberman 2011), fashion recommendation (Burton et al. 2012), etc.

Although a proper use of crowdsourcing requires a careful task design and quality control setup, recent studies show that it can approximate the distance function between the objects using crowd judgments (Yi et al. 2012; Chen et al. 2018; Chang, Kittur, and Hahn 2016). Some of

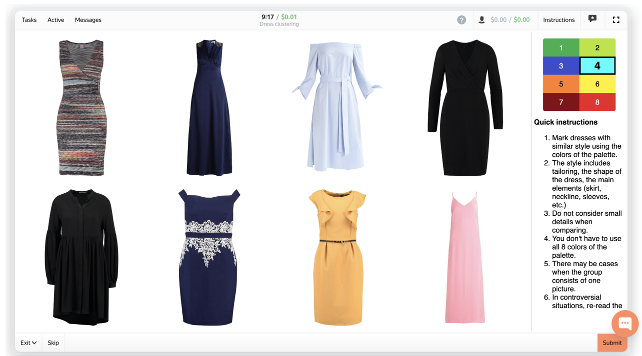

Figure 1: Clustering Task. The worker uses a color palette to highlight similar objects with same color; similar to Gomes et al. (2011).

these papers are theoretical (Mazumdar and Saha 2017; Raman and Varshney 2017), evaluate performance on synthetic datasets (Korlakai Vinayak and Hassibi 2016), or require a prohibitively large number of human tasks to converge (Green Larsen, Mitzenmacher, and Tsourakakis 2020).

In this demonstration paper, we build a system for clustering with crowds, and evaluate it with crowds without involving any machine learning algorithms. We run our experiments on Toloka with two real world datasets, dresses from Zalando's FEIDEGGER (Lefakis, Akbik, and Vollgraf 2018) and shoes from the Toloka Shoes Dataset (Drutsa et al. 2020), and confirm the reproducibility of this method. Also, we release the source code of the built hybrid human-computer system under a free license. We picked the *clustering by style* task as it is difficult to formalize as an algorithm, yet the task itself is relatively easy for humans: each of us can tell whether the style of clothes is similar or not.

Task Design and Worker Selection

To cluster objects, it is necessary to know how similar they are to each other; in the classical formulation, the pairwise matrix of distances is given. If the matrix is not known, it can trivially be approximated with crowds by running a pairwise comparison of all object pairs (Green Larsen, Mitzenmacher, and Tsourakakis 2020). Unfortunately, it generates $\mathcal{O}(N^2)$ tasks for N objects, which is very expensive, i.e.,

Submitted to HCOMP 2022 Works-in-Progress and Demonstration Track. The first two authors made an equal contribution.

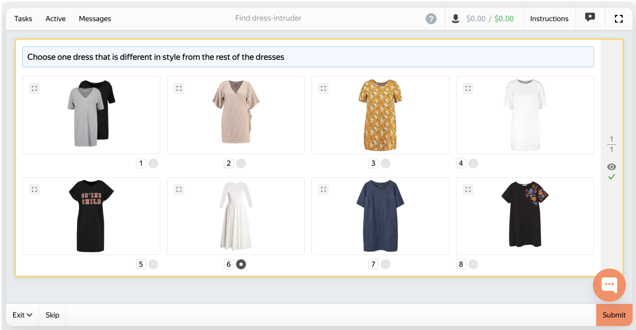

Figure 2: Evaluation Task. The worker has to select the artificially-inserted intruder element (no. 6); similar to Chang et al. (2009).

1000 objects would require roughly 500,000 comparisons, quickly impacting the annotation budget. Hence, there is a need for a task sampling method that is sufficient to divide the objects into meaningful groups as cheaply as possible.

Object Sampling. For clustering, we used the approach proposed by Gomes et al. (2011). For each task, we show M objects and ask the crowd workers to assign a color to each group from the color palette. During prototyping, we found that the optimal choice of M is between 3 and 8 as clustering a large number of objects seems to require additional concentration from the workers, resulting in mistakes, such as failing to color all similar object with the same color. In our setup, every task is completed by three different workers. We sampled each object for $V = \log_2 N \times \log_M N$ times to gather enough information on inter- and intra-relationships of the objects, allowing us to approximate the clustering.

Worker Training. Before starting, workers have to pass a training and a qualification test. The training consists of five pages of tasks, each have more pictures and requires more complex actions than the previous one. Starting with two images on the page and a step-by-step guide and ending with six pictures with more complex instructions. In the training and exam, workers receive a numerical skill value equals to their fraction of correct responses. Only those who achieved the skill value of at least 80 get access to the next step. Training tasks are obvious to everyone, so attentive workers do everything right, and we filter out those who did not understand the task at all. An example of an obvious task is "label all the high-heeled shoes with red color from palette".

Task Design. The task is formulated as *Group the objects by labeling similar ones using color palette*, and its interface is shown in Figure 1. Workers should choose one color and label similar images with it, then choose another color and make another group, etc. Since each item is completely unique, the workers are told not to pay attention to the small details when grouping clothes, but to look at the style as a whole. There is a brief instruction on each page with the main points that should be kept in mind during grouping.

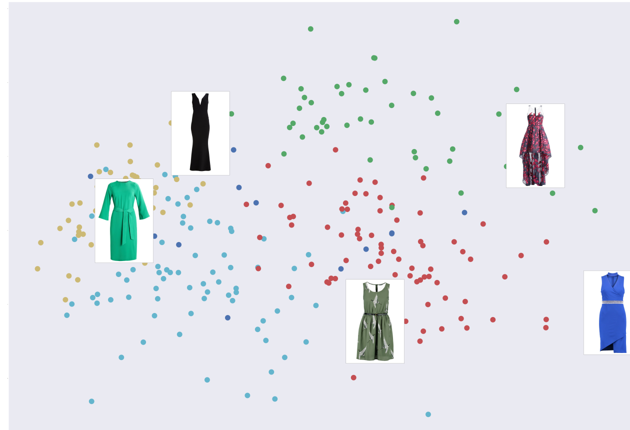

Figure 3: Cluster visualization produced by the Crowdclustering method; dots are objects, colors are clusters.

Clustering with Crowds

Since we have a sparse dataset of noisy labels for the objects, we need a special aggregation method to recover the clusters. For this, we applied and re-implemented using Python a probabilistic model called *Crowdclustering* (Gomes et al. 2011). This approach represents the objects as points in Euclidean space and also allows each worker to group objects by any attribute (e.g. color, material, shape) and works with these groups called *atomic clusters*. Then, atomic clusters are assembled into resulting clusters, the number of which is not a fixed hyper-parameter.

Quality Evaluation

Having annotated and aggregated the clusters, we evaluate the quality of them using an approach called *Intruders* (Chang et al. 2009). For each cluster, we sample from another cluster a random incorrect object called an intruder. Then, we run another crowdsourcing task, in which we ask the workers to select the out-of-style object (Figure 2). The clustering quality is a fraction of times the workers selected the intruder correctly. The quality is considered the better, the more often the workers choose this obviously incorrect object.

We ran the experiments on FEIDEGGER and Toloka Shoes in the same above-described configuration, visualization of result is shown in Figure 3. We found that for the 2000 dress images in FEIDEGGER the quality is 0.83, and for the 87 shoes images in Toloka Shoes Dataset the quality is 0.88.

Conclusion

We found that crowdsourcing allows to obtain a reasonable clustering of objects even when distances between the objects could not be measured at all. It allows using a human-understandable textual instruction instead of metric learning, while being more cost-efficient than the entire distance matrix annotation. We release a Python implementation of the pipeline at https://github.com/Toloka/crowdclustering.

References

Aroyo, L.; Dixon, L.; Thain, N.; Redfield, O.; and Rosen, R. 2019. Crowdsourcing Subjective Tasks: The Case Study of Understanding Toxicity in Online Discussions. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, 1100–1105. New York, NY, USA: Association for Computing Machinery. ISBN 9781450366755.

Ben Ayed, A.; Ben Halima, M.; and Alimi, A. M. 2014. Survey on clustering methods: Towards fuzzy clustering for big data. In *2014 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR)*, 331–336.

Burton, M. A.; Brady, E.; Brewer, R.; Neylan, C.; Bigham, J. P.; and Hurst, A. 2012. Crowdsourcing Subjective Fashion Advice Using VizWiz: Challenges and Opportunities. In *Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '12, 135–142. New York, NY, USA: Association for Computing Machinery. ISBN 9781450313216.

Chang, J.; Gerrish, S.; Wang, C.; Boyd-graber, J.; and Blei, D. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In Bengio, Y.; Schuurmans, D.; Lafferty, J.; Williams, C.; and Culotta, A., eds., *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Chang, J. C.; Kittur, A.; and Hahn, N. 2016. Alloy: Clustering with crowds and computation. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 3180–3191.

Chen, J.; Chang, Y.; Castaldi, P.; Cho, M.; Hobbs, B.; and Dy, J. 2018. Crowdclustering with partition labels. In *International Conference on Artificial Intelligence and Statistics*, 1127–1136. PMLR.

Coleman, G.; and Andrews, H. 1979. Image segmentation by clustering. *Proceedings of the IEEE*, 67(5): 773–785.

Drutsa, A.; Fedorova, V.; Ustalov, D.; Megorskaya, O.; Zerminova, E.; and Baidakova, D. 2020. Crowdsourcing Practice for Efficient Data Labeling: Aggregation, Incremental Relabeling, and Pricing. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, SIGMOD '20, 2623–2627. New York, NY, USA: Association for Computing Machinery. ISBN 9781450367356.

Gomes, R.; Welinder, P.; Krause, A.; and Perona, P. 2011. Crowdclustering. In Shawe-Taylor, J.; Zemel, R.; Bartlett, P.; Pereira, F.; and Weinberger, K., eds., *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

Green Larsen, K.; Mitzenmacher, M.; and Tsourakakis, C. 2020. Clustering with a Faulty Oracle. In *Proceedings of The Web Conference 2020*, WWW '20, 2831–2834. New York, NY, USA: Association for Computing Machinery. ISBN 9781450370233.

Grira, N.; Crucianu, M.; and Boujemaa, N. 2004. Unsupervised and semi-supervised clustering: a brief survey. *A review of machine learning techniques for processing multimedia content*, 1: 9–16.

Hamerly, G.; and Elkan, C. 2002. Alternatives to the K-Means Algorithm That Find Better Clusterings. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, CIKM '02, 600–607. New York, NY, USA: Association for Computing Machinery. ISBN 1581134924.

Jain, A. K.; and Bhattacharjee, S. 1992. Text segmentation using gabor filters for automatic document processing. *Machine Vision and Applications*, 5(3): 169–184.

Jain, A. K.; Murty, M. N.; and Flynn, P. J. 1999. Data Clustering: A Review. *ACM Comput. Surv.*, 31(3): 264–323.

Judd, D.; McKinley, P.; and Jain, A. 1998. Large-scale parallel data clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8): 871–876.

Kittur, A. 2010. Crowdsourcing, Collaboration and Creativity. *XRDS*, 17(2): 22–26.

Korlakai Vinayak, R.; and Hassibi, B. 2016. Crowdsourced Clustering: Querying Edges vs Triangles. In Lee, D.; Sugiyama, M.; Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Lefakis, L.; Akbik, A.; and Vollgraf, R. 2018. FEIDEGGER: A Multi-modal Corpus of Fashion Images and Descriptions in German. In *LREC 2018, 11th Language Resources and Evaluation Conference*.

Luon, Y.; Aperjis, C.; and Huberman, B. A. 2011. Rankr: A mobile system for crowdsourcing opinions. In *International Conference on Mobile Computing, Applications, and Services*, 20–31. Springer.

Mazumdar, A.; and Saha, B. 2017. Clustering with Noisy Queries. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Raman, R. K.; and Varshney, L. R. 2017. Budget-optimal clustering via crowdsourcing. In *2017 IEEE International Symposium on Information Theory (ISIT)*, 2163–2167. IEEE.

Rokach, L.; and Maimon, O. 2005. *Clustering Methods*, 321–352. Boston, MA: Springer US. ISBN 978-0-387-25465-4.

Yi, J.; Jin, R.; Jain, S.; Yang, T.; and Jain, A. 2012. Semi-crowdsourced clustering: Generalizing crowd labeling by robust distance metric learning. *Advances in neural information processing systems*, 25.

Yuen, M.-C.; King, I.; and Leung, K.-S. 2011. A Survey of Crowdsourcing Systems. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, 766–773.