Assessing Dawid-Skene on a Crowdsourced Dataset

Tarmo Pungas
14222515

Abstract

This report investigates the efficacy of the Dawid-Skene model in the realm of crowdsourced data annotation, contrasting its performance against the conventional majority voting method. The study aims to unravel the potential and limitations of the Dawid-Skene model in leveraging the collective intelligence of diverse individuals. Utilizing a real-life dataset, the Toloka Aggregation Relevance 2, the experiment reveals that Dawid-Skene outperforms majority voting, achieving an F1 score of 0.79 compared to 0.76. The findings contribute to the understanding of human-machine collaboration, emphasizing how modeling individual annotator accuracy enhances crowdsourced data annotations.

1 Introduction

In the dynamic landscape of artificial intelligence, the integration of human expertise has become increasingly important. Human-in-the-Loop Machine Learning focuses on the relationship between machine intelligence and human intuition. Within this paradigm, crowdsourcing is as a powerful tool, harnessing the collective wisdom of diverse individuals to tackle complex problems. This report delves into the assessment of the Dawid-Skene model in the context of a crowdsourced dataset, shedding light on its efficacy in comparison to the conventional method of majority voting.

Crowdsourcing offers a scalable and often cost-effective solution to tasks that demand human cognitive abilities. By distributing tasks to a large pool of contributors, crowdsourcing enables the aggregation of diverse perspectives, tapping into the collective intelligence of individuals with varied expertise and experiences. Furthermore, it is sometimes difficult or near impossible to determine ground truth labels with great confidence. After asking many different people, the consensus is usually highly correlated with the ground truth.

The subsequent sections will explore the specifics of the experimental design, data characteristics, and the methodology employed to evaluate the Dawid-Skene model against majority voting on a novel dataset. The hypothesis is that Dawid-Skene performs better than majority voting. This comparative analysis should offer insight of human-machine collaboration, unraveling the potential and limitations of the Dawid-Skene model in a real-world, crowdsourced setting.

2 Background & Related Work

The **Dawid-Skene model** is a probabilistic framework designed to uncover the expertise of human annotators in a crowdsourced environment [1]. Developed to address the challenges of annotator bias and inconsistency, the model enables discerning true labels from noisy annotations. By estimating both annotator accuracy and the underlying ground truth, the Dawid-Skene model unveils latent patterns in the data and allows us to evaluate and exclude annotators based on their capabilities, improving the performance of machine learning algorithms.

The model operates on the premise that each annotator has some degree of expertise, influencing the accuracy of their annotations. Formally, let t_n denote the ground truth for a given instance n. The observed annotation, denoted as $y_{n,l}$, represents the label assigned by annotator l to instance n. The probability of annotator l reporting class j when the true class is k can then be expressed as:

$$\hat{\pi}_{k,j} = \frac{\sum\limits_{n=1}^{N} \hat{p}_{n,k}^{t} \cdot y_{n,l,j}}{\sum\limits_{i=1}^{K} \sum\limits_{n=1}^{N} \hat{p}_{n,k}^{t} \cdot y_{n,l,i}}. \tag{1}$$

A common baseline method for evaluating crowdsourcing models is **majority voting**, a straightforward approach wherein the label assigned by the majority of annotators is considered the final annotation. This simplistic method, while easy to implement, does not account for individual annotator expertise and may be susceptible to noisy annotations.

Since the introduction of the original Dawid-Skene model, improvements have been suggested. **Fast Dawid-Skene** is an enhanced EM-based algorithm for sentiment classification, exhibiting faster convergence while maintaining competitive accuracy [4]. The algorithm proves particularly effective for real-time sentiment annotation, addressing the challenge of scarce labeled data in machine learning training.

Beyond the Dawid-Skene model, various approaches have been proposed to address challenges in crowdsourced data annotation. Notable among these are the **Multi-Annotator Competence Estimation (MACE)** model [2], the **Generative model of Labels, Abilities, and Difficulties (GLAD)** [5], and the **Karger, Oh, and Shah (KOS)** model [3].

MACE confronts the inherent challenges of non-expert annotation services, such as those provided by Amazon's Mechanical Turk. The model employs an item-response framework to learn, in an unsupervised manner, the trustworthiness of annotators and predict the correct underlying labels. The model shows significant improvements over standard baselines in both predicted label accuracy and trustworthiness estimates, demonstrating resilience even under adversarial conditions, making it a valuable contribution to crowdsourced data annotation.

GLAD addresses the need for large databases of hand-labeled images in modern machine learning-based computer vision systems. With a probabilistic approach, GLAD simultaneously infers image labels, annotator expertise, and image difficulty. The model surpasses majority voting and exhibits robustness to both noisy and adversarial labelers. GLAD's contribution lies in providing a principled solution to the challenges posed by varying levels of annotator expertise and image difficulty in large-scale image labeling tasks.

KOS tackles the reliability of crowdsourcing systems by considering a general model of crowdsourced tasks. It formulates the problem of minimizing the total cost, measured as the number of task assignments, to achieve a target overall reliability. The algorithm outperforms majority voting and demonstrates optimality compared to an oracle that knows the reliability of every worker. This algorithm, inspired by belief propagation and low-rank matrix approximation, provides valuable insights into minimizing costs while maintaining reliability in crowdsourced systems.

3 Methods

The dataset used in this experiment, **Toloka Aggregation Relevance 2**[1], was published by Toloka, a crowdsourcing platform developed by Yandex. The anonymized dataset contains information on annotators' binary judgments for the relevance of items to a single query. It contains the annotations of 7,138 different people for 99,319 tasks, for a total of **475,536 labels**. The data is clean and no preprocessing was needed.

The distribution of annotations is skewed, with most people having annotated less than 100 tasks (with a median of 40) but some close to 1,400 (Figure 1). Class-wise, the dataset is fairly balanced, with 58% of labels belonging to class 1 and 42% to class 0. There is also a small subset of the tasks (10%) with ground truth labels. The median accuracy of the annotators on the ground truth data is 66%.

[1] `https://toloka.ai/datasets/?datasets-category=crowdsourcing#datasets`

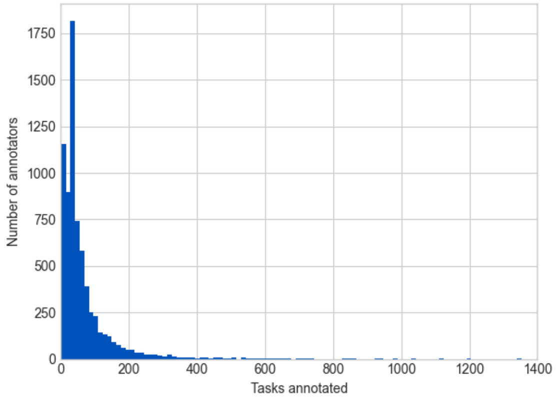


Figure 1: Histogram of annotations per person.

The **Crowd-Kit library**[2] provided the codebase for implementing both Dawid-Skene and Majority Vote models in this study. Crowd-Kit is an open-source library that offers efficient implementations of various quality control methods, including aggregation, uncertainty, and agreements. The only hyperparameter that was necessary to specify was the maximum number of expectation–maximization iterations, which was set at $n\_iter = 100$.

F1 score was selected as the performance metric. This was because of the underlying assumption that both positive and negative classes matter equally. For a given query, we wouldn't want to show irrelevant documents but also wouldn't want to leave out relevant documents. In order to assess the quality of a given annotator under Dawid-Skene, individual accuracy was calculated using the confusion matrices.

All experiments run in the project were implemented with Python, using Jupyter Notebook. The code can be freely accessed on GitHub[3].

4 Experimental Results

After training the Dawid-Skene model on the whole dataset, the **discrepancy between annotators' performance** becomes clear. Figure 2 demonstrates that many annotators perform as good or worse than chance, according to the individual accuracies computed from Dawid-Skene's confusion matrices. Without access to the methodology used to collect this data, it is impossible to determine whether the discrepancy in accuracy is due to selection bias (some annotators choosing to label only easier tasks), exploitation (perhaps the annotators were being paid per label), or simply expertise.

As for the performance of the Dawid-Skene model, it achieved an F1 score of **0.79** on the full dataset. This is a slightly better result than the baseline, majority voting, which reached an F1 score of **0.76**. The results **confirm the original hypothesis**, implying that modeling individual annotators' accuracy allows Dawid-Skene to extract more detailed information from the same amount of data.

In order to see how well the Dawid-Skene model of annotators' accuracy correlates with the ground truth, we can remove low-performing annotators, retrain the model and calculate the F1 score again. Figure 3 illustrates **how accurately the Dawid-Skene model deduces annotator quality**. In the figure, the accuracy threshold determines how many annotators we discard before training the new model. The accuracy of every annotator is determined by the Dawid-Skene model on the full dataset. When the threshold is 0, we use all data. When the threshold is 100, we discard all annotators whose accuracy is below 100.

Care should be taken when interpreting Figure 3. The F1 score is calculated using ground truth labels. However, since not all people annotated all tasks, the test data also keeps getting smaller (Figure 4b) in size as we decrease the number of annotators (Figure 4a). Therefore, Figure 3 does not provide insights into how well the models would perform when retaining only the best annotators and

[2] `https://github.com/Toloka/crowd-kit`

[3] `https://github.com/tarmopungas/hlml-project/tree/main`

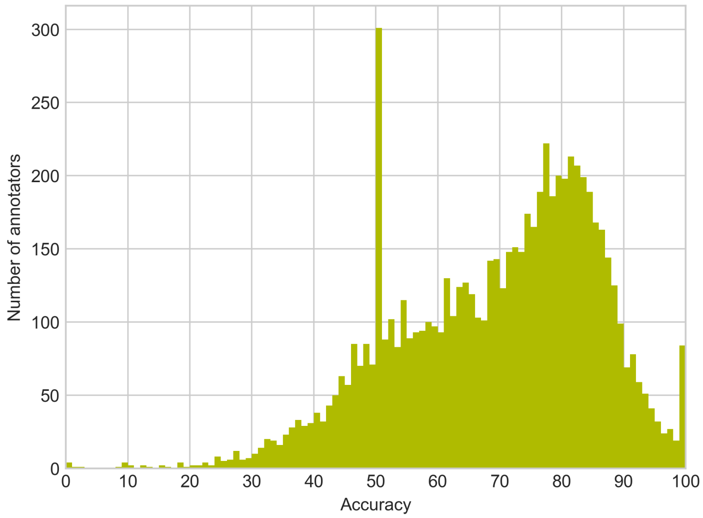


Figure 2: Histogram of annotators' accuracy according to Dawid-Skene.

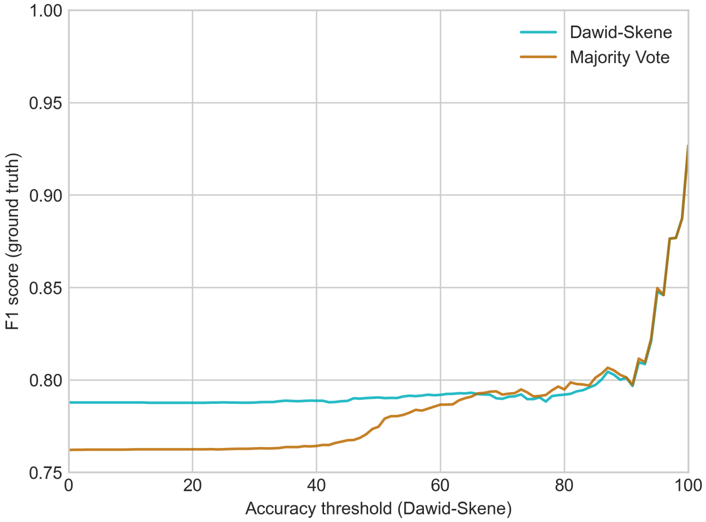
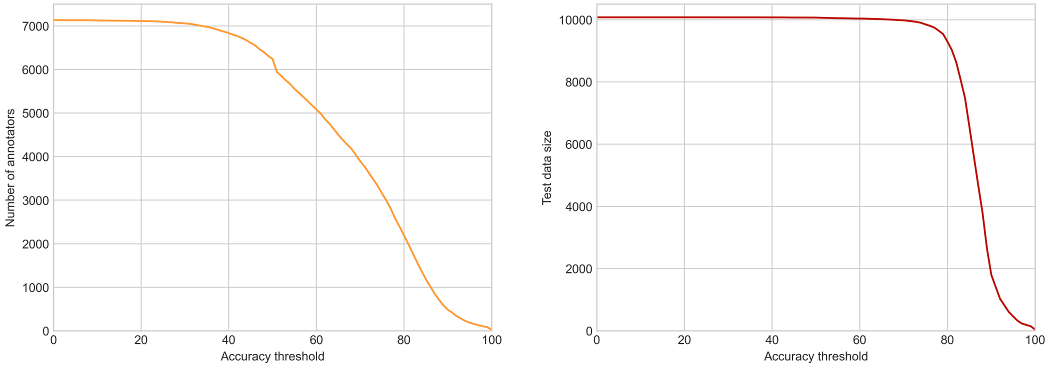


Figure 3: Model performance on different thresholds of individual annotator accuracy.

excluding the others. Instead, it shows how accurately the Dawid-Skene model assesses annotator quality. The fact that both Dawid-Skene and majority voting converge as the accuracy threshold increases, shows that the **best annotators are so accurate as to render negligible differences** between the models.

To mitigate the problem of decreasing test data, a new experiment was designed. Instead of evaluating the model on only the data annotated by any remaining annotator, the missing test data was evaluated on the base model (trained on the whole dataset), using it as a form of fallback. If the performance were to increase with this approach, it would imply that using fewer and higher quality annotators leads to a better outcome. However, the results of this experiment showed a consistently steady, albeit slightly fluctuating F1 score, suggesting that **fallback is not a viable approach** for improving the model and affirming the wisdom of the crowd (Figure 5).



(a) # of annotators with accuracy over the threshold.

(b) Size of test data per accuracy threshold.

Figure 4: Decrease of annotators and test data corresponding to Figure 3.

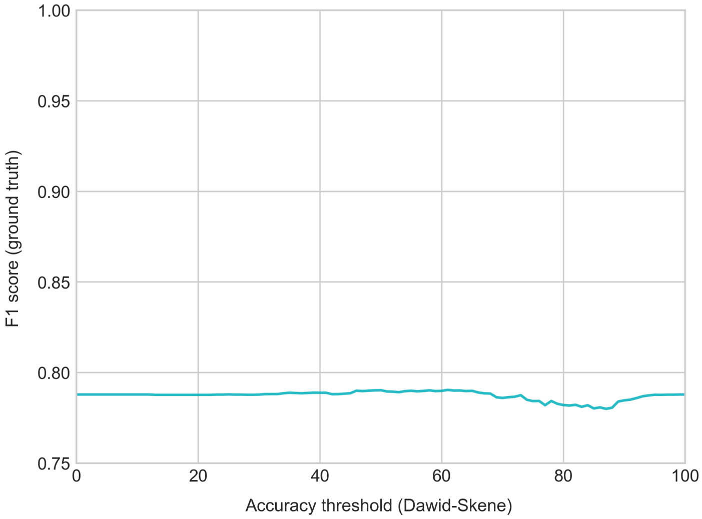


Figure 5: Dawid-Skene performance on annotator accuracy thresholds, using fallback.

5 Conclusion, Limitations, & Future Work

This study explored the efficacy of the Dawid-Skene model in the context of a crowdsourced dataset, comparing its performance to the conventional majority voting method. The Dawid-Skene model, designed to uncover the expertise of human annotators, demonstrated superior performance with a slightly higher F1 score of 0.79 compared to majority voting's score of 0.76. This outcome supports the hypothesis that modeling individual annotators' accuracy allows Dawid-Skene to extract more information from the same data.

The examination of annotators' performance revealed significant variations, indicating potential issues such as selection bias, exploitation, or varying levels of expertise among annotators. Additionally, the experiment to assess Dawid-Skene's model on annotator accuracy thresholds emphasized the model's accuracy in discerning annotator quality.

Despite the successful outcomes, this study has limitations. The dataset, while informative, is only one example and cannot represent all real-world crowdsourcing scenarios. In addition, there was limited information available regarding the methodology of collecting said data. The experiment's design, particularly in addressing diminishing test data, did not yield a viable approach for model improvement, suggesting a need for alternative strategies.

Future work could focus on refining experimental methodologies to address challenges related to annotator variability and data scarcity. Exploring alternative crowdsourcing models beyond Dawid-Skene, such as MACE, GLAD, and KOS, may offer comparative insights and offer more competitive comparisons than majority voting. Additionally, investigating other real-world applications of crowdsourcing, potentially in domains like sentiment analysis, could further validate the practicality of these models.

Appendix: ChatGPT prompts

You can access the ChatGPT prompts I used to help write this report here: `https://chat.openai.com/share/888c71b0-5dfa-4106-a9d8-3cf75bb8a56a`.

References

[1] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):20–28, 1979.

[2] Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. Learning whom to trust with MACE. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff, editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[3] David Karger, Sewoong Oh, and Devavrat Shah. Budget-optimal task allocation for reliable crowdsourcing systems. *Operations Research*, 62, 10 2011.

[4] Vaibhav B Sinha, Sukrut Rao, and Vineeth N Balasubramanian. Fast dawid-skene: A fast vote aggregation scheme for sentiment classification, 2018.

[5] Jacob Whitehill, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier R. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Neural Information Processing Systems*, 2009.