Active Learning with Human Heuristics: An Algorithm Robust to Labelling Bias

Sriram Ravichandran[+], Nandan Sudarsanam[+], Konstantinos Katsikopoulos[†],Balaraman Ravindran[+]
[+]Indian Institute of Technology, Madras
[†]University of Southampton,UK

Abstract

Active learning(AL) enables prediction algorithms to achieve better performance with fewer data points by adaptively querying an oracle for output labels. In many instances, the oracle is a human. According to behavioral sciences, humans provide labels by employing decision heuristics which tend to offer biased labels.AL algorithms trained with such labels could in turn provide incorrect predictions, which could make the decisions made by such models unfair. How would modelling the oracle with such heuristics affect the performance of AL algorithms? We investigate three human heuristics (fast-and frugal tree, tallying, and franklin's rule) combined with four active learning algorithms (entropy-based, multi-view learning, density-based, and novel density-based) and apply them to five datasets from domains such as health, wealth and sustainability. A first novel finding is that if a heuristic leads to significant labelling bias, the performance of active learning algorithms significantly drops, sometimes below random sampling. Thus, it is key to design active learning algorithms robust to labeling bias. Our second contribution is a novel density-based algorithm that achieves an overall median improvement of 31% over current algorithms when the oracle has a significant labelling bias. In sum, designing and benchmarking active learning algorithms should incorporate the modelling of human decision heuristics.

1 Introduction

AI is being used in various significant applications that affect human lives. These include recruitment, consumer lending, healthcare, criminal justice, etc. Building prediction models is crucial for automating such decision processes because it enables decisions based on data rather than relying solely on intuition or past experiences. There is an increasing need for training such models in conditions where obtaining labels is significantly more expensive than their attributes. Moreover, due to the sensitivity of the applications the trained models is also be expected to be fair i.e. devoid of bias that exists when a human makes a decision. Active learning (AL) algorithms have the leverage of choosing the data points to be queried at each instance, thereby reaching the benchmark accuracy with fewer queries (labeled instances). A typical active learner starts with a small number of labeled instances and queries for one or more unlabeled instances, then selects additional points to query based on the labels obtained from previous queries. Labeling the queried instances can be done in multiple ways and is therefore typically assumed to be an unbiased random response. For example, building a model to predict the durability of a car involves crash-testing cars to obtain labels that are highly expensive, making this a suitable application for AL algorithms. However, a substantial subset of AL-based querying involves a human annotator. For instance, A review of AL papers searched with the keyword "Active Learning" that were published during 2021-2023 across prominent venues such as Nature Communications, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Journal of Machine Learning Research and Advances in Neural Information Processing Systems shows that about 63% of the works involved the usage of human-annotated labels.Traditional literature in behavioral economics[1] highlights the deviation of the human decision-making process from rationality, which they defined as bias. Providing labels for AL should be no exception.

However, annotator bias and its implications on trained models are acknowledged in only a small subset of AL literature. For instance, Deepesh et.al.[2] noticed that behavioral biases in the oracle decrease the classification accuracy of prediction models built by at least 20%. Moreover, Burr et.al.[3], in their extensive literature survey on AL, mentioned the reliability of the labels provided by humans might be compromised due to difficulties faced in comprehending the instances that might impact the quality of the labels obtained.

This understanding resulted in development of a class of AL algorithms that considers the biases present in the human oracle.

Works belonging to this class[4, 5] considered the presence of human bias as random or a uniformly distributed error while proposing novel algorithms.J.Du et.al.[6] on the other hand, proposed an algorithm with an exploration and exploitation approach by relabeling data points that could be wrongly labeled. The oracle here was modeled based on the assumption that the probability of obtaining biased labels depends on the maximum posterior probability of an instance computed with the ground truth labels.

In all the above works, the oracle was assumed to offer incorrect responses randomly, or the label bias was synthetically injected based on certain assumptions. However, Herbert Simon, the founder of bounded rationality, argues that people must utilize approximations for the majority of tasks, including simple decision heuristics[7]. Additionally, Gigerenzer et al.[8] pointed out several human heuristics existent under bounded rationality that the human mind tends to follow as its incapable of superhuman reasoning.

The above works support that human oracle is likely to use decision strategies during annotations, and the label bias tends to result from the heuristic used. This makes it essential to study the effect of decision strategies on the active learning models since a model trained with an unfair human decision strategy could make unfair decisions.

This study contributes to the active learning literature by asserting that the decision strategy used by the oracle significantly affects the relative performance of AL algorithms, thereby necessitating the need to benchmark AL algorithms with human decision strategies. We also propose a novel AL algorithm that pioneers the birth of a new class of algorithms built based on human decision strategies.

The rest of the paper has been structured as follows. The methodology is laid forth in Section 2, including explanations of the datasets, AL algorithms, and human heuristics utilized in the study. After discussing the results in Section 3, Section 4 concludes by summarising the same.

2 Methodology

Typically, the active learner chooses the instance to obtain label(x_i) from the pool of unlabeled instances(X) sequentially based on its query strategy and queries the same to the Human. The labels thus obtained(y_i) train the AL after every query. In our study, we mimic the functionality of the human oracle using fast and frugal heuristics such as the fast and frugal tree (FFT), tallying, and a conventional heuristic(Franklin's rule). The decision strategies ensure that the bias labels provided to the oracle are not random but are based on the instance for which querying is done.[see section2.1]

To perform the experiments, we chose five labeled data sets from various domains such as Health[Cleveland Heart disease[9]], Wealth[To predict fraudulent firm[10]], Automobile[Car Condition prediction[11]], Food science[Wine Prediction[12]] and Sustainability[Biodegradable Data set[13]].

For our study, we considered the pool-based sampling scenario where the pool of instances is ranked based on the query strategy. The active learner then selects the best query based on these ranks. The AL algorithms considered were Entropy Sampling, Multi-view learning with co-testing, Conventional Density-based learning, and Novel Density-based learning[see section 2.2]

2.1 When is a Fast and Frugal Decision strategy likely to provide an unbiased label?

To get a rational understanding of situations where fast and frugal heuristics(FFT and Tallying) provide incorrect labels, We postulate the following hypothesis:

Hypothesis 1 *Data points whose attribute values are farther away from their corresponding mean attribute value are less prone to obtaining biased labels from human oracle/heuristics.*

The above hypothesis was formulated based on the intuition that the decisions made by Fast and frugal heuristics always compare the attribute values to constant values. In FFT and Tallying, this constant value tends to be the mean attribute value.

This hypothesis can be illustrated with a case where the task is to classify a car's condition based on its usage period (Let the average usage be five years). Intuitively, the human oracle would find it easier to classify cars that are 2/10 years old than a car that has been used for five years. i.e., Cars with attribute values closer to their mean.

On the datasets taken into consideration, fast and frugal heuristics were employed to produce predictions in order to test the hypothesis. Table 1 and Table 2 show that the prediction accuracy of the heuristics was significantly higher for data points that were farther away from the mean(FM) compared to data points that were closer to the mean(CM), thereby supporting our claim.

Sr.No.	Data-set Name	FM(%)	CM(%)	Overall(%)
1	Biodegradable Data set	78.74	73.33	77.02
2	Car Prediction	80.61	68.56	71.29
3	Cleveland Heart Disease Data set	95.45	83.83	84.72
4	Audit Dataset	96.5	94.4	95.7
5	Wine Dataset	100	86.7	87.07

Table 1: Accuracy of Predictions made by Tallying heuristic

Sr.No.	Data-set Name	FM(%)	CM(%)	Overall(%)
1	Biodegradable Data set	76.44	57.33	70.97
2	Car Prediction	94.1	88.5	92.59
3	Cleveland Heart Disease Data set	81.25	80.07	81.25
4	Audit Data	96.5	91.1	94.42
5	Wine Data set	100	97.1	97.75

Table 2: Accuracy of Predictions made by FFT heuristic

2.2 Novel Density-based Learning

The experimentally supported hypothesis(section 2.1) motivates the development of a query strategy that queries data points whose attribute values are farther away from their mean attribute value. It must also be noted those instances tend to have lower cosine Information density values. Existing algorithms, such as conventional density-based learning, are based on metrics directly proportional to entropy and cosine similarity. This makes them prefer querying data points more susceptible to obtaining biased labels. Hence, we consider a modified metric:

$$H(x) = -\frac{\sum_k p_k \log(p_k)}{(\frac{1}{U})\sum_{u=1}^{U} sim(x, x^u)} \tag{1}$$

As the above formula indicates, the data points are ranked based on their similarity to other unlabeled data points in the pool set($\frac{1}{U}\sum_{u=1}^{U} sim(x, x^u)$) as well as the entropy measure. U represents the pool of unlabeled instances after every query. The metric is expected to motivate the learner to query data points with high entropy and low information density, i.e.query data points that are useful and tend to obtain accurate labels.

3 Results and Discussion

The AL models were trained based on the labels produced by human heuristics. This was repeated for every heuristic-AL algorithm-decision strategy combination, and the trained model's accuracy was measured after each query. Conventional studies involve the evaluation of AL algorithms using Learning curves(Accuracy vs. data points queried). However, it is reasonably apparent to expect a decrease in the accuracy of both AL algorithms and random sampling across data points queried when labels are provided due to biased decision strategies. Thereby, evaluating algorithms based on absolute accuracy is redundant in this study.

However, the relative accuracy of AL algorithms compared to that of Random sampling would help understand the comparative effectiveness within active learning algorithms in the presence of decision strategies. Hence we introduce a particular metric, 'Leverage'[L_i], to visualize the same.

$$L_i = AL_i - RandomSampling_i \tag{2}$$

Here,AL_i and $RandomSampling_i$ represent the accuracy obtained by the respective query strategies after "i" no. of queries.

Furthermore, in order to find the relative robustness within the AL algorithms, we assess the decrease in the effectiveness of AL algorithms observed due to the influx of decision strategies i.e., drop in leverage across the learning phase[∇_i]:

$$\nabla_i = [L_i]_{Ground} - [L_i]_{DecisionStrategy} \tag{3}$$

In Eqn.3, $[L_i]_{DecisionStrategy}$ represents the active learning algorithm's leverage after obtaining labels as a result of the "Decision Strategy" for "i" queries.

The Leverage curve/Drop in leverage curve plotted based on the above help in representing both the absolute effectiveness and drop in the efficacy of AL algorithms when the fast and frugal heuristics provide significantly incorrect labels(see Appendix).

Absolute Leverage	Entropy(%)	MVL(%)	Proposed(%)	Conventional(%)	Improvement(%)
Biodegradable-FFT	1.59	1.53	2.68	-1.5	68.29
Biodegradable-Tallying	2.26	1.97	2.66	1.63	17.62
Car Rate-FFT	1.11	0.51	1.24	0.34	12.15
Car Rate-Tallying	0.52	0.36	1	-1.76	92.03
Clevland Heart-FFT	0.44	0.49	0.53	0.46	9.41
Clevland Heart-Tallying	1.76	1.66	1.46	1.75	-16.82
Wine-Tallying	3.74	3.62	3.58	3.76	-4.91

Drop in Leverage	Entropy(%)	MVL(%)	Proposed(%)	Conventional(%)	Decrease in drop(%)
Biodegradable-FFT	8.64	8.13	5.62	10.52	30.9
Biodegradable-Tallying	7.08	7.02	4.71	6.86	31.34
Car Rate-FFT	0.11	0.13	-0.45	0.51	524.92
Car Rate-Tallying	0.69	0.29	-0.17	2.46	159.01
Clevland Heart-FFT	0.5	0.58	0.83	0.2	-317.12
Clevland Heart-Tallying	-0.043	0.037	0.979	-0.356	-374.89
Wine-Tallying	2.59	2.69	2.09	2.56	18.5

Figure 1: Top-Avg. leverage of AL algorithms, Bottom-Avg. drop in Leverage of AL algorithms

Figure 1 represents the average Absolute and Drop in Leverage experienced by the AL algorithms through the learning phase(until convergence) specifically in scenarios where fast and frugal heuristics(FFT and Tallying) provided significantly incorrect labels.

The proposed density-based learning performs better than other algorithms by showing a median improvement of 11% and a median decrease in a drop of 31% compared to the best-performing algorithm. The notable reduction in drop-in leverage demonstrates the robustness of the proposed algorithm. When heuristics like Franklin's rule gave mostly close-to-ground truth labels, the algorithm was not discovered to perform the best. As a result, the suggested approach is subjected to be used only in situations where heuristics provide considerably biased labels.

4 Conclusion

The primary motive of the work was to model the oracle with human heuristics, which enabled the study of human heuristics' impact on AL algorithms. The same was achieved with three human heuristics(Fast and frugal tree(FFT), Tallying, Franklin's rule), four AL algorithms(Entropy based, Multi-view Learning, Density-based, Novel-density based), and five data sets. The performance of AL algorithms decreased considerably when human heuristics provided significantly incorrect labels. This necessitated a novel algorithm robust to bias labels provided by decision strategies. Our empirically proven hypothesis that heuristics tend to provide correct labels when queried data points with attribute values farther from the mean led to a novel density-based AL algorithm.

The proposed density-based learning algorithm improved absolute leverage by 11% in comparison to the best-performing algorithm. Moreover, the median decrease in drop-in leverage was 31% making the proposed algorithm a preferred one. The ability of the proposed algorithm to query instances that are likely to provide accurate labels and its lesser dependency on the labels obtained attributes to its good performance. On the other hand, when biased labels provided by the human heuristics were minimal, the proposed algorithm was not found useful, thereby restricting its usage in such scenarios.

In sum, the variation in the relative performance of Active Learning algorithms w.r.t decision strategies advocates the need for bench-marking algorithms in existing AL literature using the decision strategy framework proposed in the study. Moreover, the findings strongly motivate the need for a new era of algorithms in the AL domain that considers the uncertainty of the oracle while providing labels on instances, one of which has been achieved in this study.

References

[1] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

[2] Deepesh Agarwal, Obdulia Covarrubias-Zambrano, Stefan Bossmann, and Balasubramaniam Natarajan. Impacts of behavioral biases on active learning strategies. In *2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pages 256–261, 2022.

[3] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[4] Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 614–622. Association for Computing Machinery, 2008.

[5] Perry Groot, Adriana Birlutiu, and Tom Heskes. Learning from multiple annotators with gaussian processes. In Timo Honkela, Włodzisław Duch, Mark Girolami, and Samuel Kaski, editors, *Artificial Neural Networks and Machine Learning – ICANN 2011*, pages 159–164, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[6] J. Du and C. Ling. Active learning with human-like noisy oracle. *2010 IEEE International Conference On Data Mining*, pages 797–802, 2010.

[7] Herbert A. Simon. Invariants of human behavior. *Annual Review of Psychology*, 41(1):1–20, 1990.

[8] Gigerenzer, Peter Todd, Jean Czerlinski, Jennifer Davis, Gerd Gigerenzer, Daniel Goldstein, Adam Goodie, Ralph Hertwig, Ulrich Hoffrage, Kathryn Laskey, Laura Martignon, and Geoffrey Miller. *Simple Heuristics That Make Us Smart*. 01 1999.

[9] Jeroen Eggermont, Joost Kok, and Walter Kosters. Genetic programming for data classification: Partitioning the search space. volume 2, pages 1001–1005, 03 2004.

[10] N. Hooda, S. Bawa, and P. Fraudulent Firm Classification: A Rana. Case study of an external audit. *Applied Artificial Intelligence*, 32:48–64, 2018.

[11] Ivan Bratko Demsar.J Zupan. B, Marko Bohanec. Machine learning by function decomposition. In *International Conference on Machine Learning*, 1997.

[12] Olivier Y.de Vel Aeberhard. S, Danny Coomans. Improvements to the classification performance of rda. *Journal of Chemometrics*, 7, 1993.

[13] K. Mansouri, T. Ringsted, D. Ballabio, R. Todeschini, and V. Consonni. Quantitative structure–activity relationship models for ready biodegradability of chemicals. *Journal Of Chemical Information And Modeling*, 53:867–878, 2013.

A Appendix

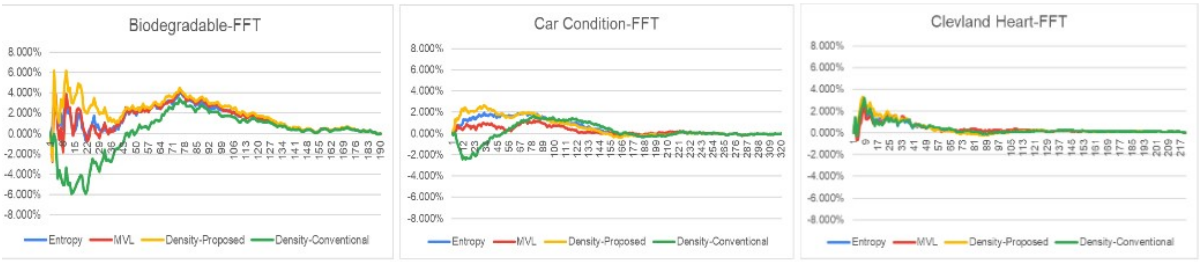

Figure 2: Leverage curves of active learning algorithms when the oracle provided labels with significant bias are a result of FFT

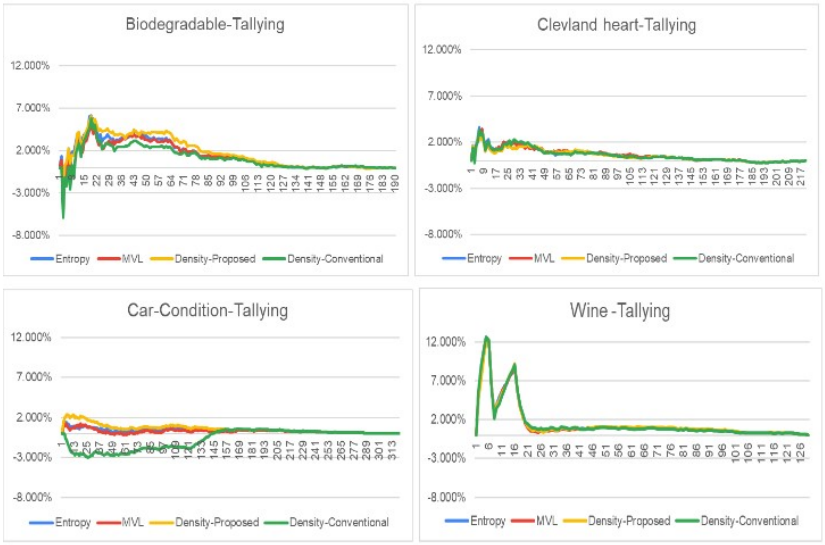

Figure 3: Leverage curves of active learning algorithms when the oracle provided labels with significant bias as a result of tallying heuristic

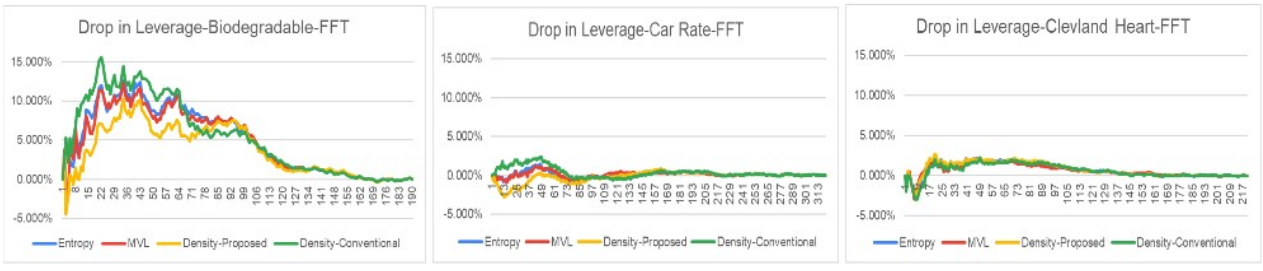

Figure 4: Drop in leverage across the learning phase of active learning algorithms for FFT

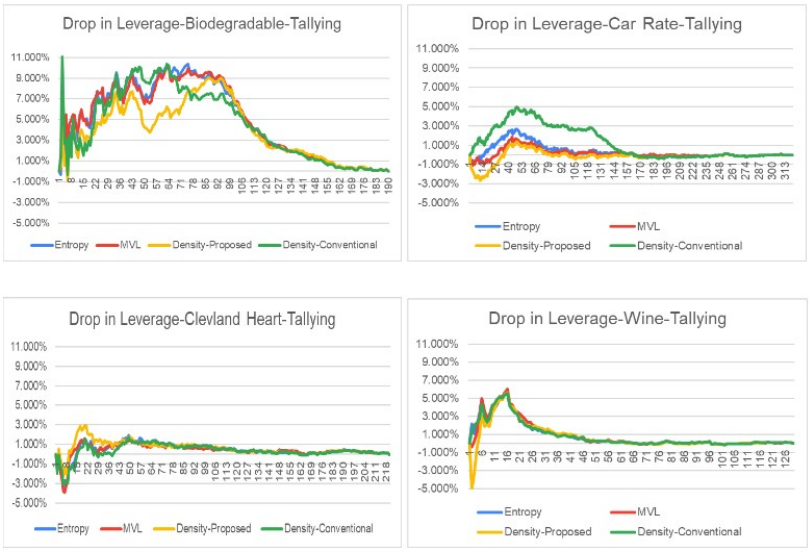

Figure 5: Drop in leverage across the learning phase of active learning algorithms for tallying