A Cross-Domain Benchmark for Active Learning

Thorben Werner, Johannes Burchert, Maximilian Stubbemann, Lars Schmidt-Thieme

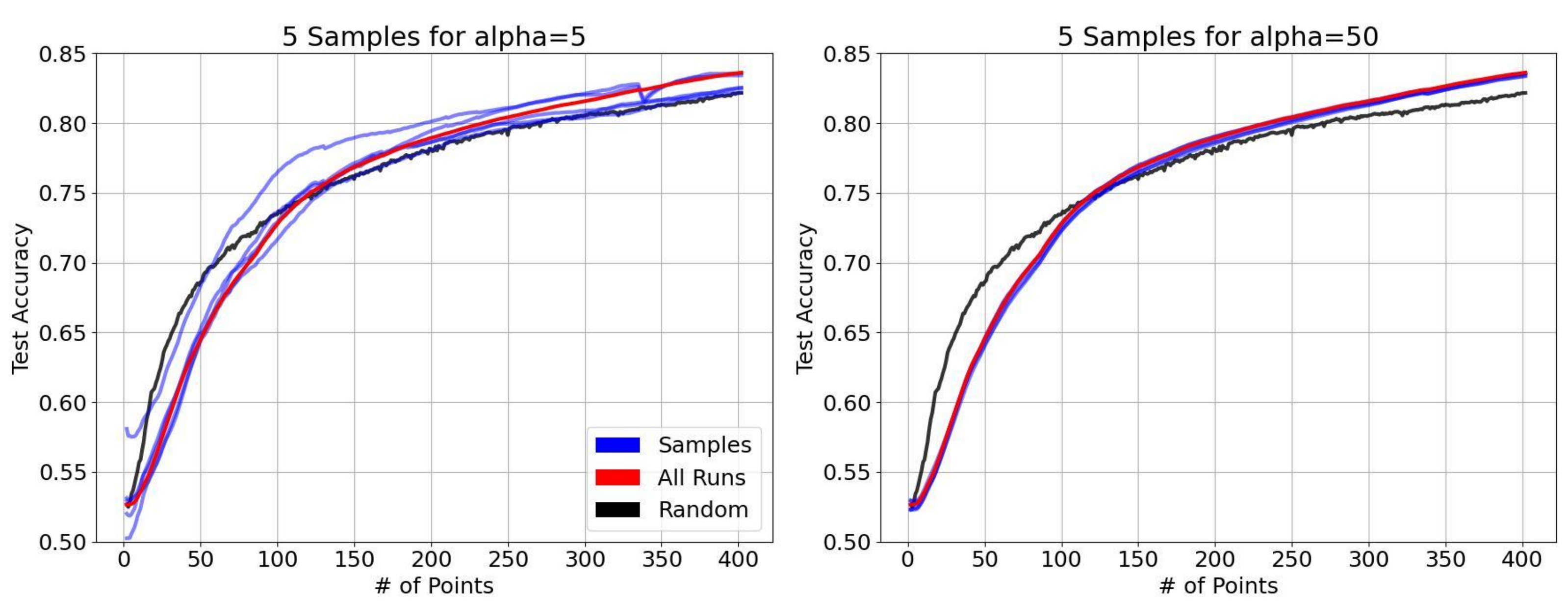
Main Findings

The image domain is an outlier

Our experiments revealed a dependency of the top-performing algorithms on the data domain. While margin sampling performs very well for Tabular, Text, and Semi-Supervised Domains, it comparatively underperforms for image data. Furthermore, for images, least confident sampling performs best, while it performs way worse for other domains. This is an important finding, as the image domain - evidently and outlier - is the most researched domain for active learning. This highlights the importance of testing AL algorithms across as many domains as possible.

3-5 repetitions are not enough to produce consistent results

We computed 100 runs of our top-performing AL method on one dataset. This allows us firstly, to obtain a very strong estimation of the "true" average performance on this particular dataset and secondly, to draw subsets from this pool of 100 runs. Setting the size of our draws to α and sampling uniformly, we can approximate a cross-validation process with α repetitions. Each of these draws (blue lines) can be interpreted as a **reported result in AL literature** where the authors employed α repetitions.

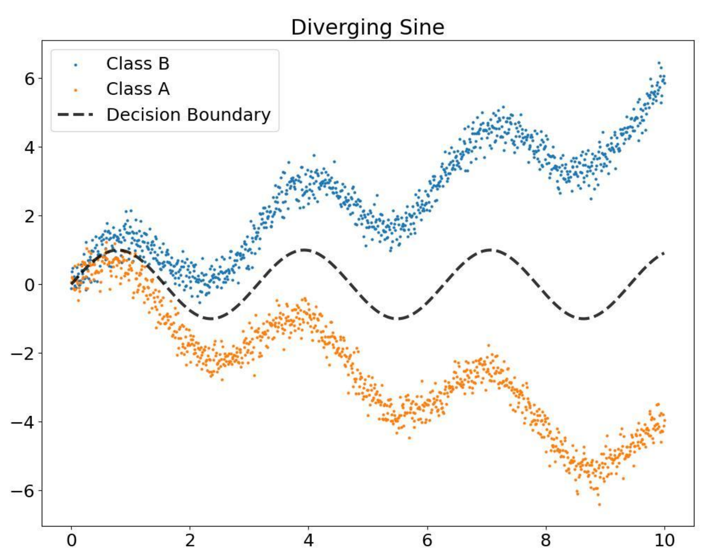


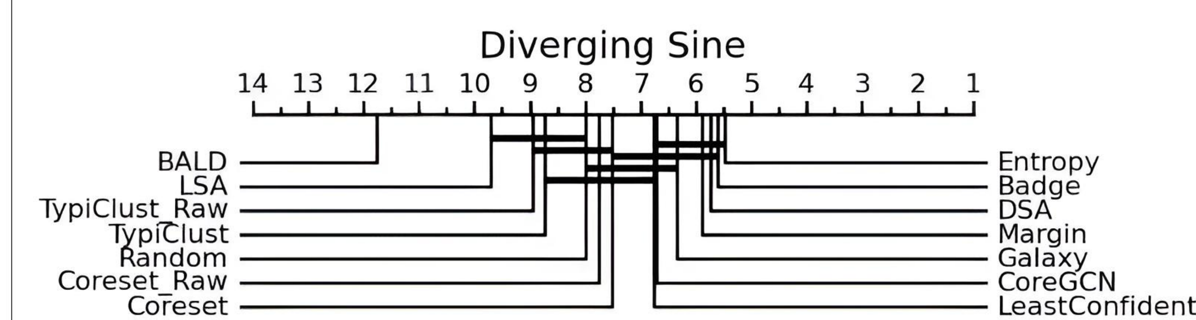
Benchmark	Sampling	# Datasets	# Algorithms	Image	Text	Tabular	Synthetic	Semi-Sup.	Oracle	Repetitions
Beck et al.	Batch	4	7	✓	-	-	-	-	-	-
Hu et al.	Batch	5	13	✓	✓	-	-	-	-	3
Zhou et al.	Batch	3	2	✓	✓	-	-	-	✓	5
Zahn et al.	Single + Batch	35	18	-	-	✓	✓	-	✓	10-100
Munjal et al.	Batch	2	8	✓	-	-	-	-	-	3
Li et al.	Batch	5	13	✓	-	-	-	✓	-	-
Rauch et al.	Batch	11	5	-	✓	-	-	-	-	5
Zhang et al.	Batch	6	7	✓	-	-	-	-	-	2-4
Bahri et al.	Batch	69	16	-	-	✓	-	-	-	2-4
Ji et al.	Batch	3	8	✓	-	-	-	-	-	-
Lueth et al.	Batch	4	5	✓	-	-	-	✓	-	3
Ours	Single + Batch	9(14)	11	✓	✓	✓	✓	✓	✓	50

Synthetic Datasets

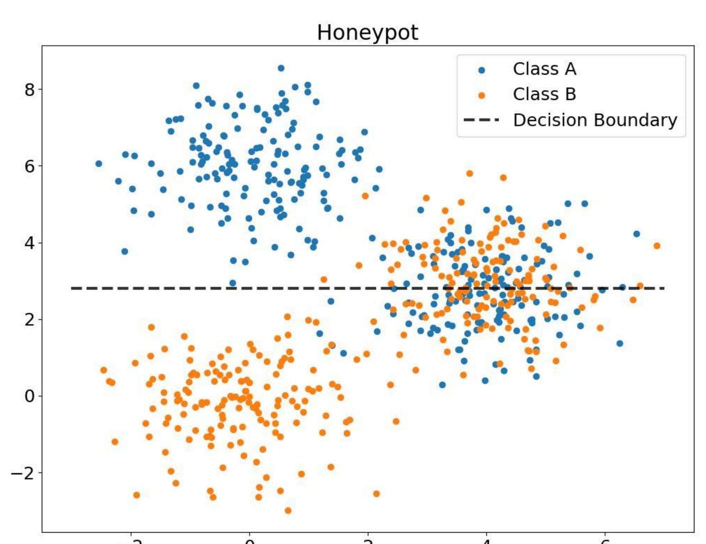
Each dataset was designed to either challenge uncertainty sampling or clustering methods

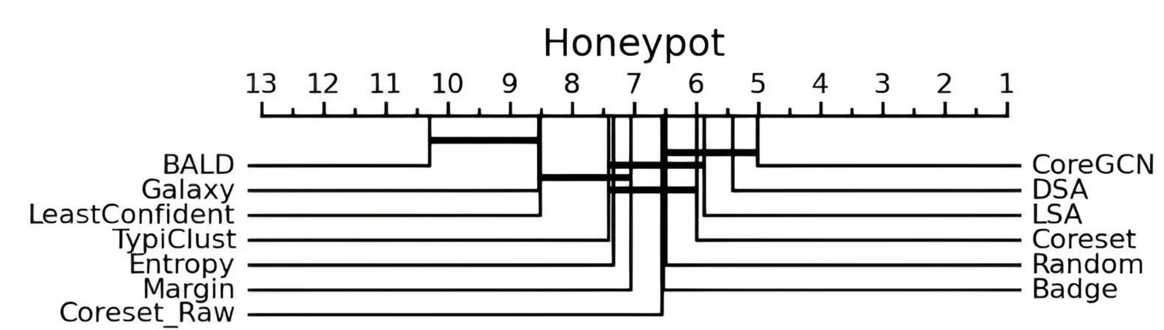
Synthetic datasets allow us to measure principled shortcomings of well-known AL algorithms. Even though these shortcomings might already be known for some algorithms, they have yet to be tested systematically.





The Diverging Sine dataset is designed to be hard to solve for clustering algorithms. This dataset needs a lot of samples on the left-hand side and progressively less towards the right. The dynamic nature of the sine functions promt clustering algorithms to sample uniformly accross X and therefore oversample the right hand side.



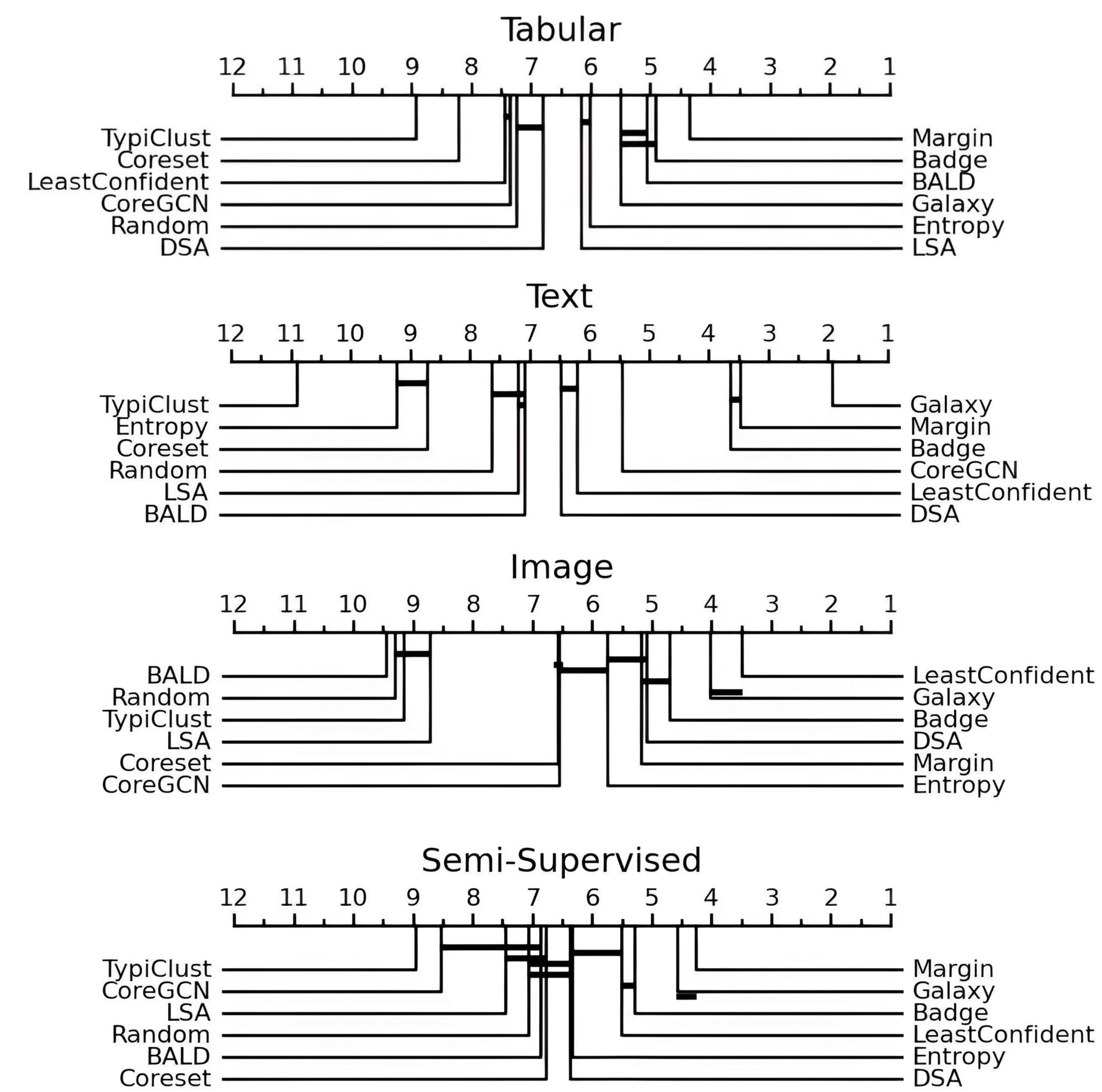


The Honeypot dataset is designed to be hard to solve for uncertainty sampling algorithms. This dataset introduces a noisy region in dataspace where the labels are random. Uncertainty sampling algorithms will heavily oversample this noisy region without improving the classifier much, while clustering algorithms will equally sample from all three regions.

Results for all Domains

Average rank of each algorithm over datasets and query sizes

Basing our evaluation on ranks allows us to average the performance of algorithms across different datasets and query sizes without risking a skew from different scales. We also employed a paired-t-test instead of the regular t-test and display the significances in Critical Difference Diagrams. The paired-t-test is made possible by specialized seeding in our framework.



Critical Difference Diagrams for each domain across all datasets and query sizes. Lower rank is better. A horizontal bar means that a group of algorithms is not significantly different, based on the paired-t-test.

References

Nathan Beck, Durga Sivasubramanian, Apurva Dani, Ganesh Ramakrishnan, and Rishabh Iyer. "Effective evaluation of deep active learning on image classification tasks." arXiv preprint arXiv:2106.15324, 2021.

Qiang Hu, Yuejun Guo, Maxime Cordy, Xiaofei Xie, Wei Ma, Mike Papadakis, and Yves Le Traon. "Towards exploring the limitations of active learning: An empirical study." In 2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE), pages 917–929. IEEE, 2021.

Yilun Zhou, Adithya Renduchintala, Xian Li, Sida Wang, Yashar Mehdad, and Asish Ghoshal. "Towards understanding the behaviors of optimal deep active learning algorithms." In International Conference on Artificial Intelligence and Statistics, pages 1486–1494. PMLR, 2021.

Xueying Zhan, Qingzhong Wang, Kuan-hao Huang, Haoyi Xiong, Dejing Dou, and Antoni B Chan. "A comparative survey of deep active learning." arXiv preprint arXiv:2203.13450, 2022.

Prateek Munjal, Nasir Hayat, Munawar Hayat, Jamshid Sourati, and Shadab Khan. "Towards robust and reproducible active learning using neural networks." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 223–232, 2022.

Bahri, Dara, et al. "Is margin all you need? An extensive empirical study of active learning on tabular data." arXiv preprint arXiv:2210.03822 (2022).

Yu Li, Muxi Chen, Yannan Liu, Daojing He, and Qiang Xu. "An empirical study on the efficacy of deep active learning for image classification." arXiv preprint arXiv:2212.03088, 2022.

Lukas Rauch, Matthias Aßenmacher, Denis Huseljic, Moritz Wirth, Bernd Bischl, and Bernhard Sick. "Activeglae: A benchmark for deep active learning with transformers." In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 55–74. Springer, 2023.

Zhang, Jifan, et al. "LabelBench: A Comprehensive Framework for Benchmarking Adaptive Label-Efficient Learning." Journal of Data-centric Machine Learning Research (2024).

Carsten Lüth, Till Bungert, Lukas Klein, and Paul Jaeger. "Navigating the pitfalls of active learning evaluation: A systematic framework for meaningful performance assessment." Advances in Neural Information Processing Systems, 36, 2024.

Yilin Ji, Daniel Kaestner, Oliver Wirth, and Christian Wressnegger. "Randomness is the root of all evil: More reliable evaluation of deep active learning." In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 3943–3952, 2023.

Code available!