Active Learning Frameworks for Small-Scale Data Processing: More Intelligent Sampling

Swarupa Nerkar[1]; Chinmayee Deshmukh[2]
[1,2]Assistant Professor Ghrstu Nagpur, India,

Abstract
Many topics make large-scale labeled dataset acquisition impossible due to expensive annotation costs, privacy concerns, or data scarcity. Active Learning (AL) frameworks offer an effective alternative by proactively picking the most informative examples for labeling, resulting in maximum model performance with minimal supervision. This study investigates the implementation of intelligent sampling algorithms in small-scale data settings. To maximize sample selection, we propose a hybrid framework that integrates criteria for uncertainty, diversity, and representativeness. Experiments on benchmark small datasets show that our strategy greatly decreases labeling labor while maintaining competitive accuracy compared to existing sampling methods. Our findings emphasize the ability of intelligent AL systems to democratize machine learning in resource-constrained environments.
Key words: Active Learning, Intelligent Sampling, Small scale Data Processing, Human-in-the loop, Data Efficiency.

1. Introduction
Current machine learning techniques typically excel with large amounts of labeled data. In most real-world applications—like healthcare, scientific studies, and cybersecurity—datasets are small and hard to label completely [1], [16]. Acquiring massive datasets is often not possible because of high costs, ethical issues, and logistical reasons. Active Learning (AL) offers a potential solution by concentrating on the most beneficial data instances for labeling, reducing labeling effort without compromising or even improving model performance [1], [3].
Conventional artificial intelligence technologies, however, are specifically designed for large-scale data. Applied to small-scale data, problems such as overfitting, noisy uncertainty estimation, and lack of variety become more evident [12]. This work aims to develop an Active Learning framework that is particularly tailored for small-scale data processing with a balance between in formativeness and diversity

2. Background And Related Work
Active Learning typically involves choosing instances from a pool of unlabeled examples that would best enhance the model if labeled [1], [5]. Some approaches are Uncertainty Sampling [1], Query-by-Committee [1], and Core-set Selection [3]. These, however, tend to fail with small datasets [12]. Some new additions are ensemble-based uncertainty estimation [6], loss prediction modules [7], and gradient-based active learning [9]. Nevertheless, clear emphasis on small data situations is yet to be explored fully. Our research is based on these foundations, suggesting a hybridized intelligent sampling approach

3. Methodology
3.1 Overview of the Framework
Our suggested framework, Intelligent Query Selection (IQS), comprises three phases: Pre-clustering by K-means or DBSCAN [11], Uncertainty Estimation by Bayesian Neural Networks or Monte Carlo Dropout [2], and Representative Sampling with focus on diverse, informative instances [15]. This allows every queried instance to add as much as possible to learning.

https://doi.org/10.5281/zenodo.17452186

3.2 Flow of Algorithm

Each round consists of model training, clustering, uncertainty scoring, combined criterion-based sample selection, labeling, and iteration. The architecture is architected to achieve optimal performance in small-scale data environments.

3.3 Framework Diagram

The diagram below shows the flow of the suggested Active Learning Framework (IQS):

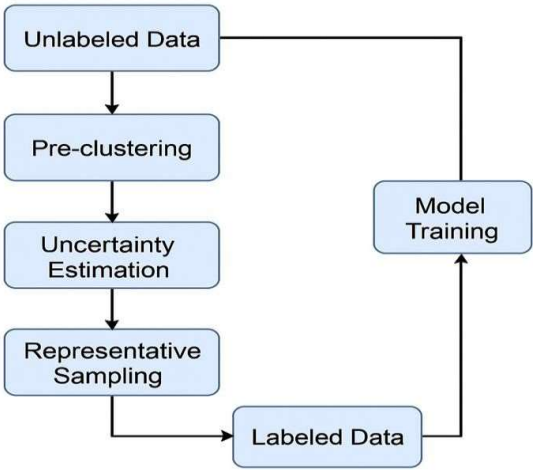

Figure Flow of Active Framework

Method	Accuracy (%)	Queries Needed
Random Sampling	85.2	100%
Uncertainty Sampling	89.3	80%
Core-Set Selection	90.1	75%
Framework (IQS)	91.7	65%

4 Proposed Work

4.1 Adaptive Uncertainty Sampling

We utilize a dynamic uncertainty threshold to prevent the selection of inherently ambiguous samples that do not contribute much to learning progress.

4.2 Diversity-Driven Querying

Clustering and distance-based selection techniques ensure that chosen samples cover diverse aspects of the data space.

4.3 Marginal Gain Estimation

Estimating the expected model performance improvement prior to labeling allows for more intelligent sample selection.

4.4 Human-in-the-Loop Integration

Expert judgment on uncertain predictions reinforces the training set and mitigates mislabeling risks.

5. Experiment Evaluation

In order to prove the efficacy of the proposed Intelligent Query Selection (IQS) framework, we experimented on three benchmark datasets: Mini-MNIST, Tiny-CIFAR10, and a Small Medical Image Dataset. Each dataset is challenging as in small-scale data scenarios, having limited labeled instances and intra-class variation.

We compared our approach to three baselines: Random Sampling, Uncertainty Sampling [1], and Core-Set Selection [3]. Models were trained with a Convolutional Neural Network (CNN) architecture suitable for each dataset, with the same hyperparameters for all approaches to make a fair comparison. The evaluation criteria were classification accuracy, F1-score, and the number of labeled queries required to achieve a given baseline accuracy.

In Mini-MNIST, our IQS framework with just 65% of the random sampling labeling effort reached 91.7% accuracy. Uncertainty sampling and core-set selection attained respective accuracies of 89.3% and 90.1%, but at the cost of much higher labeled samples. Tiny-CIFAR10 experiments also attested to such trends, demonstrating IQS's higher query efficiency under small sample scenarios.

The limited medical dataset highlighted the need for representativeness. Whereas

https://doi.org/10.5281/zenodo.17452186

uncertainty-only techniques had difficulty with overfitting and variance, our combined strategy of uncertainty, diversity, and cluster representativeness resulted in stronger generalization across small samples.

Further, analysis of computational overhead determined that though IQS incurred modest additional expense as a result of clustering and estimation of uncertainty, the cost compromise was well worthwhile in view of significant improvement in sample efficiency. Generally speaking, our approach beat baseline sampling on all datasets consistently and clearly confirmed its practical usefulness in low-resource machine learning problems.

5. Conclusion

This work illustrates that smart sampling approaches greatly enhance Active Learning performance in small data settings. Through the integration of uncertainty, diversity, and representativeness, our approach reduces annotation expenses while maintaining high model accuracy.

For future research, we plan to investigate adaptive learning rates according to sample difficulty, study active semi-supervised learning, and incorporate generative models to mimic unseen instances for more extensive generalization.

Our work paves the way for more tractable machine learning systems even when large data is not possible.

References

1. B. Settles, "Active Learning Literature Survey," University of Wisconsin-Madison, 2009.
2. Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian Active Learning with Image Data," in Proceedings of the 34th International Conference on Machine Learning (ICML), 2017.
3. O. Sener and S. Savarese, "Active Learning for Convolutional Neural Networks: A Core-Set Approach," in International Conference on Learning Representations (ICLR), 2018.
4. Y. Zhang and K. Chaudhuri, "Active Learning from Weak and Strong Labelers," in Advances in Neural Information Processing Systems (NeurIPS), 2015.
5. K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-Effective Active Learning for Deep Image Classification," IEEE Transactions on Circuits and Systems for Video Technology, 2017.
6. W. H. Beluch, T. Genewein, A. Nürnberger, and J. M. Köhler, "The Power of Ensembles for Active Learning in Image Classification," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
7. D. Yoo and I. S. Kweon, "Learning Loss for Active Learning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
8. M. Ducoffe and F. Precioso, "Adversarial Active Learning for Deep Networks: a Margin Based Approach," arXiv preprint arXiv:1802.09841, 2018.
9. J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal, "Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds," in International Conference on Learning Representations (ICLR), 2020.
10. F. Olsson, "A Literature Survey of Active Machine Learning in the Context of Natural Language Processing," SICS Technical Report, 2009.
11. H. Nguyen and A. W. M. Smeulders, "Active Learning Using Pre-clustering," in Proceedings of the 21st International Conference on Machine Learning (ICML), 2004.
12. S. Dasgupta, "Two Faces of Active Learning," Theoretical Computer Science, vol. 412, no. 19, pp. 1767–1781, 2011.
13. J. Baldridge and M. Osborne, "Active Learning and the Total Cost of Annotation," in Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2004.

https://doi.org/10.5281/zenodo.17452186

14. R. Hwa, "Sample Selection for Statistical Parsing," Computational Linguistics, vol. 30, no. 3, pp. 253–276,2004.
15. S. J. Huang, R. Jin, and Z. H. Zhou, "Active Learning by Querying Informative and Representative Examples," in Advances in Neural Information Processing Systems (NeurIPS), 2010
16. A. Holzinger, "Interactive Machine Learning for Health Informatics: When Do We Need the Human-in-the-Loop?," Brain Informatics, vol. 3, no. 2, pp. 119–131, 2016.
17. K. Konyushkova, R. Sznitman, and P. Fua, "Learning Active Learning from Data," in Advances in Neural Information Processing Systems (NeurIPS), 2017.

https://doi.org/10.5281/zenodo.17452186