Tutorial: Interactive Adaptive Learning

Marek Herde[1,†], Minh Tuan Pham[1,†], Alaa Tharwat[2,†] and Bernhard Sick[1]

[1]University of Kassel, Germany
[2]Hochschule Bielefeld, Germany

Abstract

We summarize the contents of the tutorial we present as a part of the 8th Interactive Adaptive Learning workshop. This workshop is co-located with the ECML-PKDD conference, where it takes place on September 9th, 2024 in Vilnius, Lithuania.

Keywords

active learning, uncertainty quantification, exploration-exploitation tradeoff, scikit-activeml, noisy class labels

1. Introduction

Interactive adaptive learning refers to methods that help improve the entire lifecycle of machine learning models. This includes how the models interact with human experts or other systems and how they adapt to different types of emerging data rather than just training them on a fixed dataset. This allows the models to improve and adapt over time, which is critical for many real-world applications. Active learning is the most prominent field of interactive adaptive learning [1, 2, 3]. Therefore, we explore different aspects of active learning in this tutorial and then discuss recent advances in the field in a workshop session.

This tutorial is divided into three main parts, which are described in detail in the following sections. Their titles and presenters are as follows:

1. Introduction to Uncertainty-Based Active Learning (A. Tharwat),
2. Hands-on Pool-based Active Learning via `scikit-activeml` (M. Herde),
3. Towards Pool-based Active Learning with Error-prone Annotators (M. Herde).

2. Part I – Introduction to Uncertainty-Based Active Learning

The motivation behind active learning stems from the common scenario where large amounts of unlabeled data are readily available and free, but the process of obtaining labeled data is time-consuming and expensive. In many real-world applications, such as medical image analysis, document classification, or sensor network monitoring, the cost of manual labeling or expert annotation can be prohibitively high. Traditional passive learning approaches, where the model is trained on a fixed dataset, may not be the most efficient use of limited labeling resources [4]. Active learning offers a solution to this challenge by enabling the model to actively select the most informative and/or representative samples from the unlabeled pool for labeling. By focusing the annotation effort on the most valuable data points, active learning can achieve better model performance with fewer labeled samples compared to passive learning [3]. There are two main strategies for querying points in active learning as follows:

- **Exploration-focused query strategies**: This category aims to identify representative samples that can help the model explore the underlying data distribution more effectively. By querying for diverse and representative data points, the model can gain a better understanding of the

IAL@ECML-PKDD'24: 8th Intl. Worksh. & Tutorial on Interactive Adaptive Learning, Sep. 9th, 2024, Vilnius, Lithuania
[†]These authors contributed equally.
marek.herde@uni-kassel.de (M. Herde); tuan.pham@uni-kassel.de (M. T. Pham); alaa.othman@hsbi.de (A. Tharwat); bsick@uni-kassel.de (B. Sick)

true structure of the problem domain, even with a limited labeled dataset [5]. Some common exploration-focused query strategies include (i) **Diversity-based sampling**: selecting dissimilar samples to explore different regions, (ii) **Density-based sampling**: prioritizing samples in dense areas, as they are more representative, (iii) **Clustering-based sampling**: identifying samples near cluster centroids to ensure coverage of data subgroups [3].

- **Informativeness (or exploitation)-focused strategies**: The type in which active learning employs informativeness-focused query strategies with the goal of identifying the most informative samples, i.e., the data points that, if labeled, would provide the maximum information gain to the model [6]. Some common exploitation-focused approaches include (i) **Margin-based sampling**: prioritizing samples near the decision boundary [1], (ii) **Entropy-based sampling**: selecting the most uncertain samples [7], and (iii) **Variance-based sampling**: prioritizing samples with high prediction variance [7].

However, by exploiting the model's own uncertainty about the unlabeled data, active learning can more effectively identify the samples that, if labeled, would provide the maximum information gain to improve the model's performance. The key idea behind uncertainty-based active learning is that the model's uncertainty serves as a proxy for the informativeness of a data point. Samples with higher uncertainty are more likely to be informative because they represent areas of the input space where the model's predictions are less confident or reliable [8, 9]. Common uncertainty-based approaches include

- **Uncertainty sampling**: Selecting the most uncertain samples, as they are likely to be the most informative [7].
- **Expected information gain**: Choosing samples with the highest expected information gain [7].
- **Bayesian optimization**: Using a Bayesian model to quantify uncertainty and guide sample selection [10].

Recent advances in uncertainty quantification have further extended the capabilities of active learning. The distinction between (i) epistemic uncertainty, which occurs due to lack of knowledge, and samples with high epistemic uncertainty are often the most informative for improving model understanding, and (ii) aleatory uncertainty, which captures the inherent noise or randomness in the data, and samples with high aleatory uncertainty may be less useful for model training [11, 12, 13].

By distinguishing between epistemic and aleatory uncertainty, active learning can more effectively identify the most informative samples to improve model performance with less labeled data. The integration of advanced uncertainty quantification with active learning strategies creates a powerful framework for efficient and effective model training, even with large amounts of unlabeled data [9].

3. Part II – Hands-on Pool-based Active Learning via `scikit-activeml`

Active learning is a versatile approach to reducing the labeling cost. Assumptions regarding data availability and training of the classifiers can vary greatly depending on the use case. Active learning libraries often abstract parts of active learning experiments, such as the whole experiment (`AliPy` [14] and `Baal` [15]), the classifier training (`modAL` [16] and `small-text` [17]), and data management (`libact` [18]). These abstractions help simplify active learning experiments where the use case matches the library's scope. However, it requires considerable work if the assumptions differ. The `scikit-activeml` [19] library has been conceptualized with this in mind, with its modular design inspired by and built on top of `scikit-learn` [20], a flexible general-purpose machine learning library.

The goal of `scikit-activeml` is to bridge active learning research and its application in real-world use cases. The library is flexible enough for researchers to allow for many assumptions and promotes reproducibility. For practitioners, it provides extensive documentation, many tutorials for different use cases, and examples for each query strategy with animations visualizing the strategies' behavior that can be used as a starting ground.

Algorithm 1 A basic active learning cycle example using `scikit-activeml` (v0.5.1) [19]

```
1  import numpy as np
2  from sklearn.gaussian_process import GaussianProcessClassifier
3  from sklearn.datasets import make_blobs
4  from skactiveml.pool import UncertaintySampling
5  from skactiveml.utils import unlabeled_indices, MISSING_LABEL
6  from skactiveml.classifier import SklearnClassifier
7
8  # Generate data set.
9  X, y_true = make_blobs(n_samples=200, centers=4, random_state=0)
10 y = np.full(shape=y_true.shape, fill_value=MISSING_LABEL)
11
12 # Use the first 10 samples as initial training data.
13 y[:10] = y_true[:10]
14
15 # Create classifier and query strategy.
16 clf = SklearnClassifier(
17     GaussianProcessClassifier(random_state=0),
18     classes=np.unique(y_true),
19     random_state=0
20 )
21 qs = UncertaintySampling(method='entropy', random_state=0)
22
23 # Execute active learning cycle.
24 n_cycles = 30
25 for c in range(n_cycles):
26     query_idx = qs.query(X=X, y=y, clf=clf)
27     y[query_idx] = y_true[query_idx]
28
29 # Fit final classifier.
30 clf.fit(X, y)
```

`scikit-activeml` provides query strategies for pool-based active learning in classification and regression with single or multiple annotators with varying batch sizes. Stream-based active learning [21] for classification is also supported for single annotator scenarios. In this tutorial, we focus on pool-based active learning. Algorithm 1 shows a small pool-based active learning script using `scikit-activeml`. Labeled and unlabeled data are stored together in `X` (samples) and `y` (labels), where unlabeled data is marked with a user-specified `MISSING_LABEL` constant (cf. lines 9–13). The classifier and query strategy are initialized independently and support using random seeds to ensure reproducibility (cf. lines 16–21). The for-loop shows the active learning cycle, where sample indices are queried (cf. line 26), and their corresponding missing label is replaced with the ground truth (cf. line 27). Figure 1 shows the fitted classifier, labeled, and unlabeled data after 10 and 30 cycles. Additionally, areas where it is beneficial to query more labels, according to uncertainty sampling, are highlighted in dark green.

Starting from such a basic learning cycle with uncertainty sampling as the employed query strategy, this tutorial's part outlines other popular and state-of-the-art query strategies for pool-based active learning, e.g., *core set* [22], *batch active learning by diverse gradient embeddings* (BADGE) [23], *typical clustering* (TypiClust) [24], *probability coverage* (ProbCover) [25], *clustering uncertainty-weighted embeddings* (CLUE) [26], and *contrastive active learning* (CAL) [27]. Specifically, we analyze these query strategies regarding informativeness, representativeness, and batch diversity as central concepts in pool-based active learning (cf. Section 2). Further, we introduce the differentiation between low- and high-budget active learning scenarios. Depending on the scenario, the importance of the aforementioned concepts changes. Throughout this tutorial's part, illustrations of synthetic two-dimensional datasets (cf. Fig. 1) provide a more intuitive understanding of the query strategies' main ideas and sample selection behaviors. Beyond such toy examples, we also present an empirical evaluation study as a potential application for `scikt-activeml`, where we compare query strategies' performances across tabular,

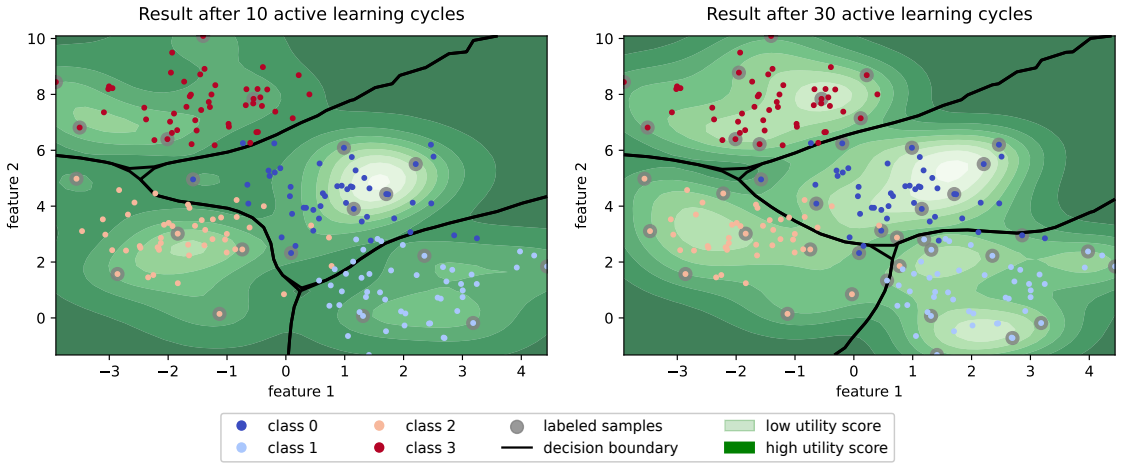

Figure 1: Visualization of the results from basic active learning cycle in Algorithm 1.

image, and text data. In doing so, we leverage feature representations (embeddings) learned by the pre-trained model *self-distillation with no labels* (DINOv2) [28] for the image and *bidirectional encoder representations from transformers* (BERT) [29] for the text data. This tutorial's part concludes with a hands-on session, where participants can access a Jupyter notebook to apply their newly acquired knowledge by implementing their own active learning experiment via `scikit-activeml`.

4. Part III – Towards Pool-based Active Learning with Error-prone Annotators

In pool-based active learning, a common assumption is that the queried class labels originate from a single omniscient annotator [30]. However, many annotation campaigns involve querying class labels from multiple humans, e.g., crowdworkers, who are prone to error for various reasons, e.g., lack of expertise, tiredness, or missing motivation [31]. As a result, the queried class labels are subject to noise. Training with such noisy class labels can strongly deteriorate the classifier's performance. With a focus on neural networks, numerous techniques have been proposed to improve the robustness against noisy class labels [32]. A common approach is the joint training of the classifier and an annotator performance model, which corrects the noisy class labels by modeling each annotator's individual performance [33]. Depending on the assumptions about the annotators' noise patterns, confusion matrices are estimated per annotator [34] or even for each sample-annotator pair [35], for example. Such techniques are typically employed to train a neural network after completing an annotation campaign. Yet, the annotators' performance estimates could be used to guide the annotator selection during an ongoing annotation campaign. In conjunction with intelligent sample selection, we refer to this scenario as pool-based active learning with multiple error-prone annotators. Corresponding query strategies [36, 37] must also balance the added exploration-exploitation trade-off when assigning annotators to provide class labels for given instances. The goal of this tutorial's part is to give a basic understanding of such challenges and outline potential baselines, which leverage common pool-based query strategies for the sample selection and performance estimates for the annotator selection.

Acknowledgments

The work of A. Tharwat was conducted within the framework of the project "SAIL: SustAInable Lifecycle of Intelligent SocioTechnical Systems" (grant no. NW21-059B). SAIL is receiving funding from the program "Netzwerke 2021", an initiative of the Ministry of Culture and Science of the State of North Rhine-Westphalia. The work of M. T. Pham was conducted within the framework of the

project “Künstliche Intelligenz zur Fremdkörperdetektion in befüllten Getränkeflaschen (KI4FKD)” (493 22_0022_2B). This project is receiving funding from the program “Distr@l”, an initiative of the Ministry for Digitalization and Innovation of the State of Hesse. The sole responsibility for the content of this publication lies with the authors.

References

[1] B. Settles, Active learning literature survey (2009).
[2] A. Tharwat, W. Schenck, A novel low-query-budget active learner with pseudo-labels for imbalanced data, Mathematics 10 (2022) 1068.
[3] A. Tharwat, W. Schenck, A survey on active learning: State-of-the-art, practical challenges and research directions, Mathematics 11 (2023) 820.
[4] D. Cohn, L. Atlas, R. Ladner, Improving generalization with active learning, Machine Learning 15 (1994) 201–221.
[5] A. Tharwat, W. Schenck, Balancing Exploration and Exploitation: A novel active learner for imbalanced data, Knowledge-Based Systems 210 (2020) 106500.
[6] A. Tharwat, W. Schenck, Using methods from dimensionality reduction for active learning with low query budget, IEEE Transactions on Knowledge and Data Engineering (2024).
[7] V.-L. Nguyen, M. H. Shaker, E. Hüllermeier, How to measure uncertainty in uncertainty sampling for active learning, Machine Learning 111 (2022) 89–122.
[8] M. H. Shaker, E. Hüllermeier, Aleatoric and epistemic uncertainty with random forests, in: Advances in Intelligent Data Analysis: International Symposium on Intelligent Data Analysis, Springer, 2020, pp. 444–456.
[9] A. Tharwat, W. Schenck, Active Learning for Handling Missing Data, IEEE Transactions on Neural Networks and Learning Systems (2024).
[10] D. Khatamsaz, B. Vela, P. Singh, D. D. Johnson, D. Allaire, R. Arróyave, Bayesian optimization with active learning of design constraints using an entropy-based approach, npj Computational Materials 9 (2023) 49.
[11] L. Wimmer, Y. Sale, P. Hofman, B. Bischl, E. Hüllermeier, Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures?, in: Uncertainty in Artificial Intelligence, PMLR, 2023, pp. 2282–2292.
[12] E. Hüllermeier, W. Waegeman, Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods, Machine Learning 110 (2021) 457–506.
[13] R. Senge, S. Bösner, K. Dembczyński, J. Haasenritter, O. Hirsch, N. Donner-Banzhoff, E. Hüllermeier, Reliable classification: Learning classifiers that distinguish aleatoric and epistemic uncertainty, Information Sciences 255 (2014) 16–29.
[14] Y.-P. Tang, G.-X. Li, S.-J. Huang, ALiPy: Active Learning in Python, arXiv preprint arXiv:1901.03802 (2019).
[15] P. Atighehchian, F. Branchaud-Charron, A. Lacoste, Bayesian active learning for production, a systematic study and a reusable library, arXiv preprint arXiv:2006.09916 (2020).
[16] T. Danka, P. Horvath, modAL: A modular active learning framework for Python, arXiv preprint arXiv:1805.00979 (2018).
[17] C. Schröder, L. Müller, A. Niekler, M. Potthast, Small-text: Active learning for text classification in python, in: Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, 2023, pp. 84–95.
[18] Y.-Y. Yang, S.-C. Lee, Y.-A. Chung, T.-E. Wu, S.-A. Chen, H.-T. Lin, libact: Pool-based Active Learning in Python, arXiv preprint arXiv:1710.00379 (2017).
[19] D. Kottke, M. Herde, T. P. Minh, A. Benz, P. Mergard, A. Roghman, C. Sandrock, B. Sick, scikit-activeml: A library and toolbox for active learning algorithms, Preprints (2021).
[20] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, G. Varoquaux, API design for

machine learning software: experiences from the scikit-learn project, in: ECML PKDD Workshop: Languages for Data Mining and Machine Learning, 2013, pp. 108–122.

[21] D. Cacciarelli, M. Kulahci, Active learning for data streams: a survey, Machine Learning 113 (2024) 185–239.

[22] O. Sener, S. Savarese, Active learning for convolutional neural networks: A core-set approach, in: International Conference on Learning Representations, 2018.

[23] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, A. Agarwal, Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds, in: International Conference on Learning Representations, 2020.

[24] G. Hacohen, A. Dekel, D. Weinshall, Active Learning on a Budget: Opposite Strategies Suit High and Low Budgets, in: International Conference on Machine Learning, 2022, pp. 8175–8195.

[25] O. Yehuda, A. Dekel, G. Hacohen, D. Weinshall, Active Learning through a Covering Lens, in: Advances in Neural Information Processing Systems, 2022, pp. 22354–22367.

[26] V. Prabhu, A. Chandrasekaran, K. Saenko, J. Hoffman, Active Domain Adaptation via Clustering Uncertainty-weighted Embeddings, in: IEEE/CVF International Conference on Computer Vision, 2021, pp. 8505–8514.

[27] K. Margatina, G. Vernikos, L. Barrault, N. Aletras, Active Learning by Acquiring Contrastive Examples, in: Conference on Empirical Methods in Natural Language Processing, 2021, pp. 650–663.

[28] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, P. Bojanowski, DINOv2: Learning robust visual features without supervision, Transactions on Machine Learning Research (2024).

[29] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019, pp. 4171–4186.

[30] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, Á. Fernández-Leal, Human-in-the-loop machine learning: a state of the art, Artificial Intelligence Review 56 (2023) 3005–3054.

[31] M. Herde, D. Huseljic, B. Sick, A. Calma, A Survey on Cost Types, Interaction Schemes, and Annotator Performance Models in Selection Algorithms for Active Learning in Classification, IEEE Access 9 (2021) 166970–166989.

[32] H. Song, M. Kim, D. Park, Y. Shin, J.-G. Lee, Learning From Noisy Labels With Deep Neural Networks: A Survey, IEEE Transactions on Neural Networks and Learning Systems (2022).

[33] M. Herde, D. Huseljic, B. Sick, Multi-annotator Deep Learning: A Probabilistic Framework for Classification, Transactions on Machine Learning Research (2023).

[34] S. Ibrahim, T. Nguyen, X. Fu, Deep Learning From Crowdsourced Labels: Coupled Cross-Entropy Minimization, Identifiability, and Regularization, in: International Conference on Learning Representations, 2023.

[35] M. Herde, L. Lührs, D. Huseljic, B. Sick, Annot-Mix: Learning with Noisy Class Labels from Multiple Annotators via a Mixup Extension, arXiv preprint arXiv:2405.03386 (2024).

[36] S. Chakraborty, Asking the Right Questions to the Right Users: Active Learning with Imperfect Oracles, in: AAAI Conference on Artificial Intelligence, 2020, pp. 3365–3372.

[37] M. Herde, D. Kottke, D. Huseljic, B. Sick, Multi-annotator Probabilistic Active Learning, in: International Conference on Pattern Recognition, IEEE, 2021, pp. 10281–10288.