Challenges of Reliable, Realistic and Comparable Active Learning Evaluation

Daniel Kottke[1], Adrian Calma[1], Denis Huseljic[1],
Georg Krempl[2], and Bernhard Sick[1]

[1)] University of Kassel
Wilhelmshöher Allee 73, 34112 Kassel, Germany
{daniel.kottke,adrian.calma,bsick}@uni-kassel.de

[2)] Otto-von-Guericke University Magdeburg
Universitätsplatz 2, 39106 Magdeburg, Germany
georg.krempl@ovgu.de

Abstract. Active learning has the potential to save costs by intelligent use of resources in form of some expert's knowledge. Nevertheless, these methods are still not established in real-world applications as they can not be evaluated properly in the specific scenario because evaluation data is missing. In this article, we provide a summary of different evaluation methodologies by discussing them in terms of being reproducible, comparable, and realistic. A pilot study which compares the results of different exhaustive evaluations suggests a lack in repetitions in many articles. Furthermore, we aim to start a discussion on a gold standard evaluation setup for active learning that ensures comparability without reimplementing algorithms.

Keywords: Evaluation, Active Learning, Classification, Semi-supervised Learning, Data Mining

1 Introduction

The field of machine *active learning* (AL) investigates how a learning algorithm can learn to solve problems (e.g., classification or regression problems) more effectively by exploiting interactions with humans (e.g., experts in a specific application field) or simulation systems which are abstractly modeled as an *oracle* [1] (Fig. 1). In many application domains, it is unproblematic to collect unlabeled data, but gathering labels may be complicated, time-consuming, or costly [18]. Furthermore, AL is based on the assumption that by allowing the *learner* to be curios (i.e., it is allowed to choose the data from which it learns), it may learn faster [39].

Pool-based AL [29] usually starts with an initially empty or very sparsely labeled set of samples, a large pool of unlabeled samples (candidates), and iteratively queries for new labels from instances of the candidate pool by "asking the right questions". For example, in every learning cycle the oracle is asked to provide labels for the most "informative" samples based on a *selection strategy*.

Thereby, it aims to improve the performance of the learning model as fast as possible. After the labels are added, the knowledge model is updated.

In this article, we focus on three critical aspects of AL evaluation which are underrepresented in current AL research:

- **Reliable Evaluation:** Reliable evaluation results require a robust and reproducible evaluation methodology. Hence, the methodology should be described in detail and should be robust to varying seeds or shuffled data.
- **Realistic Evaluation:** Evaluating an AL algorithm in a lab setting (the lack of labels is just simulated) is not realistic. Often, implications for the real world do not hold. Hence, AL methods are not very common in industrial applications. We will discuss the challenges of a real-world application.
- **Comparable Evaluation:** Current evaluation methodologies vary a lot regarding its evaluation type, performance measure, number of repetitions, etc. Ideally, presented results are directly comparable with others. Hence, this article aims to initiate a discussion for a standardized AL evaluation gold standard.

The article starts with a general overview of components taking part in an AL cycle (Sec. 2). Next, we discuss aspects of reliable evaluation (Sec. 3) and compare two methodologies in a pilot study (Sec. 4). In Sec. 5, we present unrealistic assumptions for real-world applications. Finally, we conclude the work and propose an outlook on how comparable evaluation could be made possible.

2 Active Learning in Classification Tasks

The learning cycle of AL (see Fig. 1) consists of three main components: In pool-based AL for classification tasks, we have a selection strategy, an oracle, and a classifier. The selection strategy selects the instances from the candidate pool to be labeled by the oracle such that the classifier can learn a well-suited model. This procedure repeats until a stopping criterion is reached. In AL evaluation, we normally investigate the performance of the selection strategy. Using an omniscient oracle and a pre-trained classifier, we can assure that performance

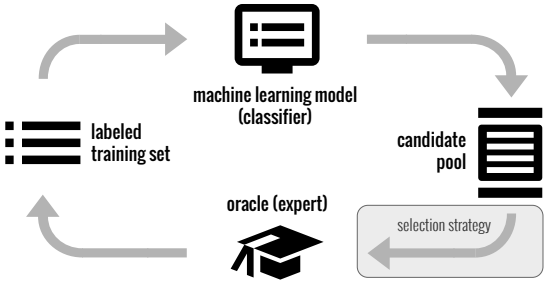

Fig. 1. Pool-based active learning cycle [39]

differences are solely induced by the selection of training instances from the candidate pool. Changing the classifier (or the parameters of the classifier) within different AL systems might lead to falsified results because of the high interdependence between the three components.

Comparing multiple classifiers in combination with AL, the selection strategy should be fixed. Comparing both, classifiers and selection strategies, one should run every combination. Unfortunately, some selection strategies solely work with specific classifiers or classifier types. Hence, it is not possible to compare these selection strategies with their individual classifiers as performance differences could be explained by the qualities of the classifiers and not the selection strategy. To face this problem, we could learn *multiple* classifiers on the selected samples. According to [42], this is subsumed under the term *label reusability*. The authors propose to use the specific classifier for the active selection (selector) and train additional classifiers for prediction (consumer). Although the authors of [42] show that the suitability of selector-consumer pairings cannot be estimated independently of the AL problem, we propose to run each selector also as a consumer for evaluation.

3 Aspects of Reliable Evaluation

Reliable evaluation is robust and reproducible. Robustness in evaluation means that changing seeds or the order of data points does not effect the results. In this section, we will point out different aspects and discuss what is done in literature.

3.1 Repetitions and Hold-Out Evaluation

In AL, we are facing classification tasks with very few training instances. When classifiers try to generalize from only a few training samples, their performance might be very sensitive to small changes. Also, the performance probably varies a lot depending on the concrete choice of instances to be labeled. Hence, lots of repetitions are needed to get a reliable trend of the performance. In Fig. 2, we clarify the nomenclature of different sets that might take part in AL.

In recent active learning articles, the number of repetitions varies between one single training-evaluation set [49] to 100 different partitionings [26]. Therefore, some authors use a k-fold cross validation [2, 5, 31] with solely one execution [31, 38] or multiple ones [2, 5]. Executing a k-fold cross validation multiple times

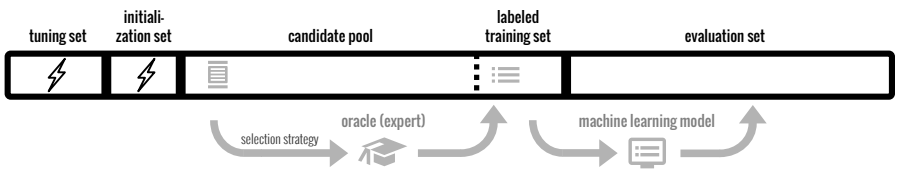

Fig. 2. Different sets used in literature for active learning.

requires different seeds among the repetitions. Others [8, 21, 30, 46] use a simple split with a fixed percentage (varying between 50% and 67%) for the candidate pool and the rest, respectively, for the evaluation set. To get rid of random effects, this is repeated multiple times.

In Sec. 4, we present a pre-study that shows the drawbacks of a single k-fold cross validation and shows the importance of multiple repetitions.

3.2 Performance Measures

Active Learning is a dynamic process which improves its model by successively adding labels to instances from the candidate pool. The aim of AL algorithms is to achieve a high performance which improves as fast as possible. Hence, we have two objectives [27, 39]:

1. achieve a high performance level (learn a good classifier) and
2. learn as fast as possible (save cost induced by annotations).

Applying Common Performance Measures to AL:

Depending on the learning problem, several performance measures [36] have been used. Usually, accuracy or error [2, 6] are used for problems with balanced misclassification cost and class priors. For unbalanced data, measures like cost, F1-Score, G-mean, Area under the Receiver Operating Characteristic-Curve (AUROC) [17, 20] (see [21, 22, 30, 48]) or H-measure [19] are more sophisticated. Usually, these performance measures are then plotted over time (resp. the number of acquired labels), which is then called learning curve (e.g., see Fig. 3).

As mentioned in the previous subsection, the results from multiple executions should be included in the evaluation by plotting standard deviations or ideally quartiles. An evaluation of means could also include the mean standard error or mean quartiles which can be determined using bootstrapping [15]. Note that quartiles are more exact as the distribution of performances given the number of acquired labels is unlikely normally distributed because these random variables are bounded (most of the time between 0 and 1).

The comparison of learning curves remains difficult as it is unclear how to combine the two objectives from above. The easiest option is to present the result for different points in time (e.g., early stage, mid stage, saturated stage) [26, 37]. Having fixed these time points, one can use comparison methods like in usual classification tasks. Note that most often, these time points and the total number of label acquisitions (when to stop learning) are chosen by the authors which could bias the results. We recommend not to stop learning before most of the AL algorithms have converged, and if possible, to also include the performance of a classifier learned on all instances as a baseline.

In reliable evaluation, statistical testing plays a essential role. Nevertheless, one should be reminded that statistical test only show if the results may also be explained by random artifacts [33], and do not show the real superiority of one's method. Nuzzo [33] claims that results should not only be reported by their statistical significance but also their effect size. Typically, statistical tests

(like the t-test or the Wilcoxon signed rank test [47]) assume to have i.i.d. random variables. Hence, the compared performance values should be drawn from the different training-evaluation combinations and not from different time points because these performance values are highly correlated and therefore *not* independent. One also could argue that even the performances across the repetitions are not independent because training and/or evaluation sets might overlap. Many use a t-test for comparing the tendencies of the mean between two algorithms [8, 21]. Due to the assumption of the mean being normally distributed, it might be better to use a parameter-free test like the Wilcoxon signed rank test [8, 22, 26, 41]. To test if an algorithm is significantly better across datasets, the Wilcoxon signed rank test might also be a good choice. An alternative to statistical testing is to present the number of won/lost trials using a simple pairwise comparison between the performances of two algorithms [26].

Active Learning Specific Performance Measures:

There also exist approaches to summarize the shape of the performance curve: The easiest approach sums up all the performance values for each time point. Often, this is called area under the learning curve [38] (also denoted as AUC[1]). This measure is proportional to the mean and hence dependent on the length of the AL process (i.e., the number of acquisitions which is often chosen manually).

More convenient is the deficiency score proposed by Yanik et al. [50]. This is determined by calculating the area between the maximal performance line and the actual learning curve which they call α for algorithm A and β for algorithm B. The deficiency of A with respect to B is then calculated using the following equation:

$$\text{deficiency}(A, B) = \frac{\alpha}{\alpha + \beta} \tag{1}$$

Another measure to calculate how fast the AL algorithm learns (2nd objective) is the Data Utilization Rate (DUR) by Reitmaier et al. [38]. They first compute the target accuracy defined as the mean (considering the performances between 80% and 100% of the total number of acquired labels) from the random strategy. The DUR is then the minimum number of samples needed by each strategy to reach this target accuracy divided by the number of samples needed by random.

3.3 Initialization of Active Learning

Some papers propose to initialize their AL cycle with some labels to be compatible to state-of-the-art implementations or as an essential part of their algorithm. The number of initialization labels varies between no label at all and 10% [30]. This choice is highly dependent on the dataset and the proposed algorithm. Unfortunately, it is often not described, how the specific values have been determined (or tuned), although this is essential for the method to succeed or fail.

[1] We do not recommend the abbrev. AUC because it can be mixed up with AUROC

The number of initial labels is relatively small when initialization is done due to compatibility issues [7, 13, 25, 37]. In some SVM implementations, the classifiers need one instance per class to predict labels. Hence, some authors added a fixed number of instances per class [43, 49, 50, 37] although this is not possible in real applications as the class labels are unknown in advance. This is even more relevant in datasets with unequal class priors as finding an instance of the minority class is especially difficult [16].

In [30, 48], the initialization step is used to have a representative sample for the dataset to find a broad decision boundary. Later, an uncertainty based method is used to refine the boundary and improve the performance. In this case, the number of samples used for initialization is critical for the active learning process. Especially, when the number of initial samples is varied across the datasets [30], one should mention how this number has been tuned.

For transparent evaluation of the selection strategy, we propose that algorithms with an initialization phase should be seen as a two step selection strategy. In the first step, labeling candidates are chosen according to an initialization strategy (e.g., random) which is stopped by a comprehensible stopping criterion. Then, the real active learning method can proceed. As this initialization phase is now part of the active learning algorithm it should be somehow evaluated (e.g., regarding robustness) and included in the learning curves [30, 37].

3.4 Parameter Tuning

Tuning parameters for classifiers is very difficult with only a few labels available. Unfortunately, these tuning procedures are often not described in great detail. Yanik et al. [50] used a grid search approach in an 5 fold cross validation after each label acquisition to tune the parameters of the SVM. Similarly, Tuia et al. [43] tune their parameters for their SVM. Both do not describe, on which data this is executed. Using a hold out tuning set [13, 27] is not valid in AL unless these additional labels are comprehensibly selected and included in the evaluation (i.e., considering them in the number of acquired labels in the learning curve). As in passive classification tasks, it is strictly forbidden to tune the parameters using the evaluation instances.

One could also argue that parameters should be adapted during learning as the number of training instances is increased by AL which affects the capability of generalization. This means, we either use a pre-trained mediocre classifier because parameters are tuned for a specific labeling situation, or we re-calibrate the parameters during learning which means that classifiers become different across selection methods which also biases the results.

Another way is to use standard parameter with normalized features (e.g. z-normalized) [25].

3.5 Proposing an AL Evaluation Methodology

In order to achieve reliable results across selection strategies, we propose the following methodology for AL evaluation:

- Use exactly the same robust classifier for every AL method when comparing and try to sync the parameters of these classifiers.
- Capture the effect of different AL methods on multiple datasets using at least 50 repetitions.
- Start with an initially unlabeled set. If you need initial training instances, sample randomly and explain how to determine the number of samples.
- Use either a clear defined stopping criterion or enough label acquisitions (sample until convergence).
- Show learning curves (incl. quartiles) with reasonable performance measures.
- Present pairwise differences in terms of significance and effect size (Wilcoxon signed rank test).

4 Pilot Study: Influence of the Number of Repetitions

The major challenge of AL evaluation is to measure the effect of improvement although the variance of results might be high: Especially in the early learning stages ($1\% - 10\%$ of the data are labeled), the classification performance varies a lot. This is where the differences across AL methods are highest. Hence, experiments have to be repeated multiple times to yield reliable results as mentioned before. In this section, we provide an exemplary evaluation methodology using a 5-fold cross validation.

For these experiments, we solely used one dataset from the UCI machine learning repository, named Mammographic Mass [3]. We chose this dataset as it is a typical representative for an AL dataset regarding the number of instances and features. For classification, we decided to use a robust classifier based on Gaussian kernel density estimation, namely a Parzen Window Classifier (PWC). Here, we only have one parameter: the bandwidth. In a pre-processing step, all categorical data has been dichotomized and all features are linearly transformed into $[0, 1]$ space. Hence, we use a standard bandwidth for the Gaussian kernel of the PWC of 0.2 as this seems to be reasonable. The AL algorithms are: Optimized Probabilistic AL [26], uncertainty sampling (Uncer) [29], an optimized version of expected error reduction from Chapelle (EER) [11], and random (Rand).

In 5-fold cross validation, we split the dataset D into 5 separate subsets ($D = D_1 \cup \ldots \cup D_5, D_i \cap D_j = \emptyset, i \neq j$) to build disjoint candidates and evaluation sets $(\mathcal{T}_i, \mathcal{E}_i)$. In this subsection, we applied AL 5 times on four of the subsets and evaluated the trained classifier on the left out subset.

Performing solely one complete 5-fold cross validation, as shown in Fig. 3, the performances might vary a lot. Furthermore, the ranking of the final performance (after 60 labels have been acquired) changes completely. The left evaluation shows OPAL being the best, followed by Expected Error Reduction, Random, and Uncertainty Sampling. Using another seed (right plot), the ranking is different: First OPAL, then Random, Uncertainty Sampling, and Expected Error

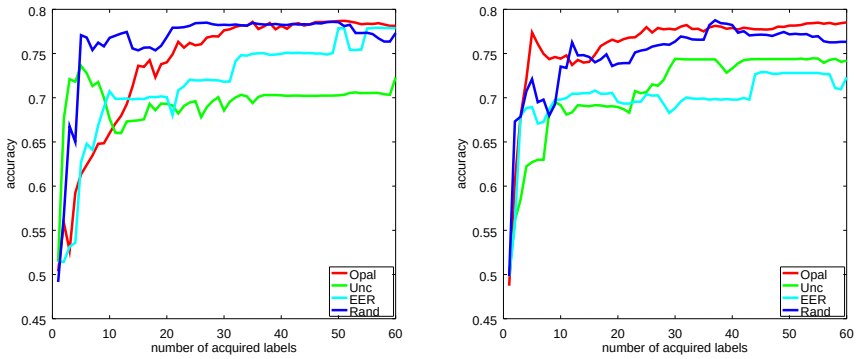

Fig. 3. Results of a 5-fold cross validation: two executions with different seeds of a complete 5-fold cross validation.

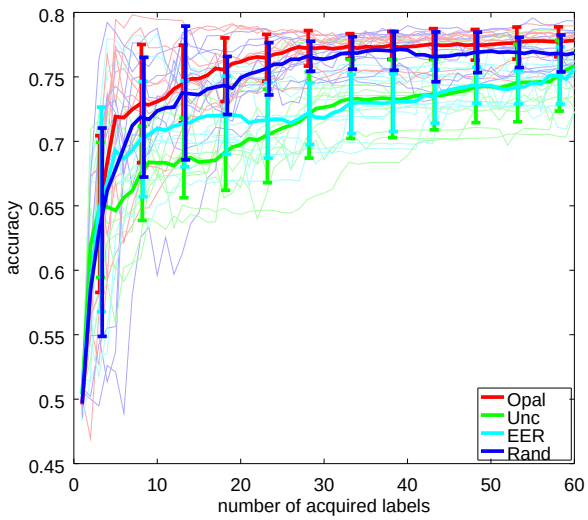

Fig. 4. Mean results of 10 times repeated 5-fold cross validations

Reduction. This clearly shows that a 5-fold cross validation evaluation for these AL methods on this dataset using a PWC is not sufficient. Similar experiments (not shown due to space restrictions) show that it is also true for other datasets and other classifiers. Repeating this 5-fold cross validation 10 times as shown in Fig. 4, provides much more stable results that are also comparable to the ones from the following experiment.

5 Challenges of realistic evaluation

Publications from companies such as Microsoft [24, 35], IBM [32], or Mitsubishi [23] show the growing interest in AL and its practical usefulness. AL has been successfully applied to solve problems such as on-road vehicle detection [40] or in recommender systems [28]. Unfortunately, these systems are highly specialized and often cannot be easily used for related problems.

In contrast to lab experiments, real active learning approaches only have one shot to learn. Hence, not the mean performance of multiple repetitions is of interest but the pairwise comparisons of the different methods. Because of high variances, it is still difficult to ensure a certain improvement of performance of one selection algorithm against others. This is the reason for many researchers arguing that random sampling is still a powerful baseline [10].

One of the main challenges to apply active learning in practice is to know when to stop querying for new label information. By now, in real-world applications, the AL process stops when a given "labeling budget" has been consumed. For example, in [40] the performance of the investigated AL approaches is done after a *fixed number* of queried samples. But, this may be a waste of resources, both in terms of time and money. Thus, the active learner should be able to asses its own performance. Here, different problems occur: a) collecting a separate evaluation dataset by randomly sampling instances is expensive, b) the collected data can not be used for performance estimation due to the sampling bias [12]. Some research work has been done to analyze when to stop the AL process besides estimating the performance directly [14, 34, 45]. It has been shown that it is possible to identify when a learning process might be saturated, but none provides information about the real classification performance.

In *dedicated collaborative interactive learning (D-CIL)* [9], different realistic applications for AL have been outlined. It addresses AL processes that are *interactive* – the information flows from humans to the active learner and vice versa, *collaborative* – multiple domain experts collaborate, and *dedicated* – a small number of benevolent domain experts interact with the active learner in order to support the selection process. Even though the oracles are impersonated by benevolent domain experts, they are still prone to error. Their labeling performance may depend on the labeler's experience, form of the day, or the complexity degree of the learning problem. In case of an *opportunistic* active learner [4], the oracles are not necessarily embodied by benevolent domain experts. Similar smart systems, simulation systems, or own sensors of the learning

system may assemble together or separately the oracle. Furthermore, there is high heterogeneity between these oracles, and their number is not fixed.

To summarize, AL research is mostly based on the following (limiting) assumptions [9]: a) the classification problem is well-defined (i.e., the number of classes and features are known in advance), b) labeled samples are available at the beginning of the learning process, c) uniform labeling cost (i.e., identical labeling costs for all samples), d) the oracle is omnipresent and omniscient, e) there exists a ground truth, based on which the performance of the active learner is evaluated. However, these assumptions often do not hold in real-world applications. Although, a large variety of specialized solutions is given which solve single problems, there is further work necessary to apply methods in a real-world setting. Here, a central aspect is the lack of comparability across different approaches which is a critical point for practitioners to apply AL in their specific domain.

6 Conclusion and Outlook

In this article, we summarized various challenges of AL evaluation with regard to being reliable, realistic, and comparable. Some of these appear naturally by the problem's definition, others are defined through the demands of real-world applications. We proposed an evaluation methodology to initialize a discussion on a gold standard for AL evaluation and provided preliminary results in a pilot study which shows the importance of many repetitions in AL which hopefully leads to comparable results without repeating whole experiments. Nevertheless, it is essential to report all details of evaluation to be able to reproduce the results of a paper. Those details have been discussed in this paper.

As future work, we plan to extend this literature overview and refine our proposed methodology. Additionally, we aim at providing a large comparison of different methodologies showing the effect of each component for different selection strategies. In this paper, we excluded the whole discussion of online algorithms and methods for evolving datastreams. Providing a valid evaluation framework for one-shot AL, is one of the goals of future research.

Our vision is to develop an evaluation system, enabling researchers and practitioners to collaborate. This system will provide a web-based user interface like OpenML [44] showing detailed information about different AL methods and their specific characteristics in relation to different tasks. In that way, we aim to standardize AL evaluation in order to simplify the steps towards practical solutions and fair comparison.

References

1. Aggarwal, C.C., Kong, X., Gu, Q., Han, J., Yu, P.S.: Active learning: A survey. In: Aggarwal, C.C. (ed.) Data Classification: Algorithms and Applications, pp. 571–606. CRC Press (2014)
2. Aldogan, D., Yaslan, Y.: A comparison study on ensemble strategies and feature sets for sentiment analysis. Lecture Notes in Electrical Engineering 363, 359–370 (2016)
3. Asuncion, A., Newman, D.J.: UCI machine learning repository (2015), `http://archive.ics.uci.edu/ml/`
4. Bahle, G., Calma, A., Leimeister, J.M., Lukowicz, P., Oeste-Reiß, S., Reitmaier, T., Schmidt, A., Sick, B., Stumme, G., Zweig, K.: Lifelong learning and collaboration of smart technical systems in open-ended environments – Opportunistic Collaborative Interactive Learning. In: International Conference on Autonomic Computing. IEEE, Würzburg, Germany (2017)
5. Bilgic, M., Getoor, L.: Active learning for networked data. Computer 411(29-30), 2712–2728 (2010)
6. Bouguelia, M.R., Belaïd, Y., Belaïd, A.: An adaptive streaming active learning strategy based on instance weighting. Pattern Recognition Letters 70, 38–44 (2016)
7. Brinker, K.: Incorporating diversity in active learning with support vector machines. In: Proceedings of the 20th International Conference on Machine Learning (ICML). pp. 59–66 (2003)
8. Cai, W., Zhang, Y., Zhou, S., Wang, W., Ding, C., Gu, X.: Active learning for support vector machines with maximum model change. In: Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases. vol. 8724 (2014)
9. Calma, A., Leimeister, J.M., Lukowicz, P., Oeste-Rei, S., Reitmaier, T., Schmidt, A., Sick, B., Stumme, G., Zweig, A.K.: From active learning to dedicated collaborative interactive learning. In: Varbanescu, A.L. (ed.) 29th International Conference on Architecture of Computing Systems, Workshop Proceedings. pp. 1–8. VDI Verlag, Nuremberg, Germany (2016)
10. Cawley, G.C.: Baseline methods for active learning. In: Active Learning and Experimental Design Workshop in Conjunction with AISTATS 2010. pp. 47–57 (2011)
11. Chapelle, O.: Active learning for parzen window classifier. In: Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics. pp. 49–56 (2005)
12. Dasgupta, S., Hsu, D.: Hierarchical sampling for active learning. In: Proceedings of the 25th International Conference on Machine learning. pp. 208–215. ACM (2008)
13. Demir, B., Persello, C., Bruzzone, L.: Batch-mode active-learning methods for the interactive classification of remote sensing images. IEEE Transactions on Geoscience and Remote Sensing 49(3), 1014–1031 (2011)
14. Dimitrakakis, C., Savu-Krohn, C.: Cost-Minimising Strategies for Data Labelling: Optimal Stopping and Active Learning, pp. 96–111. Springer Berlin Heidelberg, Berlin, Heidelberg (2008)
15. Efron, B.: Bootstrap methods: another look at the jackknife. The annals of Statistics pp. 1–26 (1979)
16. Ertekin, S., Huang, J., Bottou, L., Giles, L.: Learning on the border: Active learning in imbalanced data classification. In: Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management. pp. 127–136. CIKM '07, ACM, New York, NY, USA (2007)

17. Flach, P., Hernandez-Orallo, J., Ferri, C.: A coherent interpretation of AUC as a measure of aggregated classification performance. In: Getoor, L., Scheffer, T. (eds.) Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA. pp. 657–664. ACM, New York, NY, USA (2011)
18. Fu, Y., Zhu, X., Li, B.: A survey on instance selection for active learning. Knowledge and Information Systems 35(2), 249–283 (2013)
19. Hand, D.J.: Measuring classifier performance: a coherent alternative to the area under the roc curve. Machine Learning 77(1), 103–123 (2009)
20. Hu, B.G., Dong, W.M.: A study on cost behaviors of binary classification measures in class-imbalanced problems. arXiv preprint arXiv:1403.7100 (2014)
21. Huang, K.h., Lin, H.t.: A novel uncertainty sampling algorithm for cost-sensitive multiclass active learning. In: 2016 IEEE 16th International Conference on Data Mining (ICDM) (2016)
22. Huang, S.j., Jin, R., Zhou, Z.H.: Active learning by querying informative and representative examples. In: NIPS'10 Proceedings of the 23rd International Conference on Neural Information Processing Systems. pp. 892–900 (2010)
23. Joshi, A.J., Porikli, F., Papanikolopoulos, N.P.: Scalable active learning for multiclass image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(11), 2259–2273 (2012)
24. Kapoor, A., Horvitz, E., Basu, S.: Selective supervision: Guiding supervised learning with decision-theoretic active learning. In: Veloso, M.M. (ed.) Proceedings of the 20th International Joint Conference on Artifical Intelligence. pp. 877–882. Morgan Kaufmann Publishers Inc. (2007)
25. Kottke, D., Krempl, G., Lang, D., Teschner, J., Spiliopoulou, M.: Multi-class probabilistic active learning. In: ECAI. Frontiers in Artificial Intelligence and Applications, vol. 285, pp. 586–594. IOS Press (2016)
26. Krempl, G., Kottke, D., Lemaire, V.: Optimised probabilistic active learning (OPAL) for fast, non-myopic, cost-sensitive active classification. Machine Learning pp. 1–28 (2015)
27. Krempl, G., Kottke, D., Spiliopoulou, M.: Probabilistic active learning: Towards combining versatility, optimality and efficiency. In: Proceedings of the 17th International Conference on Discovery Science (DS), Bled. Lecture Notes in Computer Science, Springer (2014)
28. Lamche, B., Trottmann, U., Wörndl, W.: Active Learning Strategies for Exploratory Mobile Recommender Systems. In: Proceedings of the Fourth Workshop on Context-Awareness in Retrieval and Recommendation. pp. 10–17. Amsterdam, Niederlande (2014)
29. Lewis, D., Gale, W.: A sequential algorithm for training text classifiers. In: Conference on Research and Development in Information Retrieval. pp. 3–12. ACM/Springer, New York, NY (1994)
30. Li, X., Guo, Y.: Active learning with multi-label svm classification. In: Proceedings of the 23rd International Joint Conference on Artificial Intelligence (2013)
31. Longstaff, B., Reddy, S., Estrin, D.: Improving activity classification for health applications on mobile devices using active and semi-supervised learning. Proceedings of the 4th International ICST Conference on Pervasive Computing Technologies for Healthcare (2010)
32. Melville, P., Sindhwani, V.: Active dual supervision: Reducing the cost of annotating examples and features. In: Workshop on Active Learning for Natural Language Processing. pp. 49–57. Boulder, CO (2009)
33. Nuzzo, R.: Statistical errors. Nature 506(7487), 150 (2014)

34. Olsson, F., Tomanek, K.: An intrinsic stopping criterion for committee-based active learning. In: Conference on Computational Natural Language Learning. pp. 138–146. Boulder, CO (2009)
35. Paquet, U., Gael, J.V., Stern, D., Kasneci, G., Herbrich, R., Graepel, T.: Vuvuzelas & active learning for online classification. In: Workshop on Computational Social Science and the Wisdom of Crowds. pp. 1–5. Whistler, BC (2010)
36. Parker, C.: An analysis of performance measures for binary classifiers. In: Proceedings of the 11th IEEE International Conference on Data Mining (ICDM). pp. 517–526. IEEE (2011)
37. Pasolli, E., Melgani, F.: Active learning methods for electrocardiographic signal classification. IEEE Transactions on Information Technology in Biomedicine 14(6), 1405–16 (2010)
38. Reitmaier, T., Sick, B.: Let us know your decision: Pool-based active training of a generative classifier with the selection strategy 4DS. In: Information Sciences - Informatics and Computer Science Intelligent Systems Applications. vol. 230, pp. 106–131 (2013)
39. Settles, B.: Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin, Department of Computer Science (2009)
40. Sivaraman, S., Trivedi, M.M.: Active learning for on-road vehicle detection: a comparative study. Machine Vision and Applications pp. 1–13 (2011)
41. Son, Y., Lee, J.: Active learning using transductive sparse bayesian regression. Information Sciences 374, 240–254 (2016)
42. Tomanek, K., Morik, K.: Inspecting sample reusability for active learning. In: Guyon, I., Cawley, G.C., Dror, G., Lemaire, V., Statnikov, A.R. (eds.) Workshop on Active Learning and Experimental Design. JMLR Proceedings, vol. 16, pp. 169–181 (2011)
43. Tuia, D., Volpi, M., Copa, L., Kanevski, M., Munoz-Mari, J.: A survey of active learning algorithms for supervised remote sensing image classification. IEEE Journal of Selected Topics in Signal Processing 5(3), 606–617 (2011)
44. Vanschoren, J., van Rijn, J.N., Bischl, B., Torgo, L.: Openml: Networked science in machine learning. SIGKDD Explorations 15(2), 49–60 (2013)
45. Vlachos, A.: A stopping criterion for active learning. Computer Speech & Language 22(3), 295–312 (2008)
46. Wang, J., Park, E.: Active learning for penalized logistic regression via sequential experimental design. Neurocomputing 222, 183–190 (2017)
47. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics bulletin 1(6), 80–83 (1945)
48. Yan, Y., Rosales, R., Fung, G., Dy, J.G.: Active learning from crowds. Proceedings of the 28th International Conference on Machine Learning pp. 1161–1168 (2011)
49. Yang, Y., Ma, Z., Nie, F., Chang, X., Hauptmann, A.G.: Multi-class active learning by uncertainty sampling with diversity maximization. International Journal of Computer Vision 113(2), 113–127 (2014)
50. Yanik, E., Sezgin, T.M.: Active learning for sketch recognition. Computers and Graphics (Pergamon) 52, 93–105 (2015)