Active Learning

Marianne Stecklina

Otto-von-Guericke-Universität Magdeburg
marianne.stecklina@st.ovgu.de

Abstract. While unlabeled instances are often easy to obtain, labeling of instances for training is a bottleneck of classification, as it requires human effort. Active learning methods aim to reduce the number of training instances and consequently the labeling costs by actively choosing from which instances to learn. This strategy has been shown to yield good classification results with less training compared to passively learning a training set. In this report, we summarize scenarios for active learning and different approaches how to select beneficial instances for training. Furthermore, a selection of practical challenges and setting variants is discussed.

1 Introduction

Active learning, also referred to as optimal experimental design in statistics, is a subfield of machine learning. In order to understand the motivation for active learning, it is useful to have a look at classification first. Given a set of training instances with known labels, the task is to learn a classifier that is able to predict the correct label for a new instance. To obtain a high classification accuracy, those classifiers need to be trained with a large number of labeled instances. Labels are obtained by an "oracle", in most cases a human annotator. In practice, an enormous amount of unlabeled data is often available: One might think of articles in the web, images uploaded to social media platforms or video sequences automatically recorded by cameras. However, to annotate this data by humans is time-consuming and expensive. Here arises the research field of active learning.

The common approach to train a classifier with a set of already labeled instances is a rather passive way of learning. In contrast, the paradigm of active learning can be summarized as follows: A learning algorithm actively chooses the instances for training it considers most beneficial. The classifier obtains good accuracy values with less training, as it only learns from interesting or controversial instances selected by the learner, see Fig. 1. Less training means less labeling effort for annotators and therefore a reduction of costs.

The obvious questions as how to pose queries and measure the usefulness of an instance for training have been answered in many ways in the literature. This report gives an overview of these answers and discusses a selection of practical considerations and setting variants investigated in recent years. The application of active learning methods in different domains revealed many new questions and led to further research topics in the field of active learning.

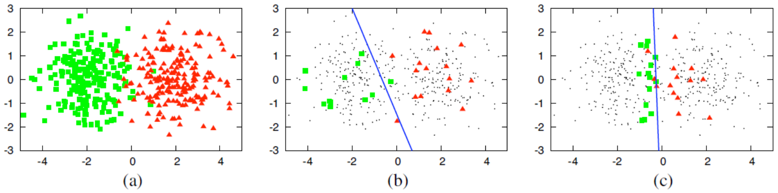

Fig. 1: This figure illustrates the classification of an artificial data set (a), where the classifier achieves an accuracy of 70% trained on 30 randomly chosen instances (b) compared to an accuracy of 90% trained on 30 activly chosen instances (c). Image taken from [26].

2 Scenarios

Active Learning addresses applications in which unlabeled data is easy to obtain, but labeling is costly and reducing the necessary amount of training instances therefore highly desirable. In the literature, a small number of scenarios is differentiated. They differ in the way the learner decides for a query. Which scenario to use depends on the specific application domain, sometimes several scenarios are possible.

2.1 Query synthesis

An early investigated active learning scenario is query synthesis [1], where the learner is allowed to generate new instances it finds helpful to query. To do so, the learner needs to know the feature dimensions and ranges of the input space. A successful application of query synthesis was proposed by King et al. [13, 14]. They developed a laboratory robot, which autonomously designs and performs experiments to study the growth of microbial strains. In this context, all possible query instances - the experiments - generated by the robot's active learning component are meaningful. If this is not the case, an unexpected problem can occur, as reported by Baum and Lang [2]. They aimed to classify handwritten characters and therefore applied query synthesis to train a neural net. The annotators were confronted with a large number of query images that contained non-existing or hybrid characters impossible to label for a human, see Fig. 2. In this case, much labeling effort is wasted on training examples without practical usefulness. This limitation is addressed by the stream- and pool-based scenarios.

2.2 Stream-based active learning

The stream-based scenario, also referred to as selective sampling [3], relies on the assumption that unlabeled data can be obtained at low cost. A new unlabeled instance is sampled in a first step and then the task for the learner is to decide, whether or not to acquire its label. How to make this decision has been investigated by different researchers: Dagan and Engelson suggested to introduce a measure of utility and to query

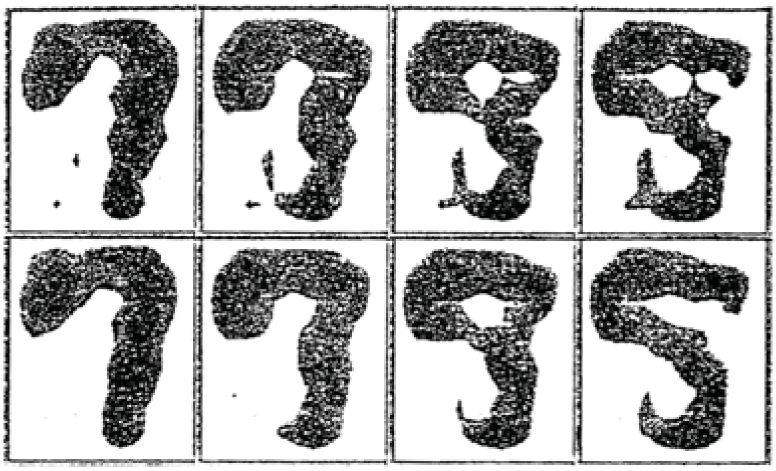

Fig. 2: Queries hard to label for humans proposed by an active learner in handwritten character classification. Image taken from [2].

promising instances according to this measure with a higher probability. Another approach by Cohn et al. explicitly defines a region of uncertainty and queries instances which fall therein.

2.3 Pool-based active learning

Huge amounts of unlabeled data often available in practice led to the idea of pool-based active learning. An instance to query is selected from a large pool of unlabeled data, typically using an utility measure that is used to compare instances in the pool. This scenario was mentioned first by Lewis and Gale [17], who used active learning for text classification. A large number of scientific works with varying application domains followed this publication. Pool-based methods were not only successfully applied in the domain of text classification [19, 31], but also for information extraction [27], image and video classification [9, 31, 37, 38] and speech recognition [32], among many others.

3 Active learning approaches

Introducing the task of active learning to choose the most beneficial instances for training, the most important question is: How can we evaluate, which instances are valuable for training? This section introduces different strategies to answer this question. Due to its practical importance, the pool-based scenario is considered here.

3.1 Uncertainty sampling

The idea of uncertainty sampling is straight forward: In each iteration, the learner queries the instance of whose label the classifier is most uncertain. Lewis and Gale [17] were the first to propose this strategy along with a simple uncertainty measure for their binary classification task. An instance is denoted uncertain if the probabilistic classifier assigns a probability close to 0.5 - which means that both classes are equally likely. For multi-class problems, a more general uncertainty measure is needed; a common practice is to use entropy [11, 10].

3.2 Query by committee

Query by committee, proposed by Seung et al. [29] and further analyzed by Freund et al. [6], emerged from a quite different motivation than uncertainty sampling. Instead of a single classifier, an ensemble of classifiers - the "committee" - is used here. Whenever those classifiers disagree on the predicted label of an instance, this instance is an interesting candidate to query. Theoretically speaking, this strategy corresponds to minimizing the version space, which in machine learning denotes the set of hypotheses consistent with the training data. The more the version space is reduced in each iteration, the faster the model converges to the true hypothesis. To obtain a query that can discard a large number of hypotheses, a measure of disagreement among the classifiers is needed. The most common choices are long-established measures from information theory, namely the entropy of committee votes and the Kullback-Leibler divergence between the single classifiers' and the committee's prediction. A modified version of the latter called Jenson-Shannon divergence is applied in the work of Melville et al. [21]. Ngai and Yarowski [22] introduced the f-compliment as a new disagreement measure and showed in experiments that the resulting learning performance is better compared to using vote entropy.

3.3 Error reduction

Error reduction [23] is motivated by the classification task itself: The best instance to query is the one which, once incorporated into the training set, most reduces the future classification error. As neither the label of a query candidate nor its influence on the future classification performance is known, the expectation of the error instead of the error itself is reduced. Successful modifications of the original algorithm can be found in the literature. Combining error reduction with semi-supervised learning, an idea of Zhu et al. [40], significantly outperformed uncertainty sampling on handwritten digit recognition and text classification. Guo and Greiner [8] assume the current likeliest label to be the true label of an instance, and change their query strategy whenever this assumption turned out to be false after asking the oracle.

3.4 Density-weighted methods

Uncertainty sampling and query by committee tend to query outliers, as these instances are likely to be uncertain or produce disagreement among a classifier committee [26]. Error reduction avoids this problem by considering the unlabeled data, but is therefore computationally intensive. Density-weighted methods aim to combine the advantages: exploit the underlying distribution and at the same time be fast. A simple heuristic is to multiply one of the former explained utility measures (e.g. uncertainty) with a density factor; this way, outliers and instances in sparse areas are downvoted. Experiments performed by different researchers [7, 19, 27, 36] showed that density-weighted approaches yield better results than methods which does not consider the underlying distribution.

4 Challenges in practice

None of the previously introduced algorithms is in general superior to the others, as large-scale comparisons show [15, 25, 27]. However, the results support the assumption that active learning is able to save costs when used instead of passive learning. In recent years, the idea of active learning spread beyond the research community and reached industry; companies like Google, IBM, Microsoft and Siemens increasingly rely on active learning methods for their real-world applications [28]. As a consequence, the focus of research moves towards solving challenges in practice, like the exemplary ones introduced in this section.

4.1 Cost-sensitive active learning

All of the above introduced methods aim to reduce the number of training instances by actively choosing queries, assuming that this achievement leads to a reduction of costs (e.g. time, money). However, this conclusion might not be true if annotation costs differ between the instances. In speech recognition for example, a long recording might be more beneficial for the learning process than a shorter one and is therefore queried, but a human annotator will spend a considerable amount of time labeling it. The term cost-sensitive active learning refers to approaches that address the problem of varying costs: different labeling costs among instances on the one hand, different misclassification costs per class on the other hand. The former mentioned laboratory robot developed by King et al. [14] considers the material costs when designing its next experiment to perform. Kapoor et al. [12] applied a cost-sensitive algorithm to a voicemail classification problem. Their utility measure combines both the labeling costs and the expected future misclassification costs. In Kapoor's work, a simple heuristic was used to estimate these costs; the work of Settles et al. [28] goes further and additionally trains a regression model to estimate the true labeling costs. A fast prediction of the future misclassification costs is the focus of OPAL, an algorithm proposed by Krempl et al. [16].

4.2 Noisy oracles

Until now, we assumed the oracle (e.g. a human annotator) to label always correct. If this is not the case, a learner can decide between querying a new instance or an already known one to confirm its label. Sheng et al. [30] observed a positive effect by querying suspect instances multiple times, their decision for a query is based on both the oracle and model uncertainty. In contrast to this work, Donmez et al. [4] also allowed for different noise levels among the oracles. The restriction to constant noise levels over time was dropped in their follow-up work [5]. In case of human annotators, Wallace et al. [33] introduced the idea to ask annotators about their certainty. This way, the model quality improved while benefiting from both experts and novices.

5 Related tasks

In the narrower sense, the term active learning refers to selecting the best instance to request a label for. In the broader sense, it covers all methods that actively choose what to learn. Active feature acquisition and active class selection are two problem settings in that sense, which are far less popular than the common active learning task.

5.1 Active feature acquisition

Active feature acquisition builds on the frequent occurrence of missing values in practice. In a medical context, missing values could be results of further diagnosis procedures, which have not been performed for the patient yet. To obtain a more reliable diagnosis, the most informative medical test should be performed next. An active feature acquisition algorithm aims to decide, which feature to request in order to achieve a better classification performance. In the described example, a feature corresponds to a medical test, the correctness of the diagnosis to the classification performance. The techniques used here are similar to those from common active learning: One could request the least certain feature [39], or the one with the highest expected information value [24]. Melville et al. [20] suggested to request features that are likely to change the model's classification.

5.2 Active class selection

In active class selection, the learner decides for a class and subsequently a new instance of that class is generated - directly opposed to the common active learning setting. Lomasky et al. [18] were the first to investigate active class selection and proposed multiple active class selection strategies: inverse, original proportion, accuracy improvement and redistricing. Wu et al. successfully applied a subset of these methods for arousal classification [35] and combined active class selection with transfer learning to train a brain-computer interface [34].

6 Conclusion

Active learning has the potential to considerably reduce the human effort needed to label training instances for classification tasks. By actively choosing the instances from which to learn, classifiers become better with less training. This report summarized the most investigated scenarios and common algorithms for the pool-based scenario, which differ in the way they evaluate the usefulness of instances. As the focus of current research moves towards applying active learning in different domains in industry, practical considerations become more import. In this report, two challenges in practice were discussed in detail. Finally, we introduced two related tasks, which also profit from an active learning manner, although their setting is different.

References

1. Angluin, D.: Queries and concept learning. Machine learning 2(4), 319–342 (1988)
2. Baum, E.B., Lang, K.: Query learning can work poorly when a human oracle is used. In: International Joint Conference on Neural Networks. vol. 8, p. 8 (1992)
3. Cohn, D., Atlas, L., Ladner, R.: Improving generalization with active learning. Machine learning 15(2), 201–221 (1994)
4. Donmez, P., Carbonell, J.G.: Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In: Proceedings of the 17th ACM conference on Information and knowledge management. pp. 619–628. Association for Computing Machinery (2008)
5. Donmez, P., Carbonell, J.G., Schneider, J.: Efficiently learning the accuracy of labeling sources for selective sampling. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 259–268. Association for Computing Machinery (2009)
6. Freund, Y., Seung, H.S., Shamir, E., Tishby, N.: Selective sampling using the query by committee algorithm. Machine learning 28(2-3), 133–168 (1997)
7. Fujii, A., Tokunaga, T., Inui, K., Tanaka, H.: Selective sampling for example-based word sense disambiguation. Computational Linguistics 24(4), 573–597 (1998)
8. Guo, Y., Greiner, R.: Optimistic active-learning using mutual information. In: IJCAI. vol. 7, pp. 823–829 (2007)
9. Hauptmann, A.G., Lin, W.H., Yan, R., Yang, J., Chen, M.Y.: Extreme video retrieval: joint maximization of human and computer performance. In: Proceedings of the 14th ACM international conference on Multimedia. pp. 385–394. Association for Computing Machinery (2006)
10. Holub, A., Perona, P., Burl, M.C.: Entropy-based active learning for object recognition. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. pp. 1–8. IEEE (2008)
11. Joshi, A.J., Porikli, F., Papanikolopoulos, N.: Multi-class active learning for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2372–2379. IEEE (2009)

12. Kapoor, A., Horvitz, E., Basu, S.: Selective supervision: Guiding supervised learning with decision-theoretic active learning. In: IJCAI. vol. 7, pp. 877–882 (2007)
13. King, R.D., Rowland, J., Oliver, S.G., Young, M., Aubrey, W., Byrne, E., Liakata, M., Markham, M., Pir, P., Soldatova, L.N., et al.: The automation of science. Science 324(5923), 85–89 (2009)
14. King, R.D., Whelan, K.E., Jones, F.M., Reiser, P.G., Bryant, C.H., Muggleton, S.H., Kell, D.B., Oliver, S.G.: Functional genomic hypothesis generation and experimentation by a robot scientist. Nature 427(6971), 247–252 (2004)
15. Körner, C., Wrobel, S.: Multi-class ensemble-based active learning. In: European Conference on Machine Learning. pp. 687–694. Springer (2006)
16. Krempl, G., Kottke, D., Lemaire, V.: Optimised probabilistic active learning (opal). Machine Learning 100(2-3), 449–476 (2015)
17. Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. In: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 3–12. Springer (1994)
18. Lomasky, R., Brodley, C.E., Aernecke, M., Walt, D., Friedl, M.: Active class selection. In: European Conference on Machine Learning. pp. 640–647. Springer (2007)
19. McCallum, A.K., Nigam, K.: Employing EM and pool-based active learning for text classification. In: Proc. International Conference on Machine Learning. pp. 359–367. Citeseer (1998)
20. Melville, P., Saar-Tsechansky, M., Provost, F., Mooney, R.: Active feature-value acquisition for classifier induction. In: Fourth IEEE International Conference on Data Mining, 2004. pp. 483–486. IEEE (2004)
21. Melville, P., Yang, S.M., Saar-Tsechansky, M., Mooney, R.: Active learning for probability estimation using jensen-shannon divergence. In: European Conference on Machine Learning. pp. 268–279. Springer (2005)
22. Ngai, G., Yarowsky, D.: Rule writing or annotation: Cost-efficient resource usage for base noun phrase chunking. In: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. pp. 117–125. Association for Computational Linguistics (2000)
23. Roy, N., McCallum, A.: Toward optimal active learning through monte carlo estimation of error reduction. International Conference on Machine Learning pp. 441–448 (2001)
24. Saar-Tsechansky, M., Melville, P., Provost, F.: Active feature-value acquisition. Management Science 55(4), 664–684 (2009)
25. Schein, A.I., Ungar, L.H.: Active learning for logistic regression: an evaluation. Machine Learning 68(3), 235–265 (2007)
26. Settles, B.: Active learning. Synthesis Lectures on Artificial Intelligence and Machine Learning 6(1), 1–114 (2012)
27. Settles, B., Craven, M.: An analysis of active learning strategies for sequence labeling tasks. In: Proceedings of the conference on empirical methods in natural language processing. pp. 1070–1079. Association for Computational Linguistics (2008)

28. Settles, B., Craven, M., Friedland, L.: Active learning with real annotation costs. In: Proceedings of the NIPS workshop on cost-sensitive learning. pp. 1–10 (2008)
29. Seung, H.S., Opper, M., Sompolinsky, H.: Query by committee. In: Proceedings of the fifth annual workshop on Computational learning theory. pp. 287–294. Association for Computing Machinery (1992)
30. Sheng, V.S., Provost, F., Ipeirotis, P.G.: 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. pp. 614–622. Association for Computing Machinery (2008)
31. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. Journal of machine learning research 2(Nov), 45–66 (2001)
32. Tur, G., Hakkani-Tür, D., Schapire, R.E.: Combining active and semi-supervised learning for spoken language understanding. Speech Communication 45(2), 171–186 (2005)
33. Wallace, B.C., Small, K., Brodley, C.E., Trikalinos, T.A.: Who should label what? instance allocation in multiple expert active learning. In: SIAM International Conference on Data Mining. pp. 176–187. Society for Industrial and Applied Mathematics (2011)
34. Wu, D., Lance, B.J., Parsons, T.D.: Collaborative filtering for brain-computer interaction using transfer learning and active class selection. PloS one 8(2), e56624 (2013)
35. Wu, D., Parsons, T.D.: Active class selection for arousal classification. In: Affective Computing and Intelligent Interaction, pp. 132–141. Springer (2011)
36. Xu, Z., Akella, R., Zhang, Y.: Incorporating diversity and density in active learning for relevance feedback. In: European Conference on Information Retrieval. pp. 246–257. Springer (2007)
37. Yan, R., Yang, J., Hauptmann, A.: Automatically labeling video data using multi-class active learning. In: Ninth IEEE International Conference on Computer Vision. pp. 516–523. IEEE (2003)
38. Zhang, C., Chen, T.: An active learning framework for content-based information retrieval. IEEE transactions on multimedia 4(2), 260–268 (2002)
39. Zheng, Z., Padmanabhan, B.: On active learning for data acquisition. In: IEEE International Conference on Data Mining. pp. 562–569. IEEE (2002)
40. Zhu, X., Lafferty, J., Ghahramani, Z.: Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions. In: International Conference on Machine Learning 2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining. vol. 3 (2003)