Active Learning: Applications, Foundations and Emerging Trends (Tutorial)

Georg Krempl
University Magdeburg, Germany
georg.krempl@ovgu.de

Vincent Lemaire
Orange Labs, France
vincent.lemaire@orange.com

Edwin Lughofer
University Linz, Austria
edwin.lughofer@jku.at

Daniel Kottke
University Kassel, Germany
daniel.kottke@uni-kassel.de

1 Introduction

Active learning optimizes the interaction between artificial data mining/machine learning systems and humans. For example, its techniques are used for selecting the most relevant information to be requested, or for determining the most informative experiment to be performed. With increasing volumes of data, contrasting limited human supervision capacities, the optimization of this interaction has become even more important. Hence, active sampling and data acquisition techniques could contribute to the design and modeling of highly intelligent learning systems.

This tutorial presents the basic techniques for pool-based and on-line active learning for streams. It contains a summary of the common concepts like version space partitioning, uncertainty sampling and decision theoretic approaches, and shortly mentions the connection between reinforcement learning and active learning. We show how these concepts can be used in data streams and on-line applications, and discuss the main challenges of stream active learning. Finally, we evaluate frameworks for pool-based and stream-based active learning to validate if a method is applicable for a specific demand.

2 Basic Strategies in Active Learning

Adaptive sampling [Sin06] and selective sampling [Roy01] are the two main scenarios to set active learning [Set12]. This tutorial focus on selective sampling. After defining the active learning problem in detail [Set12], we discuss three basic techniques: The selection criterion of methods using *version space partitioning*, is based on the disagreement between hypothesis [Ruf89], e.g., the disagreement within an ensemble of classifiers. *Uncertainty sampling* is an information theoretic approach, selecting instances based on the classifier's uncertainty, e.g., instances near the decision boundary. Decision theoretic approaches are among others *expected error reduction*, which simulate each label outcome for each label candidate, or *probabilistic active learning* [Kre14]. The latter uses local statistics of each candidate to estimate its gain in performance.

3 Reinforcement Learning and Active Learning

(Online) Reinforcement learning tries to improve itself by interacting with the environment while active learning tries to solve the traditional supervised learning problem with a human in the loop. Seeing the human as a part of the environment, as an agent [Kap05], it/he has to find a policy that maps states of the world to the actions to be taken in those states. When the agent has to learn the action to pick the next label candidate, one way to do this is to consider the informativeness of this sample as a form of reward and the current set of labeled and unlabeled data as the observation [Set08].

Copyright © by the paper's authors. Copying permitted for private and academic purposes.

In: G. Krempl, V. Lemaire, E. Lughofer, and D. Kottke (eds.): Proceedings of the Workshop Active Learning: Applications, Foundations and Emerging Trends, AL@iKNOW 2016, Graz, Austria, 18-OCT-2016, published at http://ceur-ws.org

4 On-line Active Learning for Streams

In this section, we provide an overview on *on-line* active learning concepts for data streams. These algorithms use different reliability concepts for instance selection, e.g., *conflict* and *ignorance* [Lug12]. We point out special challenges for on-line algorithms and stream active learning and discuss possible solutions. Thereby, a specific emphasis will be placed on active learning in connection with *evolving (fuzzy) systems* [Lug16], which are able to expand their structure on the fly to properly react on ignorance cases and thus to reduce uncertainty in the version space. A collection of some successful practical application examples, e.g., from quality control systems, viscose production, or image classification systems, gives an idea of how to use these methods in practice.

5 Evaluation of Active Learning

Comparing active learning methods is challenging, as the results of common evaluation methodologies like 5- or 10-fold cross-validation are often not reliable and therefore insufficient. Hence, we propose to use an evaluation framework using multiple randomly (seed-based) generated train and evaluation sets (typically 50 or more) [Kre14], which also allow pairwise comparisons. The second part of this topic concentrates on stream evaluation and shows why a separation of the temporal and spatial component of a stream active learner is mandatory in order to compare two active approaches [Kot15]. Furthermore, we discuss different forms of visualization like learning curves (performance vs. budget) and performance curves (performance vs. time) and show how the definition of the target function can affect the evaluation.

References

[Kap05] "Reinforcement learning for active model selection." A. Kapoor and R. Greiner. ACM SIGKDD Workshop on Utility-based Data Mining, 2005

[Kot15] "Probabilistic Active Learning in Datastreams." Daniel Kottke, Georg Krempl, Myra Spiliopoulou. International Symposium on Intelligent Data Analysis. Springer International Publishing, 2015.

[Kre14] "Optimised probabilistic active learning (OPAL)." Georg Krempl, Daniel Kottke, Vincent Lemaire Machine Learning 100.2-3 (2015): 449-476.

[Lug12] "Single-pass active learning with conflict and ignorance." Edwin Lughofer. Evolving Systems 3 (4), 251-271, 2012.

[Lug16] "Evolving Fuzzy Systems — Fundamentals, Reliability, Interpretability and Useability." Edwin Lughofer. in: Handbook of Computational Intelligence, World Scientific, pp. 67-135, 2016.

[Roy01] "Toward optimal active learning through sampling estimation of error reduction" Roy, N., McCallum, A. In: Proc. 18th International Conference on Machine Learning, pp. 441-448, 2001

[Ruf89] "What good are experiments?" Ritchey A. Ruff, T. G. Dietterich. Proceedings of the 6th International Workshop on Machine Learning, 1989.

[Set08] "An Analysis of Active Learning Strategies for Sequence Labeling Tasks." Burr Settles, Mark Craven Empirical Methods on Natural Language Processing (EMNLP) 2008

[Set12] "Active Learning." Burr Settles. Number 18 in Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan and Claypool Publishers, (2012).

[Sin06] "Active learning for adaptive mobile sensing networks" Singh, A., Nowak, R., Ramanathan, P. in International conference on Information processing in sensor networks, pp. 60–68, 2006