

Evaluation and Experimental Design in Data Mining and Machine Learning: Motivation and Summary of EDML 2019

Eirini Ntoutsi* Erich Schubert† Arthur Zimek‡ Albrecht Zimmermann§

1 Motivation

A vital part of proposing new machine learning and data mining approaches is evaluating them empirically to allow an assessment of their capabilities. Numerous choices go into setting up such experiments: how to choose the data, how to preprocess them (or not), potential problems associated with the selection of datasets, what other techniques to compare to (if any), what metrics to evaluate, etc. and last but not least how to present and interpret the results. Learning how to make those choices on-the-job, often by copying the evaluation protocols used in the existing literature, can easily lead to the development of problematic habits. Numerous, albeit scattered, publications have called attention to those questions and have occasionally called into question published results, or the usability of published methods [11, 4, 2, 9, 12, 3, 1, 5]. At a time of intense discussions about a reproducibility crisis in natural, social, and life sciences, and conferences such as SIGMOD, KDD, and ECML PKDD encouraging researchers to make their work as reproducible as possible, we therefore feel that it is important to bring researchers together, and discuss those issues on a fundamental level.

An issue directly related to the first choice mentioned above is the following: even the best-designed experiment carries only limited information if the underlying data are lacking. We therefore also want to discuss questions related to the availability of data, whether they are reliable, diverse, and whether they correspond to realistic and/or challenging problem settings.

2 Topics

In this workshop, we mainly solicited contributions that discuss those questions on a fundamental level, take stock of the state-of-the-art, offer theoretical arguments, or take well-argued positions, as well as actual evaluation papers that offer new insights, e.g., question published results, or shine the spotlight on the character-

istics of existing benchmark data sets. As such, topics include, but are not limited to

- Benchmark datasets for data mining tasks: are they diverse/realistic/challenging?
- Impact of data quality (redundancy, errors, noise, bias, imbalance, ...) on qualitative evaluation
- Propagation/amplification of data quality issues on the data mining results (also interplay between data and algorithms)
- Evaluation of unsupervised data mining (dilemma between novelty and validity)
- Evaluation measures
- (Automatic) data quality evaluation tools: What are the aspects one should check before starting to apply algorithms to given data?
- Issues around runtime evaluation (algorithm vs. implementation, dependency on hardware, algorithm parameters, dataset characteristics)
- Design guidelines for crowd-sourced evaluations

3 Contributions

The workshop featured a mix of invited speakers, a number of accepted presentations with ample time for questions since those contributions were expected to be less technical, and more philosophical in nature, and an extensive discussion on the current state, and the areas that most urgently need improvement, as well as recommendations to achieve those improvements.

3.1 Invited Presentations Four invited presentations enriched the workshop with focused talks around the problems of evaluation in unsupervised learning.

The first invited presentation by Ricardo J. G. B. Campello, University of Newcastle, was on “*Evaluation of Unsupervised Learning Results: Making the Seemingly Impossible Possible*”. Ricardo elaborated on the specific difficulties in the evaluation of unsupervised

*Leibniz University Hannover, Germany
& L3S Research Center, Germany

†Technical University Dortmund, Germany

‡University of Southern Denmark, Denmark

§University Caen Normandy, France

data mining methods (namely clustering and outlier detection) and reported on some recent solutions and improvements, with special focus on the first internal evaluation measure for outlier detection [6].

The second invited presentation by Kate Smith-Miles, University of Melbourne, was on “*Instance Spaces for Objective Assessment of Algorithms and Benchmark Test Suites*”, describing attempts to characterize data sets in a way to allow a map of the landscape of varying problems that shows where which algorithms perform good and this way also to identify areas where no good algorithm is available. This approach has been applied to characterize optimization problems [7] and classification problems [8]. It would be interesting to see this also on unsupervised learning problems.

The third invited presentation by Bart Goethals, University of Antwerp, reported on “*Lessons learned from the FIMI workshops*”, a series of workshops that Bart run with others roughly 15 years ago, focusing on the runtime behavior of algorithms for frequent pattern mining [4, 2]. Bart highlighted the various problems encountered in these attempts, for example the difficulty in assessing truly algorithmic merits as opposed to implementation details.

The fourth invited presentation by Miloš Radovanović, University of Novi Sad, reported on observations regarding “*Clustering Evaluation in High-Dimensional Data*” and an apparent bias that is shown by some evaluation indices w.r.t. the dimensionality of the data [10].

3.2 Contributed Papers The submitted papers discussed a variety of problems around the topic of the workshop.

In “*EvalNE: A Framework for Evaluating Network Embeddings on Link Prediction*”, Alexandru Mara, Jeffrey Lijffijt, and Tijl De Bie describe an evaluation framework for benchmarking existing and potentially new algorithms in the targeted area, motivated by a observed lack of reproducibility.

Martin Aumüller and Matteo Ceccarello contributed a study on “*Benchmarking Nearest Neighbor Search: Influence of Local Intrinsic Dimensionality and Result Diversity in Real-World Datasets*”, in which they study the influence of intrinsic dimensionality on the performance of approximate nearest neighbor search.

In their contribution “*Context-Driven Data Mining through Bias Removal and Incompleteness Mitigation*”, Feras Batarseh and Ajay Kulkarni describe case studies for the use of context to overcome obstacles based on data quality (or a lack thereof) and thereby to improve the quality achieved in the corresponding data mining application.

Based on the instance space analysis techniques for optimization and for classification problems as discussed earlier in the invited presentation by Kate Smith-Miles, in “*Instance space analysis for unsupervised outlier detection*” Sevvandi Kandanaarachchi, Mario Munoz and Kate Smith-Miles discuss an approach to extend these techniques to the unsupervised and therefore more challenging problem of outlier detection.

The contribution “*Characterizing Transactional Databases for Frequent Itemset Mining*” by Christian Lezcano and Marta Arias proposes a list of metrics to capture representativeness and diversity of benchmark datasets for frequent itemset mining.

3.3 Program Committee The workshop would not have been possible without the generous help and the time and effort put into reviewing submissions by

- Martin Aumüller, IT University of Copenhagen
- James Bailey, University of Melbourne
- Roberto Bayardo, Google
- Christian Borgelt, University of Salzburg
- Ricardo J. G. B. Campello, University of Newcastle
- Sarah Cohen-Boulakia, Université Paris-Sud
- Ryan R. Curtin, Symantec Corporation
- Tijl De Bie, University of Gent
- Marcus Edel, Freie Universität Berlin
- Bart Goethals, University of Antwerp
- Markus Goldstein, Hochschule Ulm
- Nathalie Japkowicz, American University
- Daniel Lemire, University of Quebec
- Philippe Lenca, IMT Atlantique
- Helmut Neukirchen, University of Iceland
- Jürgen Pfeffer, Technical University Munich
- Miloš Radovanović, University of Novi Sad
- Protiva Rahman, Ohio State University
- Mohak Shah, LG Electronics
- Kate Smith-Miles, University of Melbourne
- Joaquin Vanschoren, Eindhoven University of Technology
- Ricardo Vilalta, University of Houston
- Mohammed Zaki, Rensselaer Polytechnic Institute

4 Conclusions

To summarize, the submitted papers as well as the discussion had a main focus on unsupervised evaluation. But we also touched other topics and agreed that the richness of topics and questions is asking for a continuation to a workshop series. Some main points of the discussion were:

- Dataset complexity is important. So far, the community mainly focused on building more complex methods, however evaluating existing and new methods on appropriate benchmarks reflecting the real world complexity is necessary for scientific advance.
- In general, awareness of reviewers should be raised regarding evaluation aspects, full-range evaluation, reproducibility, embracing negative results etc.

These aspects are important for the furthering of maturity of data mining as a scientific effort. However, it seems still very hard to publish papers concerning issues around evaluation in main stream venues. We need a critical mass to change the current status quo.

Evaluation is a huge domain and only few aspects have been covered at EDML 2019. Data-related issues like sample representativeness, redundancy, bias, non-stationary data etc. have not been discussed. From a learning method perspective, it would be also interesting to investigate similar questions in the context of deep neural networks, that are currently dominating the research in the data mining/machine learning areas. These are possible candidate focus areas for future workshops. We plan to continue EDML as a series.

Finally, we wish to express our appreciation of the presented work as well as of interest and vivid participation of the audience.

References

- [1] D. Basaran, E. Ntoutsis, and A. Zimek. Redundancies in data and their effect on the evaluation of recommendation systems: A case study on the amazon reviews datasets. In *SDM*, pages 390–398. SIAM, 2017.
- [2] R. J. Bayardo Jr., B. Goethals, and M. J. Zaki, editors. *FIMI '04, Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations, Brighton, UK, November 1, 2004*, volume 126 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2005.
- [3] G. O. Campos, A. Zimek, J. Sander, R. J. G. B. Campello, B. Micenkova, E. Schubert, I. Assent, and M. E. Houle. On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study. *Data Min. Knowl. Discov.*, 30(4):891–927, 2016.
- [4] B. Goethals and M. J. Zaki, editors. *FIMI '03, Frequent Itemset Mining Implementations, Proceedings of the ICDM 2003 Workshop on Frequent Itemset Mining Implementations, 19 December 2003, Melbourne, Florida, USA*, volume 90 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2003.
- [5] H. Kriegel, E. Schubert, and A. Zimek. The (black) art of runtime evaluation: Are we comparing algorithms or implementations? *Knowl. Inf. Syst.*, 52(2):341–378, 2017.
- [6] H. O. Marques, R. J. G. B. Campello, A. Zimek, and J. Sander. On the internal evaluation of unsupervised outlier detection. In *SSDBM*, pages 7:1–7:12. ACM, 2015.
- [7] M. A. Muñoz and K. A. Smith-Miles. Performance analysis of continuous black-box optimization algorithms via footprints in instance space. *Evolutionary Computation*, 25(4), 2017.
- [8] M. A. Muñoz, L. Villanova, D. Baatar, and K. Smith-Miles. Instance spaces for machine learning classification. *Machine Learning*, 107(1):109–147, 2018.
- [9] D. Sidlauskas and C. S. Jensen. Spatial joins in main memory: Implementation matters! *PVLDB*, 8(1):97–100, 2014.
- [10] N. Tomašev and M. Radovanović. Clustering evaluation in high-dimensional data. In M. E. Celebi and K. Aydin, editors, *Unsupervised Learning Algorithms*. Springer, 2016.
- [11] Z. Zheng, R. Kohavi, and L. Mason. Real world performance of association rule algorithms. In *KDD*, pages 401–406. ACM, 2001.
- [12] A. Zimmermann. The data problem in data mining. *SIGKDD Explorations*, 16(2):38–45, 2014.