

Beyond Adaptation: Understanding Distributional Changes

Edited by

Georg Kreml^{*1}, Vera Hofer², Geoffrey Webb³, and
Eyke Hüllermeier⁴

- 1 Algorithmic Data Analysis, Department of Information and Computing Sciences, Utrecht University, The Netherlands g.m.kreml@uu.nl
- 2 Department of Statistics and Operations Research, Karl-Franzens-University Graz, Austria vera.hofer@uni-graz.at
- 3 Data Science, Monash University, Australia geoff.webb@monash.edu
- 4 Intelligent Systems and Machine Learning, Paderborn University, Germany eyke@upb.de

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 20372 “Beyond Adaptation: Understanding Distributional Changes”. It was centered around the aim to establish a better understanding of the causes, nature and consequences of distributional changes. Four key research questions were identified and discussed in during the seminar. These were the practical relevance of different scenarios and types of change, the modelling of change, the detection and measuring of change, and the adaptation to change.

The seminar brought together participants from several distinct communities in which parts of these questions are already studied, albeit in separate lines of research. These included data stream mining, where the focus is on concept drift detection and adaptation, transfer learning and domain adaptation in machine learning and algorithmic learning theory, change point detection in statistics, and the evolving and adaptive systems community. Therefore, this seminar contributed to stimulate research towards a thorough understanding of distributional changes.

Seminar September 6–11, 2020 – <http://www.dagstuhl.de/20372>

2012 ACM Subject Classification Theory of computation → Machine learning theory, Mathematics of computing → Time series analysis, Computing methodologies → Multi-task learning, Computing methodologies → Learning under covariate shift, Computing methodologies → Lifelong machine learning

Keywords and phrases Statistical Machine Learning, Data Streams, Concept Drift, Non-Stationary Non-IID Data, Change Mining, Dagstuhl Seminar

Digital Object Identifier 10.4230/DagRep.10.4.1

Edited in cooperation with Anastasiia Novikova[†], Maritime Graphics, Fraunhofer IGD, Rostock, Germany, anastasiia.novikova@igd-r.fraunhofer.de

* <https://orcid.org/0000-0002-4153-2594>

† <https://orcid.org/0000-0002-9724-8576>



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 3.0 Unported license

Beyond Adaptation: Understanding Distributional Changes, *Dagstuhl Reports*, Vol. 10, Issue 4, pp. 1–36

Editors: Georg Kreml, Vera Hofer, Geoffrey Webb, and Eyke Hüllermeier



DAGSTUHL
REPORTS

Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany


1 Executive Summary

Georg Kreml (Utrecht University, The Netherlands, g.m.kreml@uu.nl)

Vera Hofer (Graz University, Austria, vera.hofer@uni-graz.at)

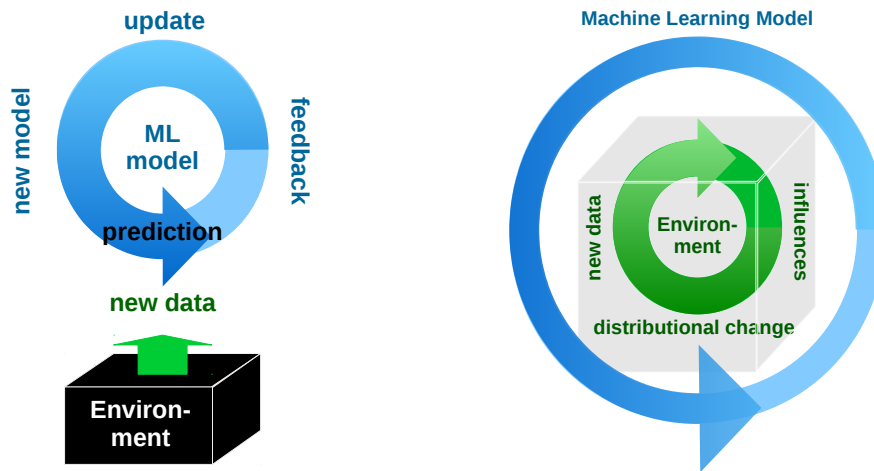
Eyke Hüllermeier (Paderborn University, Germany, eyke@upb.de)

Geoffrey I. Webb (Monash University, Australia, geoff.webb@monash.edu)

License  Creative Commons BY 3.0 Unported license

© Georg Kreml, Vera Hofer, Geoffrey Webb, Eyke Hüllermeier

The world is dynamically changing and non-stationary. This is reflected by the variety of methods that have been developed to detect changes and adapt to them. These contributions originate from various communities, including statistics, machine learning, data mining, and the evolving and adaptive systems community. Nevertheless, most of this research views the changing environment as a black-box data generator, to which models are adapted (Fig. 1).



■ **Figure 1** Black-Box Model Adaptation.

■ **Figure 2** Understanding Distributional Change.

The aim of this seminar was to put the focus on the distributional change itself, i.e., to make the process itself more transparent and a subject of research in its own right (Fig. 2). In its endeavour to understand causes, nature and consequences of distributional change, the seminar brought together researchers from communities in which related questions have already been studied, albeit in separate lines of research. These include *data stream mining*, *time series and sequence analysis*, *domain adaptation* and *transfer learning*, *subgroup discovery* and *emerging pattern mining*.

Data stream mining studies data that arrives either one-by-one or in batches over time, and where the data generating process is often non-stationary. This requires computationally efficient approaches that are capable to detect and adapt to distributional changes. In this literature, the latter are commonly denoted as *concept drift*, *population drift*, or *shift*. Related to this seminar are in particular the problems of identifying change or irregularities in data streams, such as *outlier detection* [1], *anomaly detection* [2], *change detection* [3], *change diagnosis* [4], *change mining* [5], *drift mining* [6], and *drift understanding* [7].

Time series analysis studies data observed over a time course typically exhibiting some time dependencies. The correspondence of distributional change in this literature are *distributional structural breaks* or *change points*. Thus, of particular interest are the problems of *statistical change point analysis*, see e.g. the books by [8, 9, 10] or recent survey

articles [12, 13]. A different line of research focuses on smallest detection delay for changes in sequentially observed data, see e.g. the recent books by [14, 15]. Of recent interest are also methods for the localization of multiple change points also known as *data segmentation* methodology, see e.g. the recent survey articles [16, 17, 18, 19, 20]. Of further interests is an *early classification of time series* [21].

Domain adaptation and **transfer learning** study the problem of transferring knowledge between domains or tasks. While there is not necessarily a temporal relationship between domains or tasks, distributional differences between domains are studied under the notion of *dataset shift*. Related problems of particular interest are *lifelong learning* [22] and *unsupervised domain adaptation* [23].

Subgroup discovery studies the problem of finding subgroups that show an unusual distribution for a target variable. There is not necessarily a temporal relationship between subgroup. Of particular interest is *exceptional model mining*, which studies the problem of finding subgroups, where a model fitted to that subgroup is somehow exceptional [24]. Another related area is *emerging pattern mining* [25] for identifying emerging trends in time-stamped databases.

Topics Discussed in the Seminar

The seminar identified several key research questions around understanding distributional changes:

1. Understanding the practical relevance of different scenarios and types of change.
2. How to model such types of change effectively.
3. How to detect, verify, and measure types of change.
4. How to effectively adapt prediction models to the different types of change.
5. How to establish bounds for distributional change, or for predictive performance under change.
6. How to visualise change, and how to highlight individual types of change.(interactively).
7. How to evaluate techniques for the above questions.

Due to the limited time, discussion has focused mostly on the first four research questions, with plans to address the remaining questions in a follow-up seminar.

Program Overview

This one-week seminar was structured such that plenary sessions formed a frame around parallel break-out group sessions. It was opened with plenary sessions on Monday and Tuesday morning, where four tutorial served to establish a common vocabulary and understanding between the participants from the different communities. In the subsequent four half-days, 13 spotlight talks were organised, each followed by discussions in break-out groups, and each closing by a short bring-back plenary session. The seminar closed by two plenary sessions on Friday morning, where action plans for further steps on research and collaboration were discussed.

Outcomes

As detailed in the description of the sessions below, and in particular for the plenary session, differences in the terminology, concepts and common assumptions used in the different communities were identified as an important challenge towards common understanding of distributional changes. Therefore, a potential follow-up collaboration will focus on a joint publication that provides a mapping of terms and concepts. In particular, it should work out the notion of change (and representation) in data streams and time series, as well as in domain adaptation with multiple temporally connected source domains.

References

- 1 Shiblee Sadik and Le Gruenwald. Research issues in outlier detection for data streams. *ACM SIGKDD Explorations Newsletter*, 15(1):33–40, 2014.
- 2 Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.
- 3 Dang-Hoan Tran, Mohamed Medhat Gaber, and Kai-Uwe Sattler. Change detection in streaming data in the era of big data: models and issues. *ACM SIGKDD Explorations Newsletter*, 16(1):30–38, 2014.
- 4 Charu C. Aggarwal. On change diagnosis in evolving data streams. *IEEE Transactions on Knowledge and Data Engineering*, 17(5):587–600, 2005.
- 5 Mirko Böttcher, Frank Höppner, and Myra Spiliopoulou. On exploiting the power of time in data mining. *ACM SIGKDD Explorations Newsletter*, 10(2):3–11, 2008.
- 6 Vera Hofer and Georg Kreml. Drift mining in data: A framework for addressing drift in classification. *Computational Statistics and Data Analysis*, 57(1):377–391, 2013.
- 7 Geoffrey Webb, Loong Kuan Lee, Bart Goethals, and Francois Petitjean. Analyzing concept drift and shift from sample data. *Data Mining and Knowledge Discovery*, 32(5):1179–1199, 2018.
- 8 Miklós Csörgö and Lajos Horváth. *Limit Theorems in Change-point Analysis*, volume 18. John Wiley & Sons Inc, 1997.
- 9 Jie Chen and Arjun K Gupta. *Parametric statistical change point analysis: with applications to genetics, medicine, and finance*. Springer Science & Business Media, 2nd edition, 2011.
- 10 B. E. Brodsky and B. S. Darkhovsky. *Nonparametric Methods in Change-point Problems*. Springer, 1993.
- 11 John AD Aston and Claudia Kirch. *Detecting and estimating changes in dependent functional data*. *Journal of Multivariate Analysis*, 109:204–220, 2012.
- 12 Alexander Aue and Lajos Horváth. Structural breaks in time series. *Journal of Time Series Analysis*, 34:1–16, 2013.
- 13 Lajos Horváth and Gregory Rice. Extensions of some classical methods in change point analysis. *TEST*, 23:1–37, 2014.
- 14 Alexander Tartakovsky, Igor Nikiforov, and Michele Basseville. *Sequential analysis: Hypothesis testing and changepoint detection*. CRC Press, 2014.
- 15 Alexander Tartakovsky. *Sequential Change Detection and Hypothesis Testing: General Non-iid Stochastic Models and Asymptotically Optimal Rules*. CRC Press, 2019.
- 16 Idris A Eckley, Paul Fearnhead, and Rebecca Killick. Analysis of changepoint models. In David Barber, A. Taylan Cemgil, and Silvia Chiappa, editors, *Bayesian Time Series Models*, chapter 10, pages 205–224. Cambridge University Press, Cambridge, 2011.
- 17 Samaneh Aminikhanghahi and Diane J Cook. A survey of methods for time series change point detection. *Knowledge and Information Systems*, 51:339–367, 2017.
- 18 Haeran Cho and Claudia Kirch. Data segmentation algorithms: Univariate mean change and beyond. *arXiv preprint arXiv:2012.12814*, 2020.

- 19 Paul Fearnhead and Guillem Rigau. Relating and comparing methods for detecting changes in mean. *Stat*, page e291, 2020.
- 20 Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020.
- 21 Zhengzheng Xing, Jian Pei, and Eamonn Keogh. A brief survey on sequence classification. *ACM Sigkdd Explorations Newsletter*, 12(1):40–48, 2010.
- 22 Anastasia Pentina and Ruth Uerner. Lifelong learning with weighted majority votes. In *Neural Information Processing Systems*, volume 29, pages 3619–3627, 2016.
- 23 Shai Ben-David and Ruth Uerner. On the hardness of domain adaptation and the utility of unlabeled target samples. In *International Conference on Algorithmic Learning Theory*, pages 139–153, 10 2012.
- 24 Wouter Duivesteijn and Julia Thaele. Understanding where your classifier does (not) work. In Albert Bifet, Michael May, Bianca Zadrozny, Ricard Gavaldà, Dino Pedreschi, Francesco Bonchi, Jaime Cardoso, and Myra Spiliopoulou, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 250–253, Cham, 2015. Springer International Publishing.
- 25 Petra Kralj Novak, Nada Lavrač, and Geoffrey I Webb. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, 10(2), 2009.

2 Table of Contents

Executive Summary

<i>Georg Krempl, Vera Hofer, Geoffrey Webb, Eyke Hüllermeier</i>	2
--	---

Tutorial Talks

Data Stream Mining and Concept Drift Adaptation <i>Mykola Pechenizkiy</i>	8
An Overview on Domain Adaptation and Modelling of Dataset Shift <i>Ruth Urner, Shai Ben-David</i>	8
Beyond Time Series Stationarity: Smooth and Abrupt Changes <i>Claudia Kirch</i>	9
Novelty Detection <i>Pavlo Mozharovskyi</i>	10

Spotlight Talks and Breakout Group Discussions


Novelty, Change, and Evaluation <i>João Gama</i>	12
Monitoring and Forecasting Changes in Feature Distributions Over Time <i>Mark Last</i>	13
Multi-Target Prediction on Data Streams <i>Sašo Džeroski</i>	15
Temporal Density Extrapolation <i>Vera Hofer</i>	16
Challenges of Applying Concept Drift Detection in Real-World Applications <i>Yun Sing Koh</i>	17
Understanding Concept Drift <i>Loong Kuan Lee</i>	19
Learning Under Concept Drift With Zero Ground Truth <i>Indrė Žliobaitė</i>	21
Time Series Classification <i>Geoffrey I. Webb</i>	22
Uncertainty in Labeling – What Can We Learn from Experiments? <i>Myra Spiliopoulou</i>	24
Recovery Analysis for Adaptive Learning from Non-stationary Data Streams <i>Eyke Hüllermeier</i>	25
Online Linear Discriminant Analysis for Data Streams with Concept Drift <i>Sarah Schnackenberg</i>	27
Classification, Calibration, and Quantification: A Study of Dataset Shift <i>Dirk Tasche</i>	28
Prediction-Dependent Drift <i>Georg Krempl</i>	29

Labelless Detection and Explanation of Concept Drift <i>Mykola Pechenizkiy</i>	32
Plenary Discussion	
Plenary Discussion <i>All participants of this seminar.</i>	33
Acknowledgements	35
Remote Participants	36
Participants	36

3 Tutorial Talks

3.1 Data Stream Mining and Concept Drift Adaptation

Mykola Pechenizkiy (TU Eindhoven – Eindhoven, The Netherlands, m.pechenizkiy@tue.nl)

License  Creative Commons BY 3.0 Unported license
 © Mykola Pechenizkiy

In the real world data often arrives in streams and evolves over time. Concept drift in supervised learning means that the relation between the input data and the target variable changes. Therefore, in many real-world applications the learning models need to adapt to the anticipated changes. In this tutorial we provide an introduction to the area of concept drift in data mining and machine learning research. First, we characterize the adaptive learning process, categorize existing strategies for (reactive) handling of concept drift in the most assumed setting – unpredictable changes happen in hidden contexts that are not observable to the adaptive learning system. Then, we consider other operational settings that commonly occur in practice, but have been underexplored in academia. In particular, we provide motivation for approaches that can handle concept drift proactively, and do not require (immediate) knowledge of the ground truth labels.


References

- 1 Indrè Žliobaitė, Mykola Pechenizkiy, and Joao Gama. An overview of concept drift applications. In Nathalie Japkowicz and Jerzy Stefanowski, editors, *Big Data Analysis: New Algorithms for a New Society*, page 91–114. Springer, 2016.
- 2 João Gama, Indrè Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):1–44, 2014.
- 3 Geoffrey I Webb, Roy Hyde, Hong Cao, Hai Long Nguyen, and Francois Petitjean. Characterizing concept drift. *Data Mining and Knowledge Discovery*, 30(4):964–994, 2016.
- 4 Geoffrey Webb, Loong Kuan Lee, Bart Goethals, and Francois Petitjean. Analyzing concept drift and shift from sample data. *Data Mining and Knowledge Discovery*, 32(5):1179–1199, 2018.
- 5 Igor Goldenberg and Geoffrey I Webb. *Survey of distance measures for quantifying concept drift and shift in numeric data*. Knowledge and Information Systems, pages 1–25, 2018.
- 6 Igor Goldenberg and Geoffrey Webb. PCA-based drift and shift quantification framework for multidimensional data. Knowledge and Information Systems, 62:2835–2854, 07 2020

3.2 An Overview on Domain Adaptation and Modelling of Dataset Shift

Ruth Urner (York University – Toronto, Canada, ruth@eecs.yorku.ca)

Shai Ben-David (University of Waterloo, – Waterloo, Canada, shai@cs.uwaterloo.ca)

License  Creative Commons BY 3.0 Unported license
 © Ruth Urner, Shai Ben-David
 Video <http://videlectures.net/DagstuhlSeminar2020/>
 Videorecording of the tutorial by Ruth Urner, Shai Ben-David

Data shift is a common problem in machine learning. It arises when the data generating process at test time differs from the data that the model was trained on. Transfer learning is an umbrella name for tools aiming to address data shift by utilizing the data from the training phase rather than addressing the test-time environment from scratch. In this tutorial we address some of the basic challenges, algorithmic solution tools and inherent limitations

of transfer learning. All of these aspects vary with the setup (from domain adaptation, when the training and target data distributions are both fixed, to lifelong learning, when changes in the data keep occurring and need to be addressed continuously) and with the type of prior knowledge, or assumptions, that the learner relies on.

The tutorial describes several learning paradigms, such as model distillation, importance reweighting of training data, embedding the different data sources into a joint feature space in which they can be viewed as similar, active learning and “meta-learning” – learning invariances of that help speed up learning of yet unseen problems.

We also address the problem of detecting when a data shift occurs and understanding the nature of such a shift.


An important lesson from the theoretical analysis is that there are inherent limitations to what transfer learning can achieve – some No Free Lunch theorems that underscore the prerequisites of strong prior knowledge about the nature of the shift before one can provide any performance guarantees.

References

- 1 Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Vaughan. A theory of learning from different domains. in *Machine Learning*, 79:151–175, 2010.
- 2 Shai Ben-David, Tyler Lu, Teresa Luu, and Dávid Pál. Impossibility theorems for domain adaptation. in *Journal of Machine Learning Research – Proceedings Track*, 9:129–136, 01 2010.
- 3 Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *NIPS*, volume 19, pages 137–144, 01 2006.
- 4 Shai Ben-David and Ruth Urner. On the hardness of domain adaptation and the utility of unlabeled target samples. In *International Conference on Algorithmic Learning Theory*, pages 139–153, 10 2012.
- 5 C. Berling and R. Urner. Active nearest neighbors in changing environments. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of JMLR Workshop and Conference Proceedings, page 1870–1879. JMLR, 2015.
- 6 Tongtong Fang, Nan Lu, Gang Niu, and Masashi Sugiyama. Rethinking importance weighting for deep learning under distribution shift. arxiv, 12 2020. arXiv preprint:2006.04662.
- 7 Anastasia Pentina and Ruth Urner. Lifelong learning with weighted majority votes. In *Neural Information Processing Systems*, volume 29, pages 3619–3627, 2016.
- 8 Shai Ben-David and Ruth Urner. Domain adaptation—can quantity compensate for quality? In *Annals of Mathematics and Artificial Intelligence*, 70(3):185–202, 2014

3.3 Beyond Time Series Stationarity: Smooth and Abrupt Changes

Claudia Kirch (Magdeburg University – Magdeburg, Germany, claudia.kirch@ovgu.de)

License  Creative Commons BY 3.0 Unported license
© Claudia Kirch

Video <https://www2.math.uni-magdeburg.de/owncloud/index.php/s/CRpk4Jf5buoCsfl>
Videorecording of the tutorial by Claudia Kirch

In many applications data are collected during their time course where it can no longer be assumed that today's observations are independent from yesterday's. Because these dependencies have to be taken into account for any meaningful statistical analysis, the field of time series analysis aims at investigating, modelling and mathematically analysing them.

This is particularly challenging in nonparametric i.e. model-free statistics, where likelihood approximations can be used in bootstrapping and Bayesian analysis – both methods from computational statistics that aim at quantifying uncertainty. This includes so-called locally stationary time series that are not stationary but exhibit a slowly changing structure which can be seen as approximately stationary locally, i.e. in a small environment of every point in time.

Of particular interest are non-stationarities caused by structural breaks in the data generating mechanism, so called *change points*. The detection and localisation of such change points has a long tradition in time series analysis and statistics. Classical theory deals with the *detection* of at most one change point in a fully observed data set. Even in this situation methodology for more complex situations beyond univariate mean changes such as distributional changes in the time series structure (modelled e.g. via neural networks) or high-dimensional data sets – both functional and panel data – are of recent interest. Another recent line of research deals with the estimation of multiple change points which is also known as *data segmentation* problem in the literature. Finally, sequential or online methods are shortly discussed where new data arrives steadily and after each new observation a decision has to be made whether or not a change has occurred. The mathematical tools required for this kind of statistical analysis are indeed very different from the classical a-posteriori or offline change point detection. The aim of this tutorial was to give a short overview/introduction into the above topics.

References

- 1 John AD Aston and Claudia Kirch. Detecting and estimating changes in dependent functional data. in *ournal of Multivariate Analysis*, 109:204–220, 2012.
- 2 Haeran Cho and Claudia Kirch. Data segmentation algorithms: Univariate mean change and beyond. arXiv preprint arXiv:2012.12814, 2020.
- 3 Franziska Häfner and Claudia Kirch. Moving fourier analysis for locally stationary processes with the bootstrap in view. In *Journal of Time Series Analysis*, 38(6):895–922, 2017.
- 4 Claudia Kirch and J Tadjuidje Kamgaing. Detection of change points in discrete valued time series. In *Handbook of Discrete Valued Time series*. Springer Berlin Heidelberg, 2014.
- 5 Claudia Kirch and Joseph Tadjuidje Kamgaing. On the use of estimating functions in monitoring time series for change points. In *Journal of Statistical Planning and Inference*, 161:25–49, 2015.
- 6 Claudia Kirch, Matthew C Edwards, Alexander Meier, Renate Meyer, et al. Beyond whittle: Nonparametric correction of a parametric likelihood with a focus on bayesian time series analysis. In *Bayesian Analysis*, 14(4):1037–1073, 2019.
- 7 Alexander Meier, Claudia Kirch, and Haeran Cho. mosum: A package for moving sums in change point analysis. In *Journal of Statistical Software* (to appear), 2019.

3.4 Novelty Detection

Pavlo Mozharovskiy (Télécom Paris – Paris, France, pavlo.mozharovskiy@telecom-paris.fr)

License © Creative Commons BY 3.0 Unported license

© Pavlo Mozharovskiy

Video <http://videlectures.net/DagstuhlSeminar2020/>

Videorecording of the tutorial by Pavlo Mozharovskiy

Novelty detection [1] is a branch of machine learning which aims at identifying single or grouped observations that exhibit behavior unknown during the model training. Be it measurement errors, disease development, severe weather, production quality default(s)

(items) or failed equipment, financial frauds or crisis events, their on-time identification, isolation and explanation constitute an important task in almost any branch of industry and science. In this tutorial, we will focus on the task of identification of these novel events, the task closely related to anomaly detection [2]. Roughly speaking, the state-of-the-art on novelty detection can be split into two categories: statistical methods and neural-network based approaches. Here, we address statistical multivariate and functional novelty detection.

When the data are presented in a form of a table that contains properties of individuals (a typical structure of a data base), multivariate novelty detection methods should be employed. As a first go, we address three non-parametric methods, which can also be seen as extensions of the existing classification methodology. *One-class support vector machines* [4] aim at detecting a minimal set excluding abnormal observations. *Local outlier factor* [3] uses neighborhood relation to decide whether an observation or a (small) cluster of them is distant from the majority of the data. *Isolation forest* [5] employs random (one-dimensional) cuts to separate far-lying points faster than those in the dense data regions.

Among non-parametric methods, *data depth* [6] occupies today a special place. Given an observation, it measures how typical (or deep) this observation is with respect to other available observations of the same nature. Multivariate data depth possesses such attractive properties as robustness and affine invariance. In the current, we discuss the concept of data depth in the multivariate settings, review most common notion of the depth, and address the question of identifying novel observations by means of the central depth regions.

If the data are functions of an argument, *e.g.*, time (such as time series), projection on a multivariate sub-basis and then applying a multivariate technique or functional novelty detection methods [7] can be in use. Here, we consider two approaches from this last family: *Integrated functional data depth* generalizes the multivariate depth to functional spaces using averaging, and allows for identification of novel functional observations. While for certain functional depth notions identification of isolated novelties can constitute problems, these can be dealt with when using *functional isolation forest* with a proper dictionary.

Practical part of this tutorial exemplifies identification of novelties for simulated and real-world multivariate and functional data with codes provided in both R and Python.


References

- 1 Marco Pimentel, David Clifton, Lei Clifton, and L. Tarassenko. A review of novelty detection. *Signal Processing*, 99:215–249, 06 2014.
- 2 Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.
- 3 Markus Breunig, Hans-Peter Kriegel, Raymond Ng, and Joerg Sander. Lof: Identifying density-based local outliers. In *Proc. of the 2000 ACM SIGMOD int. conf. on Management of data*, volume 29, pages 93–104, 06 2000.
- 4 Bernhard Schölkopf, John Platt, John Shawe-Taylor, Alexander Smola, and Robert Williamson. Estimating support of a high-dimensional distribution. *Neural Computation*, 13:1443–1471, 07 2001.
- 5 F. T. Liu, K. M. Ting, and Z. Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422, 2008.
- 6 Karl Mosler. Depth statistics. *Robustness and Complex Data Structures: Festschrift in Honour of Ursula Gather*, pages 17–34, 07 2012.
- 7 Mia Hubert, Peter Rousseeuw, and Pieter Segaeert. Multivariate functional outlier detection. *Statistical Methods and Applications*, 24:177–202, 07 2015.

4 Spotlight Talks and Breakout Group Discussions

4.1 Novelty, Change, and Evaluation

João Gama (University of Porto – Porto, Portugal, jgama@fep.up.pt)

License  Creative Commons BY 3.0 Unported license
© João Gama

Main reference Elaine R. Faria, Isabel J. C. R. Gonçalves, André C. P. L. F. de Carvalho, João Gama: “Novelty detection in data streams”, *Artif. Intell. Rev.*, Vol. 45(2), pp. 235–269, 2016.

URL <http://dx.doi.org/10.1007/s10462-015-9444-8>

4.1.1 Abstract of Spotlight Presentation

Novelty Detection (ND) refers to the automatic identification of unforeseen phenomena embed in a large amount of normal data. The ND task consists of training a model from a training set with examples from a small subset of the possible classes. This model is used to classify test examples, where examples from new classes can appear. Novelty detection makes it possible to recognize novel profiles (concepts) in unlabelled data, which may indicate the appearance of a new concept, a change that occurred in known concepts, or the presence of noise. The discovery of new concepts has increasingly attracted the attention of the knowledge discovery community, usually under the terms of novelty detection [2] or open set recognition [1]. The terms one-class classification [7], and anomaly detection [6] are also frequently used. Most ND algorithms [4] work in two phases. The first phase is offline. Algorithms learn from labeled data a characteristic model for each class. The second phase is online. The current decision model analyses each unlabelled example from the stream. If the model covers the example, it is classified in one of the known classes; otherwise, it is classified as *unknown* and stored in a short term memory. From time to time, the unlabelled examples stored in the short memory are analyzed to identify dense regions in the instance space. These regions are considered that correspond to novel concepts that emerged from the test data.


Note: This spotlight talk has been discussed within the plenary sessions, see Section 5 for the results of these discussions.

References

- 1 Chuanxing Geng, Sheng-Jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–1, 2020.
- 2 Elaine R. Faria, Isabel J. C. R. Gonçalves, André C. P. L. F. de Carvalho, and João Gama. Novelty detection in data streams. *Artif. Intell. Rev.*, 45(2):235–269, 2016.
- 3 Elaine Faria, Isabel Goncalves, Joao Gama, and Andre de Carvalho. Evaluation of multiclass novelty detection algorithms for data streams. In *IEEE Transactions on Knowledge and Data Engineering*, 27:1–1, 11 2015.
- 4 Elaine Faria, Andre de Carvalho, and Joao Gama. Minas: multiclass learning algorithm for novelty detection in data streams. *Data Mining and Knowledge Discovery*, 30, 08 2015.
- 5 Mohammad Masud, Jing Gao, Latifur Khan, Jiawei Han, and Bhavani M. Thuraisingham. Classification and novel class detection in concept-drifting data streams under time constraints. In *IEEE Trans. on Knowl. and Data Eng.*, 23(6):859–874, June 2011.
- 6 Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.
- 7 David Martinus Johannes Tax. *One-class classification: Concept learning in the absence of counter-examples*. PhD thesis, Technische Universiteit Delft, 2001.

4.2 Monitoring and Forecasting Changes in Feature Distributions Over Time

Mark Last (Ben-Gurion University of the Negev – Be'er Sheva, Israel, mlast@bgu.ac.il)

License  Creative Commons BY 3.0 Unported license
© Mark Last

4.2.1 Abstract of Spotlight Presentation

Feature ranking and selection reduces the data acquisition and storage requirements of machine learning algorithms, decreases training and inference times, improves the accuracy of the induced models, and facilitates their interpretability. However, in a dynamic data stream, feature importance may change over time, either gradually or abruptly. To address this issue, we need to continuously monitor and forecast features distribution and their effect on model classification performance. In case of a concept drift, online learning algorithms [1] replace some of the previously selected features with more relevant ones and update the model accordingly. However, the online learning algorithms do not monitor feature distributions over time. Online Feature Selection (OFS) allows to dynamically rank features with respect to a specific classifier that uses a small feature subset of a fixed size [2]. In [3], Heterogeneous Ensemble with Feature Drift for Data Streams integrates traditional feature selection into an ensemble and adopts a modification of the Fast Correlation-Based Filter (FCBF) algorithm so it dynamically updates the selected relevant feature subset of a data stream. Adaptive Boosting for Feature Selection (ABFS) [4] uses a combination of boosting and decision stumps in order to select features. Feature drift is detected by monitoring the error distribution of each decision stump. In [5], the authors proposed an unsupervised approach for feature ranking and selection. It is based on constructing and maintaining a sketch matrix that shrinks the original data in orthogonal vectors. The feature importance score is calculated by regression analysis, where the spectral embedding of the dataset is used as the dependent variable. So far, most OFS algorithms have been evaluated only on stationary data streams, where the values of all instance features are assumed to arrive together. In dynamic data streams with partially available feature values and class labels, specific monitoring objectives may include:

- Explain feature and concept drifts
- Improve data and model quality
- Save data collection and storage efforts
- Enhance the efficiency and effectiveness of online learning methods
- Handle delayed labeling
- Predict future changes in feature distribution and ranking

References

- 1 Mark Last. Online classification of nonstationary data streams. *Intelligent data analysis*, 6(2):129–147, 2002.
- 2 Steven CH Hoi, Jialei Wang, Peilin Zhao, and Rong Jin. Online feature selection for mining big data. In *Proceedings of the 1st international workshop on big data, streams and heterogeneous source mining: Algorithms, systems, programming models and applications*, pages 93–100, 2012.
- 3 Jan N van Rijn, Geoffrey Holmes, Bernhard Pfahringer, and Joaquin Vanschoren. The online performance estimation framework: heterogeneous ensemble learning for data streams. In *Machine Learning*, 107(1):149–176, 2018.

- 4 Jean Paul Barddal, Fabrício Enembreck, Heitor Murilo Gomes, Albert Bifet, and Bernhard Pfahringer. Boosting decision stumps for dynamic feature selection on data streams. *Information Systems*, 83:13–29, 2019.
- 5 Hao Huang, Shinjae Yoo, and Shiva Prasad Kasiviswanathan. Unsupervised feature selection on data streams. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1031–1040, 2015.

4.2.2 Results of the Working Group Discussion

Discussion Participants

- Amir Abolfazli, University of Hannover
- Shai Ben-David, University of Waterloo
- Georg Kreml, Utrecht University
- Mark Last, Ben-Gurion University of the Negev
- Myra Spiliopoulou, Magdeburg University
- Jerzy Stefanowski, Poznan University of Technology
- Dirk Tasche, Swiss Financial Market Supervisory Authority FINMA
- Andreas Theissler, Hochschule Aalen

4.2.2.1 Identified Research Gaps

Recent works in the task of data stream classification have considered concept drift as a change in the likelihood of a feature for a class and also a change in the feature space. However, a definition taking into account the above-mentioned changes along with a change in the joint distribution of features is missing.

4.2.2.2 Formal Definitions

We need mathematical definitions of several basic concepts, such as *feature importance* and *feature relevance* as a function of time. These definitions may extend the existing definitions in the static framework. We also need to define *feature drift* types with respect to individual features and feature interactions. Alphabet of (types of) drifts should be invariant over time. The relations of feature and concept drifts should also be considered.

4.2.2.3 Detecting Feature Drifts

Once the feature drift types are defined, we need algorithms for feature drift detection, probably including change point detection. Aspects to consider: measuring feature drifts in real-world settings, using sliding windows, choosing temporal resolution. We will probably need to optimize a sequence of two models: one detecting a feature drift, the other one classifying our input data. Unfortunately, availability of real-world data with known feature drifts as ground truth is still limited.

4.2.2.4 Explaining Feature Drifts

Users will not like the underlying models changing frequently, so explaining why models had to be changed/adapted due to feature drifts is highly important. Useful explanations may refer to the reason behind the change, spatio-temporal location of change in a data stream, effect on distribution, etc. We need objective measures of feature drifts explainability and interpretability. Occam’s razor: the users may prefer a model or a set of models that give the most consistent explanation over the longest timespan, but this model might be the one with

the most complex explanation. A relevant article on model and variable selection using the Minimum Description Length Principle: [12]. Feature drift explanation may be interesting but not necessarily actionable.

4.3 Multi-Target Prediction on Data Streams

Sašo Džeroski (Jozef Stefan Institute – Ljubljana, Slovenia, saso.dzeroski@ijs.si)

License © Creative Commons BY 3.0 Unported license
© Sašo Džeroski

Joint work of Aljaž Osojnik, Panče Panov, Sašo Džeroski

Main reference Aljaž Osojnik, Panče Panov, Sašo Džeroski: “Multi-label classification via multi-target regression on data streams”, *Mach. Learn.*, Vol. 106(6), pp. 745–770, 2017.

URL <http://dx.doi.org/10.1007/s10994-016-5613-5>

4.3.1 Abstract of Spotlight Presentation

Starting from tree-based regression methods for data streams, we have developed a number of approaches for on-line multi-target prediction. These cover different multi-target prediction tasks such as multi-target regression, multi-label classification and hierarchical versions of these tasks. These also cover a range of tree-based methods, including individual decision trees, option trees and tree ensembles (bagging and random forests). Finally, we have recently also addressed the task of semi-supervised multi-target prediction on data streams. We give a quick overview of these developments, based on on-line learning of predictive clustering trees, and discuss further research in this area (incl. change detection and feature ranking).

References

- 1 Aljaž Osojnik, Panče Panov, and Sašo Džeroski. Multi-label classification via multi-target regression on data streams. *Machine Learning*, 106(6):745–770, 2017.
- 2 Aljaž Osojnik, Panče Panov, and Sašo Džeroski. Tree-based methods for online multi-target regression. In *Journal of Intelligent Information Systems*, 50(2):315–339, 2018
- 3 Aljaž Osojnik, Panče Panov, and Sašo Džeroski. Incremental predictive clustering trees for online semi-supervised multi-target regression. In *Machine Learning*, 109(11):2121–2139, 2020.

4.3.2 Results of the Working Group Discussion

Discussion Participants

- Sašo Džeroski, Jozef Stefan Institute
- Johannes Fürnkranz, Johannes Kepler University, Linz
- Eyke Hüllermeier, Paderborn University
- Mykola Pechenizkiy, TU Eindhoven
- Arno Siebes, Utrecht University
- Jerzy Stefanowski, Poznan University of Technology

4.3.2.1 Problems Discussed

The group discussed several topics related to multi-target prediction (MTP) on data streams, including different degrees of supervision (e.g., fully supervised, semi-supervised and unsupervised learning; also learning with delayed supervision). A major topic of discussion was drift detection and adaptation, where different contexts of MTP on data streams were considered (different types of outputs, different degrees of supervision and different loss functions).

For example, decomposable (Hamming) and non-decomposable (subset 0/1) losses (say for multi-label classification, MLC) were discussed together with computational aspects of evaluating them in a data stream setting. Another topic discussed was the topic of clustering on data streams, where the tasks of clustering of data points and clustering of features (also called parallel data streams) can be considered, as well as the task of bi-clustering. Finally, the task of clustering in the presence of drift on data streams was identified as particularly relevant.


4.3.2.2 Conclusions

The interaction between change detection/adaptation in MTP on data streams and the (many) loss functions that can be considered was identified as particularly important. One can (and should) monitor the many performance measures available (e.g., in MLC), but then needs to decide when changes in the individual measures mean a change overall. Different options here include the detection of an overall change if any of the measures changes, on one hand, or only if all measures change, on the other hand.

Some of the measures may be too insensitive to change, such as Hamming loss in MLC. Others, such as subset 0/1 loss may be too sensitive. Luckily, other measures lay somewhere in-between on the spectrum of sensitivity (e.g., F1 score, Jacquard, ranking loss) and may be most suitable for use in practice.

4.4 Temporal Density Extrapolation

Vera Hofer (Graz University – Graz, Austria, vera.hofer@uni-graz.at)

License  Creative Commons BY 3.0 Unported license
© Vera Hofer

4.4.1 Abstract of Spotlight Presentation

In an evolving data stream data on a continuous feature X arrives sequentially (in chunks or instance by instance). The distribution of X is non stationary, i.e. the densities in feature space $f(x|t)$ can change over time. We address the question of estimating the density $f(x|t)$ of X at a future time point t given data on the development of X over time. Our model is based on a basis representation of the densities $f(x|t)$ at time t where normalised basis functions are given at fixed positions. The drift model is expressed by the time dependency of the coefficients in the basis representation. To guarantee that the weights satisfy the sum-to-1 constraint at any time point, a compositional data approach is applied. The weights are transformed according to an isometric-log-ratio transformation prior to the polynomial regression in time. The model is estimated by means of a weighted maximum likelihood approach where the weights adjust the model with respect to aging effects. The density forecast requires an extrapolation of weights, i.e. an extrapolation of ilr-transformed weights and a backtransformation into weights of basis functions. We found that the model performance depends on the nature of changes. In particular, the polynomial regression is weak in reacting to fast changes over time. As a remedy a drift model for the coefficients based on time series model may be considered.

References

- 1 Georg Kreml, Dominik Lang, and Vera Hofer. Temporal density extrapolation using a dynamic basis approach. In *Data Mining and Knowledge Discovery*, 33, Special Issue of the ECML/PKDD 2019 Journal Track(5):1323–1356, 2019.

4.4.2 Results of the Working Group Discussion

Discussion Participants

- Gerhard Gößler, Graz University
- Vera Hofer, Graz University
- Claudia Kirch, Magdeburg University
- Mykola Pechenizkiy, TU Eindhoven
- Sarah Schnackenberg, formerly TU Dortmund University
- Dirk Tasche, Swiss Financial Market Supervisory Authority FINMA
- Andreas Theissler, Hochschule Aalen

4.4.2.1 Alternatives to polynomial regression in drift model

A drift model based on a polynomial regression of the basis coefficients does not react fast enough to changes in a highly dynamic environment. It does not allow the detection of change points. Time series models seem to be a useful alternative to polynomial regression of the basis coefficients. Since a drift model for time series forecasting runs automatically/without user interaction and is robust, such a time series model for the basis coefficients of the densities need to fulfill certain requirements: Time series parameters should be found automatically, i.e. without tuning/selection by hand. The model needs to account for the dependency of the time series of the basis coefficients. A compositional data approach is required to guarantee that the sum-1-constraint is satisfied.

Instead of a time series approach which has certain disadvantages, alternative basis representations can be used. The Gaussian basis functions could be replaced by Bernstein polynomials. The basis coefficients could be estimated by a nonlinear change point regression model.

4.4.2.2 Conclusions

The drift model needs to run automatically/without user interaction and to be robust. Change point regression models will be highly appropriate since they can also detect jumps.

4.5 Challenges of Applying Concept Drift Detection in Real-World Applications

Yun Sing Koh (The University of Auckland – Auckland, New Zealand, ykoh@cs.auckland.ac.nz)

License  Creative Commons BY 3.0 Unported license
© Yun Sing Koh

4.5.1 Abstract of Spotlight Presentation

There are many examples of real-world stream learning applications, such as industrial process controls, air monitoring sensors, spam detection, fraud detection, medical sensor data, traffic monitoring. Many of these main applications of stream learning has produced a huge quantity of data continuously in real-time. There is a multitude of challenges when applying

concept drift detection in a real-world context. Many of these are well-known problems from missing data to the difficulty of obtaining labelled data. In this talk, I will detail some of the paper-cuts we have noticed that are slightly unusual. This includes real-world data stream applications that have complex combinations of many types of concept drift.

References

- 1 Hamish Huggard, Yun Sing Koh, Gillian Dobbie, and Edmond Zhang. Detecting concept drift in medical triage. Jimmy Huang, Yi Chang, Xueqi Cheng, Jaap Kamps, Vanessa Murdock, Ji-Rong Wen, and Yiqun Liu, editors, *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, SIGIR 2020, Virtual Event, China, July 25-30, 2020, pages 1733–1736. ACM, 2020.
- 2 Indrè Žliobaitė, Mykola Pechenizkiy, and Joao Gama. An overview of concept drift applications. In Nathalie Japkowicz and Jerzy Stefanowski, editors, *Big Data Analysis: New Algorithms for a New Society*, page 91–114. Springer, 2016.
- 3 Tegjot Singh Sethi and Mehmed Kantardzic. On the reliable detection of concept drift from streaming unlabeled data. In *Expert Systems with Applications*, 82:77–99, 2017.
- 4 Georg Kreml, Indrè Žliobaitė, Dariusz Brzezinski, Eyke Hüllermeier, Mark Last, Vincent Lemaire, Tino Noack, Ammar Shaker, Sonja Sievi, Myra Spiliopoulou, and Jerzy Stefanowski. Open challenges for data stream mining research. In *SIGKDD Explorations*, 16(1):1–10, 2014. Special Issue on Big Data.
- 5 Eyke Hüllermeier. Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *International Journal of Approximate Reasoning*, 55(7):1519–1534, 2014. Special issue: Harnessing the information contained in low-quality data sources.

4.5.2 Results of the Working Group Discussion

Discussion Participants

- Amir Abolfazli, University of Hannover
- Vera Hofer, Graz University
- Eyke Hüllermeier, Paderborn University
- Yun Sing Koh, The University of Auckland
- Myra Spiliopoulou, Magdeburg University
- Jerzy Stefanowski, Poznan University of Technology
- Dirk Tasche, Swiss Financial Market Supervisory Authority FINMA
- Andreas Theissler, Hochschule Aalen

4.5.2.1 Discussed Problem

Key question: How to reliably capture concept drift when model rebuilt is expensive? This question brought up further discussion/questions. This included discussion on whether we can infer the gain of retraining or estimating after retraining. Is there a possibility of locating a break-even point when adapting model versus retraining makes sense? When do we re-weight and adapt vs discarding a model in presence of drift? The base machine learning algorithm may affect the choice. For example, naïve Bayes may learn new information easier compared to Deep Learning which would be harder to forget new information. For example, the nature of the drift or shift might be considered. In the case of a hard shift, forgetting and discarding the model may be better than adaptation. Myra Spiliopoulou noted that gradual forgetting and gradual learning two problems, whereby gradual drift (knowing) when it is happening and adapting. Another point of discussion was “Can we still learn incrementally if there are novel classes?” There is a potential difference of target versus feature space sparsity.

The other discussion/question geared towards: “Can we trust the ground truth labels in the datasets?” Often correct labelling is not clear, even for the experts labelling the data. Further discussion by Eyke Hüllermeier on superset learning, allowing for “weak labelling” would be beneficial [5]. The other related research is in fuzzy set whether there is a plausible label or not. This includes generalisation the loss function of the learning with penalty scoring for optimistic superset loss.

4.5.2.2 Conclusions

The discussion open up further discussion of the area:

- Can tolerate the quality degradation of a model? Is there a possibility of locating a break even point when adapting model versus retraining makes sense?
- When do we re-weight and adapt vs discarding a model in presence of drift?
- Can we still learn incrementally if there are novel classes?
- Can we trust the ground truth labels in the datasets?

4.6 Understanding Concept Drift

Loong Kuan Lee (Monash University – Clayton, Australia, loong.kuan.lee@gmail.com)

License  Creative Commons BY 3.0 Unported license
© Loong Kuan Lee

4.6.1 Abstract of Spotlight Presentation

One way to better understand distributional changes is to figure out how large the difference between the 2 distributions before and after the change period is, also known as drift magnitude. The drift magnitude can be obtain by measuring the divergence between these 2 distributions.

However, in most cases, we do not have any information about these 2 distributions other than samples from these distributions. Furthermore, most divergences we can use to measure drift magnitude take time exponential to the number of variables in the distributions to compute. Therefore, we propose a method to estimate the divergence between the 2 distributions using sample data while avoiding this exponential time complexity w.r.t the number of variables. This crux of this method relies on the use of decomposable models to produce an estimate that decomposes the high-dimensional population distributions into products of lower dimensional distributions.

References

- 1 Alon Orlitsky, Narayana P. Santhanam, Krishnamurthy Viswanathan, and Junan Zhang. On modeling profiles instead of values. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, pages 426–435. AUAI Press, 2004.
- 2 François Petitjean and Geoffrey I. Webb. Scaling log-linear analysis to datasets with thousands of variables. In *Proceedings of the 2015 SIAM International Conference on Data Mining*, pages 469–477. Society for Industrial and Applied Mathematics, 2015.
- 3 Geoffrey I. Webb and François Petitjean. A Multiple Test Correction for Streams and Cascades of Statistical Hypothesis Tests. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD '16*, pages 1255–1264. ACM Press, 2016.

- 4 Geoffrey Webb, Loong Kuan Lee, Bart Goethals, and Francois Petitjean. Analyzing concept drift and shift from sample data. *Data Mining and Knowledge Discovery*, 32(5):1179–1199, 2018.
- 5 Mohammad S. Rahman and Gholamreza Haffari. A Statistically Efficient and Scalable Method for Exploratory Analysis of High-Dimensional Data. *SN Computer Science*, 1(2):64, 2020.

4.6.2 Results of the Working Group Discussion

Discussion Participants

- Vera Hofer, Graz University
- Georg Krempl, Utrecht University
- Loong Kuan Lee, Monash University
- Arno Siebes, Utrecht University
- Geoffrey I. Webb, Monash University

4.6.2.1 Estimation of high-dimensional distributions

In principle, any distribution that can be modelled by a graphical model can be modelled by a decomposable model. In the worst case this will lead to a saturated, or fully connected model, where all the variables in the model are connected to each other. However, in situations similar to this worst case scenario, where the treewidth, i.e. size of the largest maximal clique, of the learnt decomposable models are large, our method for estimating these high-dimensional distributions can start to face issues. This is because, currently, we are just using the empirical distribution, which is not very sample efficient, to estimate the probabilities of the maximal cliques and minimal separators of the decomposable models. Therefore, some discussion was had on alternative ways to estimate the probabilities of the maximal cliques and minimal separators of the learnt decomposable models from the given samples.

Arno Siebes speculated that it might be possible to derive computationally efficient approximations to high-dimensional distributions by aggregating the distributions over key positive and negative examples.

Furthermore, there exists a body of work for estimating probability distributions in a sample efficient manner. A promising method for distribution estimation is the Profile Maximum Likelihood (PML) estimator which maximizes the probability of the observed profile, i.e. the number of symbols appearing any given number of times [1].

4.6.2.2 Extension to numerical data

Currently, the method we proposed for estimating divergences from sample data only works for categorical data. However, it might be ideal to extend the approach to numerical data. A potential roadblock to making this extension is that the method we use to efficiently learn decomposable models from data, Chordalysis, only works on categorical data [2, 3].

Geoff Webb pointed out that recently an extension was made to Chordalysis that is able to learn decomposable models from numerical data [5].

4.6.2.3 Conclusions

The use of decomposable models to aid in estimating the divergence between 2 distributions using data sampled from them has some promise in alleviating the difficulties that come with high-dimensional distributions and data. The specific method presented in the Spotlight Presentation is only a first step in this direction.

4.7 Learning Under Concept Drift With Zero Ground Truth

Indrė Žliobaitė (University of Helsinki – Helsinki, Finland, indre.zliobaite@helsinki.fi)

License © Creative Commons BY 3.0 Unported license

© Indrė Žliobaitė

Main reference Indrė Žliobaitė: “Concept drift over geological times: predictive modeling baselines for analyzing the mammalian fossil record”, *Data Min. Knowl. Discov.*, Vol. 33(3), pp. 773–803, 2019.

URL <http://dx.doi.org/10.1007/s10618-018-0606-6>

4.7.1 Abstract of Spotlight Presentation

Fossils are the remains organisms from earlier geological periods preserved in sedimentary rock. The global fossil record documents and characterizes the evidence about organisms that existed at different times and places during the Earth’s history. One of the major directions in computational analysis of such data is to reconstruct environmental conditions and track climate changes over millions of years. Distribution of fossil animals in space and time make informative features for such modeling, yet concept drift presents one of the main computational challenges. As species continuously go extinct and new species originate, animal communities today are different from the communities of the past, and the communities at different times in the past are different from each other. The fossil record is continuously increasing as new fossils and localities are being discovered, but it is not possible to observe or measure their environmental contexts directly, because the time is gone. Labeled data linking organisms to climate is available only for the present day, where climatic conditions can be measured. The approach is to train models on the present day and use them to predict climatic conditions over the past. But since species representation is continuously changing, transfer learning approaches are needed to make models applicable and climate estimates to be comparable across geological times. Here we discuss predictive modeling settings for such paleoclimate reconstruction from the fossil record. We compare and experimentally analyze three baseline approaches for predictive paleoclimate reconstruction: (1) averaging over habitats of species, (2) using presence-absence of species as features, and (3) using functional characteristics of species communities as features. Our experiments on the present day African data and a case study on the fossil data from the Turkana Basin over the last 7 million of years suggest that presence-absence approaches are the most accurate over short time horizons, while species community approaches, also known as ecometrics, are the most informative over longer time horizons when, due to ongoing evolution, taxonomic relations between the present day and fossil species become more and more uncertain.

References

- 1 Indrė Žliobaitė. Concept drift over geological times: predictive modeling baselines for analyzing the mammalian fossil record. *Data Mining and Knowledge Discovery*, 33:773 – 803, 2019.
- 2 O. Oksanen, I. Žliobaitė, J. Saarinen, A.M. Lawing, and M. Fortelius. A humboldtian approach to life and climate of the geological past: Estimating palaeotemperature from dental traits of mammalian communities. In *Journal of Biogeography*, 46(8):1760–1776, 2019.
- 3 I. Žliobaitė, H. Tang, J. Saarinen, M. Fortelius, J. Rinne, and J. Rannikko. Dental ecometrics of tropical africa: linking vegetation types and communities of large plant-eating mammals. In *Evolutionary Ecology Research*, 19:127–147, 2018.
- 4 I. Žliobaitė, M. Fortelius, and N. Chr. Stenseth. Reconciling taxon senescence with the red queen’s hypothesis. In *Nature*, 552:92–95, 2017
- 5 M. Fortelius, I. Žliobaitė, F. Kaya, F. Bibi, R. Bobe, L. Leakey, M. Leakey, D. Patterson, J. Rannikko, and L. Werdelin. An ecometric analysis of the fossil mammal record of the turkana basin. In *Philosophical Transactions B*, 371(1698):1–13, 2016.

4.7.2 Working Group Discussion

Discussion Participants

- Sašo Džeroski, Jozef Stefan Institute
- Gerhard Gößler, Graz University
- Vera Hofer, Graz University
- Claudia Kirch, Magdeburg University
- Georg Krempf, Utrecht University
- Mark Last, Ben-Gurion University of the Negev
- Loong Kuan Lee, Monash University
- Mykola Pechenizkiy, TU Eindhoven
- Jerzy Stefanowski, Poznan University of Technology
- Indrè Žliobaitė, University of Helsinki

4.8 Time Series Classification

Geoffrey I. Webb (Monash University – Clayton, Australia, geoff.webb@monash.edu)

License © Creative Commons BY 3.0 Unported license
© Geoffrey I. Webb

Joint work of Angus Dempster, Francois Petitjean, Geoffrey I. Webb

Main reference Angus Dempster, Francois Petitjean, Geoffrey I. Webb: “ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels”, *Data Min. Knowl. Discov.*, Vol. 34(5), pp. 1454–1495, 2020.

URL <http://dx.doi.org/10.1007/s10618-020-00701-z>

4.8.1 Abstract of Spotlight Presentation

Time series classification is a fundamental data science task, providing understanding of dynamic processes as they evolve over time. The recent introduction of ensemble techniques has revolutionised this field, greatly increasing accuracy, but at a cost of increasing already burdensome computational overheads. Driven by the challenge of global analysis of earth observations over time [1], the Monash Time Series Analytics Group is developing new time series classification technologies that achieve the same accuracy as recent state-of-the-art developments, but with many orders of magnitude greater efficiency and scalability [3, 4, 5, 2]. These make time series classification feasible at hitherto unattainable scale.

The most recent and most scalable of these approaches is Rocket [2], which exploits convolutional filters, popularized by deep learning. There are many different aspects of a series that might be relevant to its classification, such as frequency, amplitude, variance and global or local shape. Convolutional filters provide a single framework which can extract a wide range of such features. Rocket uses random convolutional filters to extract a large number of features which are sufficient for a simple linear classifier to obtain state-of-the-art accuracy in classification.

References

- 1 Charlotte Pelletier, Geoffrey I. Webb, and Francois Petitjean. Temporal convolutional neural network for the classification of satellite image time series. *Remote Sensing*, 11(5), 2019.
- 2 Angus Dempster, Francois Petitjean, and Geoffrey I. Webb. Rocket: Exceptionally fast and accurate time series classification using random convolutional kernels. *Data Mining and Knowledge Discovery*, 34:1454 – 1495, 2020.

- 3 Ahmed Shifaz, Charlotte Pelletier, Francois Petitjean, and Geoffrey I Webb. Ts-chief: A scalable and accurate forest algorithm for time series classification. *Data Mining and Knowledge Discovery*, 34(3):742–775, 2020.
- 4 Hassan Ismail Fawaz, Benjamin Lucas, Germain Forestier, Charlotte Pelletier, Daniel F. Schmidt, Jonathan Weber, Geoffrey I. Webb, Lhassane Idoumghar, Pierre-Alain Muller, and Francois Petitjean. Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, 34:1936–1962, 2020.
- 5 Chang Wei Tan, François Petitjean, and Geoffrey I. Webb. Fastee: Fast ensembles of elastic distances for time series classification. *Data Mining and Knowledge Discovery*, 34(1):231–272, 2020.

4.8.2 Results of the Working Group Discussion

Discussion Participants

- Antoine Cornuéjols, AgroParisTech
- Sašo Džeroski, Jozef Stefan Institute
- Gerhard Gößler, Graz University
- Vera Hofer, Graz University
- Sarah Schnackenberg, formerly TU Dortmund University
- Arno Siebes, Utrecht University
- Jerzy Stefanowski, Poznan University of Technology
- Andreas Theissler, Hochschule Aalen
- Geoffrey I. Webb, Monash University

4.8.2.1 Further information

Geoff Webb’s group’s papers on both time series classification and earth observation can all be found at <http://i.giwebb.com/research/scalable-time-series-classifiers/>.

4.8.2.2 Sensitivity to hyperparameters

There was a discussion about hyper-parameters. At least for the type of time series in the UCR archive, the approach does not appear to be highly sensitive to hyper parameters.

4.8.2.3 Diversity in data stream ensembles

Jurek Stefanowski raised the issues of the role of diversity in data stream ensembles. There are only a few papers on this topic e.g. https://link.springer.com/chapter/10.1007/978-3-319-46307-0_15.

4.8.2.4 Stochasticity

Arno Siebes enquired about the role of stochasticity in Rocket. Geoff Webb told him that all meta parameters of the current feature generation method (i.e., length of the convolution, dilation value, value of the threshold and whether to use padding) are chosen at random for each feature generated. It is an interesting question to what extent this stochasticity is essential to the performance of the method and there is current work on a more deterministic approach to adding the features.

4.8.2.5 Rocket for anomaly detection

Andreas Theissler raised the issue of whether the ROCKET approach be used for other tasks beyond classification and in particular for anomaly detection.

Geoff Webb suggested two potential approaches.

- one-class setting: inside ROCKET learns some sort of distribution, so not using the classifiers' output but the continuous value for a given observation could work — that could be interpreted as an anomaly score.
- two-class setting (highly imbalanced two-class classification problem). While ROCKET was not tested for class imbalance, it might be possible to use the use the random convolutions to model the 'normal' distribution and to use that model for anomaly detection.


4.8.2.6 Conclusions

The discussion identified four open questions on which further research is needed:

- How best to apply the framework in a multivariate context?
- What is the contribution of stochasticity to Rocket's performance? Is it possible to develop a deterministic equivalent to the stochastic approach of Rocket?
- Does the general approach generalize to other data types beyond time series?
- Does the general approach generalize to other tasks beyond classification?

4.9 Uncertainty in Labeling – What Can We Learn from Experiments?

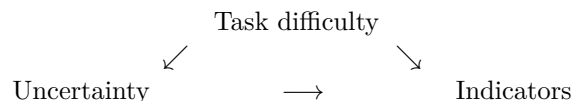
Myra Spiliopoulou (Magdeburg University – Magdeburg, Germany, myra@ovgu.de)

License  Creative Commons BY 3.0 Unported license
© Myra Spiliopoulou

4.9.1 Abstract of Spotlight Presentation

In supervised learning, we need reliable labels. When humans assign labels, they may be uncertain – either because they lack the necessary expertise or because the labeling task is hard or even unsolvable. Epistemic uncertainty can be reduced if more crowdworkers are asked to deliver a label for a given task. Aleatoric uncertainty cannot. So, it is essential to know which of the two cases holds for a given task – preferably before asking many crowdworkers.

This spotlight talk is about investigating the interplay between uncertainty of the crowdworkers and inherent difficulty of the tasks. Since both uncertainty and difficulty are not observable, we discuss the role of observables, of indicators, in experimental settings, as in:



We discuss the role of crowdworker disagreement and the potential of stress measurement for task difficulty assesment.

References

- 1 Nir Nissim, Mary Regina Boland, Yuval Elovici, George Hripcsak, Yuval Shahar, and Robert Moskovitch. Improving condition severity classification with an efficient active learning based framework. In *Journal of Biomedical Informatics*, 61, 03 2016.
- 2 Stefan Rübiger, Myra Spiliopoulou, and Yucel Saygin. How do annotators label short texts? toward understanding the temporal dynamics of tweet labeling. In *Information Sciences*, 457, 05 2018.
- 3 Neetha Jambigi, Tirtha Chanda, Vishnu Unnikrishnan, and Myra Spiliopoulou. Assessing the difficulty of labelling an instance in crowdworking. In *Proc. of the Workshop on Evaluation and Experimental Design in Data Mining and Machine Learning (EDML 2020)*, 2020.
- 4 A. Abolfazli, A. Brechmann, S. Wolff, and M. Spiliopoulou. Machine learning identifies the dynamics and influencing factors in an auditory category learning experiment. In *Scientific reports*, 10:1–12, 2020.
- 5 S. Rübiger, G. Gezici, Y. Saygin, and M. Spiliopoulou. Predicting worker disagreement for more effective crowd labeling. In *Proc. of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 179–188. IEEE, 2018.

4.9.2 Working Group Discussion

Discussion Participants

- Gerhard Gößler, Graz University
- Yun Sing Koh, The University of Auckland
- Myra Spiliopoulou, Magdeburg University
- Dirk Tasche, Swiss Financial Market Supervisory Authority FINMA
- Andreas Theissler, Hochschule Aalen

4.10 Recovery Analysis for Adaptive Learning from Non-stationary Data Streams

Eyke Hüllermeier (Paderborn University – Paderborn, Germany, eyke@upb.de)

License  Creative Commons BY 3.0 Unported license
© Eyke Hüllermeier

Main reference Ammar Shaker, Eyke Hüllermeier: “Recovery analysis for adaptive learning from non-stationary data streams: Experimental design and case study”, *Neurocomputing*, Vol. 150, pp. 250–264, 2015.

URL <https://doi.org/10.1016/j.neucom.2014.09.076>

4.10.1 Abstract of Spotlight Presentation

The extension of machine learning methods from static to dynamic environments has received increasing attention in recent years; in particular, a large number of algorithms for learning from so-called data streams has been developed. An important property of dynamic environments is non-stationarity, i.e., the assumption of an underlying data generating process that may change over time. Correspondingly, the ability to properly react to so-called concept change is considered as an important feature of learning algorithms. In this presentation, we propose a new type of experimental analysis, called recovery analysis, which is aimed at assessing the ability of a learner to discover a concept change quickly, and to take appropriate measures to maintain the quality and generalization performance of the model. Recovery analysis can be instantiated for different types of supervised learning problems, including classification and regression. As a practical application, recovery analysis is used to compare model-based and instance-based approaches to learning on data streams.

References

- 1 Ammar Shaker and Eyke Hüllermeier. Recovery analysis for adaptive learning from non-stationary data streams: Experimental design and case study. *Neurocomputing*, 150:250–264, 2015.
- 2 Ammar Shaker and Eyke Hüllermeier. Recovery analysis for adaptive learning from non-stationary data streams. In *Proceedings of the CORES 2013 Special Session on Data Stream Classification and Big Data Analytics, volume 226 of Advances in Intelligent and Soft Computing*, page 289–298. Springer, 2013.
- 3 Dariusz Brzezinski and Jerzy Stefanowski. Prequential AUC for classifier evaluation and drift detection in evolving data streams. In *International Workshop on New Frontiers in Mining Complex Patterns*, pages 87–101. Springer, 2014.
- 4 Anand M. Narasimhamurthy and Ludmila I. Kuncheva. A framework for generating data to simulate changing environments. In *Artificial Intelligence and Applications*, pages 415–420, 2007.

4.10.2 Results of the Working Group Discussion

Discussion Participants

- Amir Abolfazli, University of Hannover
- Antoine Cornuéjols, AgroParisTech
- Sašo Džeroski, Jozef Stefan Institute
- Eyke Hüllermeier, Paderborn University
- Georg Kreml, Utrecht University
- Mark Last, Ben-Gurion University of the Negev
- Loong Kuan Lee, Monash University
- Eirini Ntoutsi, University of Hannover
- Mykola Pechenizkiy, TU Eindhoven
- Dirk Tasche, Swiss Financial Market Supervisory Authority FINMA

4.10.2.1 Discussed Problem or Approach

One important discussion point centered around the question how to generate the (semi-synthetic) data streams in recovery analysis, which is indeed an important aspect of the approach. On the one side, the data should be sufficiently “realistic” and exhibit properties of real-world scenarios. On the other side, the protocol of recovery analysis assumes the data to have certain “idealized” properties. One very interesting proposal that came up during the discussion was the use generative models. Such models could be trained on real-world data first, making sure to reflect characteristic properties of that data, and then used for sampling the data streams for recovery analysis. By playing with parameters of the models, properties of the streams could be controlled in a convenient manner.

Another interesting idea, namely the use of recovery analysis for feature analysis, was brought up by Georg Kreml. More specifically, the idea is to analyze the “robustness” of features in the context of learning from data streams. In many applications, the usefulness and predictive power of individual features varies in the course of time (a point that was also made in the presentation by Mark Last). One could imagine, for example, a “non-stationary” feature having a high predictive performance under certain conditions or in certain time windows, but a relatively low performance in other periods, with transitions in the form of shifts between these periods. How does such a feature compare with a “stationary” feature the performance of which is moderate throughout? Or, more generally, how does a model (e.g., in the context of an ensemble) using non-stationary features compare with a model

using stationary features? Recovery analysis may provide a suitable basis for analyzing questions of this kind, because non-stationary features will cause (repeated) shifts, and one would expect to observe alternating phases of increasing and decreasing performance, whereas stationary features will show a more stable behavior. Going beyond a qualitative analysis, it would perhaps even be possible to quantify the “robustness” of a feature.

4.10.2.2 Conclusions

In summary, there was an agreement that recovery analysis is an interesting approach for analyzing the performance of machine learning algorithms in the context of data streams, with possible extensions and generalizations in various directions. Apart from further methodological developments, there was also a consensus that a practical and easy-to-use implementation of recovery analysis in a software package is important to popularize the approach.

4.11 Online Linear Discriminant Analysis for Data Streams with Concept Drift

Sarah Schnackenberg (TU Dortmund University – Dortmund, Germany, schnackenberg@statistik.tu-dortmund.de)

License © Creative Commons BY 3.0 Unported license
© Sarah Schnackenberg

Main reference Sarah Schnackenberg: “Online Diskriminanzanalyse für Datensituationen mit Concept Drift”, Dissertation TU Dortmund, 2020.

URL <http://dx.doi.org/10.17877/DE290R-21919>

4.11.1 Abstract of Spotlight Presentation

When focusing on data streams, due to the time component the underlying distribution of the observations can change over time, that means the data can be subject to concept drift.

For handling data streams a range of different online algorithms for various (classification) methods have already been developed (see e.g. [4, 3, 2] for algorithms for online discriminant analysis). Many of them deal with the problem of concept drift and can adapt to changing distributions through e.g. stronger weighting of new observations in the update step [3]. However, the forecasting quality of the resulting classifier of most of the methods can still be improved if the underlying distribution continuously changes further on.

The talk presents the idea of a general extension for existing methods for online discriminant analysis [6, 1]. The time-dependent trend of the expected values of the classes is modelled (and approximated locally linearly) by local linear regression models on sliding windows. With these regression models the forthcoming distribution of the features can be predicted and the predictions replace the original estimators in the continuously updated classification rule of the discriminant analysis in order to improve the forecasting quality.

Note: This spotlight talk has been discussed within the plenary sessions, see Section 5 for the results of these discussions.


References

- 1 Sarah Anna Schnackenberg. *Online Diskriminanzanalyse für Datensituationen mit Concept Drift*. Dissertation, TU Dortmund University, 2020.
Available from <http://dx.doi.org/10.17877/DE290R-21919>.

- 2 Christoforos Anagnostopoulos, Dimitris K. Tasoulis, Niall M. Adams, Nicos G. Pavlidis, and David J. Hand. Online linear and quadratic discriminant analysis with adaptive forgetting for streaming classification. *Statistical Analysis and Data Mining*, 5(2):139–166, 2012.
- 3 Ludmila I. Kuncheva and Catrin O. Plumptre. Adaptive Learning Rate for Online Linear Discriminant Classifiers. In Niels da Vitoria Lobo, Takis Kasparis, Fabio Roli, James T. Kwok, Michael Georgiopoulos, Georgios C. Anagnostopoulos, and Marco Loog, editors, *Structural, Syntactic, and Statistical Pattern Recognition*, volume 5342 of *Lecture Notes in Computer Science*, pages 510–519, Berlin, Heidelberg, 2008. Springer.
- 4 Shaoning Pang, Seiichi Ozawa, and Nikola Kasabov. Incremental Linear Discriminant Analysis for Classification of Data Streams. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 35(5):905–914, 2005.
- 5 Shaoning Pang, Seiichi Ozawa, and Nikola Kasabov. Chunk Incremental LDA Computing on Data Streams. In Jun Wang, Xiaofeng Liao, and Zhang Yi, editors, *Advances in Neural Networks – ISNN 2005*, volume 3497 of *Lecture Notes in Computer Science*, pages 51–56, Springer, Berlin, Heidelberg, 2005.
- 6 Sarah Schnackenberg, Uwe Ligges, and Claus Weihs. Online Linear Discriminant Analysis for Data Streams with Concept Drift. *Archives of Data Science, Series A (Online First)*, 5(1):02, 2018.

4.12 Classification, Calibration, and Quantification: A Study of Dataset Shift

Dirk Tasche (Swiss Financial Market Supervisory Authority FINMA – Bern, Switzerland, dirk.tasche@gmx.net)

License  Creative Commons BY 3.0 Unported license
© Dirk Tasche

Main reference Dirk Tasche: “Exact Fit of Simple Finite Mixture Models”, *Journal of Risk and Financial Management*, Vol. 7(4), pp. 150–164, 2014.

URL <https://doi.org/10.3390/jrfm7040150>

4.12.1 Abstract of Spotlight Presentation

What happens if the true dataset shift type is prior probability shift but the prevalence of the positive class is estimated under an assumption of covariate shift (see [2] for the definitions of these shifts)? We present a simple inequality for the estimation error which shows that the size of the change of the prevalence between training set and test set is always underestimated (Corollary 6 of [4]). The degree of underestimation decreases with increasing predictive power of the posterior class probabilities on the training set. We also discuss a possible application to the estimation of bounds for the change of the positive class prevalence under general dataset shift.

References

- 1 A. Bella, C. Ferri, J. Hernandez-Orallo, and M.J. Ramírez-Quintana. Quantification via probability estimators. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 737–742. IEEE, 2010.
- 2 J.G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N.V. Chawla, and F. Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, 2012.
- 3 Clayton Scott. A Generalized Neyman-Pearson Criterion for Optimal Domain Adaptation. In *Proceedings of Machine Learning Research, 30th International Conference on Algorithmic Learning Theory*, volume 98, pages 1–24, 2019.

- 4 Dirk Tasche. Exact fit of simple finite mixture models. *Journal of Risk and Financial Management*, 7(4):150–164, 2014.

4.12.2 Results of the Working Group Discussion

Discussion Participants

- Vera Hofer, Graz University
- Georg Kreml, Utrecht University
- Dirk Tasche, Swiss Financial Market Supervisory Authority FINMA

4.12.2.1 Open research questions

1. Theory: How to generalise the inequality of Corollary 6 of [4] to the multi-class case?
2. Application: Can the inequality be used to estimate bounds for the change of the positive class prevalence under general dataset shift?

Rationale: Anecdotal evidence suggests that covariate shift and prior probability shift to some extent are extreme dataset shifts (least vs. greatest change of positive class prevalence).

Hence, if \tilde{q} is a reasonable estimate of the true prevalence q under an assumption of prior probability shift and \hat{q} denotes the probability average estimator [1] of q under a covariate shift assumption, then perhaps $\min(\hat{q}, \tilde{q})$ and $\max(\hat{q}, \tilde{q})$ provide a reasonable range estimate for q ?

Empirical evidence from rating agency data so far is not too encouraging.

4.12.2.2 Conclusions

There is no simple answer to question 1. Further research on Covariate Shift with Posterior Drift (CSPD, [3]) might support progress on question 2.

4.13 Prediction-Dependent Drift

Georg Kreml (Utrecht University – Utrecht, The Netherlands, g.m.kreml@uu.nl)

License  Creative Commons BY 3.0 Unported license
© Georg Kreml

Joint work of Georg Kreml, Jelsma, Tineke

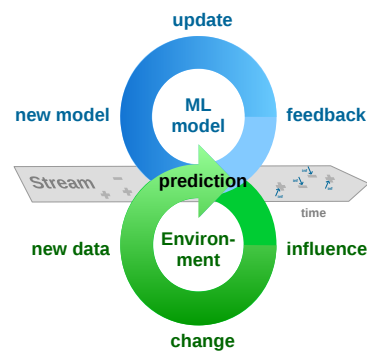
Main reference Georg Kreml, David Bodnar, and Anita Hrubos: “When learning indeed changes the world: Diagnosing prediction-induced drift”. In Tijl De Bie, Elisa Fromont, and Matthijs van Leeuwen, editors, *Advances in Intelligent Data Analysis XIV – 14th Int. Symposium, IDA 2015*, volume 9385 of LNCS, page XXII–XXIII. Springer, 2015.

URL <https://doi.org/10.1007/978-3-319-24465-5>

4.13.1 Abstract of Spotlight Presentation

The predominant paradigm for learning machine learning-based prediction systems is that they act as observers in their environment. However, as their decisions are put into action, this neglects the influence they potentially might play in their environment. For example, a company predicted to be high risk might face higher financing costs, or a region patrolled regularly might be avoided by criminals. This might lead to self-fulfilling or self-defeating prophecies.

This talk proposes the new paradigm of *influential machine learning*. Therein, *feedback loops* exist between a machine learner’s predictions and the subsequent data they receive: predictions have an *influence* on the statistical population under study, and thereby trigger *changes* in its characteristics. This raises several fundamental questions, such as:



■ **Figure 3** Machine Learning under Influential Predictions.

- Under which circumstances and in which ML applications could such influence occur?
- How to describe the influence mechanisms, and how to model them statistically?
- How to detect, assess and verify this influence?

While these questions are, to the best of our knowledge, not resolved yet, they relate to several existing lines of research. Most notably, in the widely studied problem of non-stationary and streaming data, several tasks and approaches have been proposed to identify change or irregularities[3, 4]. Nevertheless, all this research studies change that is (implicitly) assumed to be independent of previous predictions. In adversarial machine learning[2], the focus of interest is hardening a machine learning system against attacks by an adversary. This includes so-called evasion attacks, where the adversary deliberately alters the characteristics of fraudulent instances, such that they get subsequently misclassified.

Thus, this talk will provide a first model of such an influence mechanism as well as some preliminary results of an influence detection approach on synthetic data. These indicate challenges in particular when detecting self-defeating prediction influence.

References

- 1 Georg Kreml, David Bodnar, and Anita Hrubos. When learning indeed changes the world: Diagnosing prediction-induced drift. In Tijl De Bie, Elisa Fromont, and Matthijs van Leeuwen, editors, *Advances in Intelligent Data Analysis XIV – 14th Int. Symposium, IDA 2015*, volume 9385 of *LNCIS*, page XXII–XXIII. Springer, 2015.
- 2 Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2017.
- 3 João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):1–44, 2014.
- 4 Geoffrey I Webb, Roy Hyde, Hong Cao, Hai Long Nguyen, and Francois Petitjean. Characterizing concept drift. *Data Mining and Knowledge Discovery*, 30(4):964–994, 2016.
- 5 Qihua Liu, Xiaoyu Zhang, Liyi Zhang, and Yang Zhao. The interaction effects of information cascades, word of mouth and recommendation systems on online reading behavior: an empirical investigation. *Electronic Commerce Research*, 43, 2018.

4.13.2 Results of the Working Group Discussion

Discussion Participants

- Amir Abolfazli, University of Hannover
- Sašo Džeroski, Jozef Stefan Institute
- Georg Kreml, Utrecht University

- Mark Last, Ben-Gurion University of the Negev
- Mykola Pechenizkiy, TU Eindhoven
- Sarah Schnackenberg, formerly TU Dortmund University
- Arno Siebes, Utrecht University
- Myra Spiliopoulou, Magdeburg University
- Jerzy Stefanowski, Poznan University of Technology
- Andreas Theissler, Hochschule Aalen

4.13.2.1 Real-World Data Issue

A starting point for research into influential predictions is the identification of real-world applications, where such influence mechanisms are likely to play a role. This allows to discuss possible feedback loop mechanisms with domain experts, to develop simulations to generate synthetic data, and to collect real-world data with known, or at least suspected, prediction influence. This will help in a later step of evaluating detection and mitigation approaches. During the working group discussion, several applications were suggested. In pilot experiments on prediction influence in [1], data from neurobiological learning experiments was used, which offers a well-controlled environment. Further applications could include data from known medical interventions, as suggested by Arno Siebes, or data from predictive maintenance applications, as suggested by Myra Spiliopoulou. Another promising application are recommender systems, as suggested by Mykola Pechenizkiy. Here, recent studies have shown that predictions change user preferences and affect sales diversity [5].

4.13.2.2 Approach

In a pilot study [1] and in ongoing experiments with synthetic, generated data, an approach has been developed that splits the instances based on their predicted class label, as well as on their actual label. Following the classification of a first chunk of instances, in subsequent time steps the distributional change in the neighbourhood of each previously classified instance is attributed to the corresponding cells in the classification confusion table. This allows to compare the observed aggregated values against those, who were to be expected under assumed independence between distributional change and previous classification.

While first results on synthetically generated data are promising in particular for detecting self-fulfilling influence, a particular challenge is self-defeating influence with unknown time lags between the moments of prediction and of the manifestation of influence.

4.13.2.3 Conclusions

The existence of feedback loops, i.e., of some dependence between previous predictions and subsequent distributional changes, might have relevance for several machine learning applications and should be further investigated.

From a methodological point of view, this is currently a largely un(der)studied problem. First preliminary results indicate that influence mechanisms that result in a self-defeating prediction pattern might be particularly challenging to detect. Therefore, in order to address the third research question, further development of prediction influence detection approaches is needed.

4.14 Labelless Detection and Explanation of Concept Drift

Mykola Pechenizkiy (TU Eindhoven – Eindhoven, The Netherlands, m.pechenizkiy@tue.nl)

License © Creative Commons BY 3.0 Unported license
© Mykola Pechenizkiy

Main reference Shihao Zheng, Simon B van der Zon, Mykola Pechenizkiy, Cassio P de Campos, Werner van Ipenburg, Hennie de Harder, and Rabobank Nederland: “Labelless concept drift detection and explanation”. In *NeurIPS 2019 Workshop on Robust AI in Financial Services: Data, Fairness, Explainability, Trustworthiness, and Privacy*, 2019.

URL <http://www.win.tue.nl/mpechen/publications/pubs/Zheng2019.pdf>

4.14.1 Abstract of Spotlight Presentation

The common classification models are assumed to be trained on data that are sufficient and representative of the underlying unknown distribution. However, in real-world scenarios, the joint distribution of features and labels is not stationary but drifting from time to time. This phenomenon, referred to as concept drift, can deteriorate the predictive performance of existing classification model used e.g. in fraud detection and even make it obsolete. Numerous concept drift detection methods have been developed to detect drifts and adapt the model so as to recover from the influence of concept drift. However, most existing concept drift detection methods have an over-optimistic assumption that the true labels will be available after the classifier makes decisions on new coming instances so that they can track concept drift by monitoring the real-time accuracy. Besides, the localization and interpretation of concept drift are also important. Localizing drift positions and providing interpretable concept drift information would help improve usability and trustworthiness in model adaptation process but existing methods that use accuracy to track concept drift cannot provide in-depth explanations on the root causes of the drift. To address the issues mentioned above, we propose a Labelless CONcept Drift Detection and Explanation Framework (L-CODE). It requests labels only when we need to update the model and uses the Shapley values as a proxy to the joint distribution of features and labels. Our method tracks change on each feature separately, which is more efficient, but we can still obtain multivariate changes based on the multivariate nature of Shapley values. Except for drift detection, we provide three-level visualizations to explain the detected drift in different granularities. Our method can outperform other state-of-the-art labelless drift detection methods on benchmark datasets but cannot beat the methods that require labels. For experiment on Rabobank transaction dataset, we demonstrate insightful explanations on the causes of detected drift.

Note: This spotlight talk has been discussed within the plenary sessions, see Section 5 for the results of these discussions.

References

- 1 Shihao Zheng, Simon B van der Zon, Mykola Pechenizkiy, Cassio P de Campos, Werner van Ipenburg, Hennie de Harder, and Rabobank Nederland. Labelless concept drift detection and explanation. In *NeurIPS 2019 Workshop on Robust AI in Financial Services: Data, Fairness, Explainability, Trustworthiness, and Privacy*, 2019.

5 Plenary Discussion

5.1 Plenary Discussion

All participants of this seminar.

License © Creative Commons BY 3.0 Unported license
© All participants of this seminar.

5.1.0.1 Ambiguities in Terminology, Concepts and Common Assumptions

There are differences in the terminology, concepts and common assumptions that are used in the different communities to describe distributional change. This starts already in the characterisation of of change types. As pointed out by Barbara Hammer, a link between domains is missing.

In **data stream mining**, for example, the classification into sudden, gradual, incremental, and recurring drift is common [1]. Another is the distinction based on which distribution is subject to change [2]. However, even within the data stream mining community there is ambiguity around some terms. An example is *virtual concept drift*, which originally was defined in [3, page 3] as not occurring in reality, but rather “*in the computer model reflecting this reality.*”, for example due to representation language failing to identify all relevant features, or when a skewed order of training examples results in an uneven distribution of instances over the training sequence. In contrast, [4, page 143] related this term to the problem of *sampling shift* discussed in [5]. As noted in [7], this ambiguous term is nowadays often used to denote changes in the feature distribution that do not affect the posterior class probabilities. However, it is less ambiguous to refer to this as *covariate shift* or *covariate drift*, the term used for example in [6] and [2], respectively. Another difference is whether time in data streams is defined as discrete or continuous, and there have been controversies about differences between incremental and gradual drift, with a detailed taxonomy having been proposed in [8].

In the classical **time series analysis** literature, instances are assumed to exhibit some dependency in time such that they are not independent of each other. However, in the literature on *multiple change points* aka *data segmentation* often independence of the observations or even independence and Gaussianity are assumed [9]. While not necessarily allowing for time-dependencies all of these algorithms use the fact that this is *ordered data* (i.e. the most basic definition of time series) to define change points.

A first conclusion was made that the **assumption of temporal dependency** seems to be an important difference between these two fields. In time series, the assumption of temporal dependency is usually important. In contrast, in data streams there might be temporal dependencies, but not necessarily. In case of no time dependency, an entire field of algorithms for time series becomes pointless (the ones that model the time dependencies, e.g. a simple AR-model). It was noted that pointing out these differences would be a worthy contribution to the existing literature.

Furthermore, for developing a mapping of change types to suggested models, precise and unambiguous definitions of types of distributional change are needed. Therefore, several suggestions were made in the discussions:

- A **characterisation** of distributional change must be **independent of the observed time window**.
- The assumed **invariants** behind a categorisation must be clearly identified and stated.
- The **dependency between techniques and the types of distributional change** with their assumptions needs to be made clear.

5.1.0.2 Equivalent Notions of Drift

In machine learning, distributional change commonly refers to the fact that the probability distribution changes in between two time points – yet this property cannot necessarily be observed in the case of a continuously changing distribution since only few observations can be attributed to a specific point in time. In practice, drift detection is therefore often based on the decrease of a model accuracy which has been trained over a period of time, or change of time series characteristics. A step towards developing a common understanding of distributional change is the embedding of drift processes into continuous time and the development of equivalent notions of drift. This has been done in a recent work by [10, 11]. Therein, a drift process can be formalized over continuous time based on measurability properties. This framework enables an investigation under which assumptions popular notions of drift, in particular the existence of change points, changed model accuracy, and distributional changes are equivalent, and it derives a further equivalent notion of drift based in independence of variables. The latter characteristics opens novel possibilities for drift detection and drift explanation technologies.

5.1.0.3 Evaluation

Several questions concerning evaluation arise in the context of distributional change:

- How to evaluate?
- How long to wait before evaluating?
- What means *interpretable*? Is the *utility* of an interpretation a good indicator?
- How to consider in evaluations the relationship between time scales and types of drift, to techniques and their performance.
- Which data to use in evaluations?
- How to consider the timing of information in the evaluation? For example, there has been some literature on verification latency or delayed labelling [12, 13], which also allows to bridge literature from data stream mining and domain adaptation.

Evaluation has been discussed in further detail during the breakout group discussion on recovery analysis, see subsection 4.10 and [15] as well as [14], and [16].

5.1.0.4 Conclusion

This seminar highlighted the need to better integrate the different communities in the areas of data stream mining, statistical change point detection and early classification in time series, transfer learning and domain adaptation, adversarial machine learning and exceptional model mining. In order to foster a common understanding and to facilitate cross-fertilisation, the development of a common vocabulary is key. Therefore, the participants have discussed the idea of a joint position paper, which aligns the concepts and terms used in the different communities to each other, and identifies overlaps and research gaps.

References

- 1 Indrè Žliobaitė, Mykola Pechenizkiy, and Joao Gama. An overview of concept drift applications. In Nathalie Japkowicz and Jerzy Stefanowski, editors, *Big Data Analysis: New Algorithms for a New Society*, page 91–114. Springer, 2016.
- 2 J.G. Moreno-Torres, T. Raeder, R. Alaiz-Rodriguez, N.V. Chawla, and F. Herrera. A unifying view on dataset shift in classification. *Pattern Recognition*, 45(1):521–530, 2012.
- 3 Gerhard Widmer and Miroslav Kubat. Effective learning in dynamic environments by explicit context tracking. In *Proceedings of the European Conference on Machine Learning*, volume 667 of *Lecture Notes In Computer Science*, page 227–243. Springer, 1993.
- 4 Gladys Castillo. *Adaptive Learning Algorithms for Bayesian Network Classifiers*. PhD thesis, Universidade de Aveiro, Departamento de Matemática, 2006.

- 5 Marcos Salganicoff. Tolerating concept and sampling shift in lazy learning using prediction error context switching. *Artificial Intelligence Review*, 11(1–5):133–155, 1997.
- 6 Amos Storkey. When training and test sets are different: characterising learning transfer. In *Dataset Shift in Machine Learning*, page 1–28. MIT Press, 2009.
- 7 João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):1–44, 2014.
- 8 Geoffrey I Webb, Roy Hyde, Hong Cao, Hai Long Nguyen, and Francois Petitjean. Characterizing concept drift. *Data Mining and Knowledge Discovery*, 30(4):964–994, 2016.
- 9 Haeran Cho and Claudia Kirch. Data segmentation algorithms: Univariate mean change and beyond. *arXiv preprint arXiv:2012.12814*, 2020.
- 10 Fabian Hinder, André Artelt, and Barbara Hammer. A probability theoretic approach to drifting data in continuous time domains. *arXiv preprint arXiv:1912.01969*, 2019.
- 11 Fabian Hinder, André Artelt, and Barbara Hammer. Towards non-parametric drift detection via dynamic adapting window independence drift detection (DAWIDD). In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4249–4259. PMLR, 2020.
- 12 Vera Hofer and Georg Kreml. Drift mining in data: A framework for addressing drift in classification. *Computational Statistics and Data Analysis*, 57(1):377–391, 2013.
- 13 Georg Kreml. The algorithm APT to classify in concurrence of latency and drift. In João Gama, Elizabeth Bradley, and Jaakko Hollmén, editors, *Advances in Intelligent Data Analysis X*, volume 7014 of *Lecture Notes in Computer Science*, page 222–233. Springer, 2011.
- 14 Ammar Shaker and Eyke Hüllermeier. Recovery analysis for adaptive learning from non-stationary data streams: Experimental design and case study. *Neurocomputing*, 150:250–264, 2015.
- 15 Anand M. Narasimhamurthy and Ludmila I. Kuncheva. A framework for generating data to simulate changing environments. In *Artificial Intelligence and Applications*, pages 415–420, 2007.
- 16 Indrė Žliobaitė, Albert Bifet, Jesse Read, Bernhard Pfahringer, and Geoff Holmes. Evaluation methods and decision theory for classification of streaming data with temporal dependence. *Machine Learning*, 98(3):455–482, 2015.

Acknowledgements

We would like to thank the colleagues who have contributed to discussions prior to this seminar. In particular, we would like to thank:

- | | | |
|--|--|--|
| ■ Niall M. Adams
Imperial College London, GB | ■ Albert Bifet
University of Waikato –
Hamilton, NZ & Télécom
ParisTech, FR | ■ Bernhard Pfahringer
University of Waikato –
Hamilton, NZ |
| ■ Harish S. Bhat
University of California –
Merced, US | | ■ Robi Polikar, Rowan
University |

Remote Participants

- Amir Abolfazli
University of Hannover, DE
- Shai Ben-David
University of Waterloo, CA
- Antoine Cornuéjols
AgroParisTech, FR
- Sašo Džeroski
Jozef Stefan Institute –
Ljubljana, SI
- Johannes Fürnkranz
Johannes Kepler University –
Linz, AT
- João Gama,
University of Porto PT
- Gerhard Gößler
Graz University, AT
- Vera Hofer
Graz University, AT
- Eyke Hüllermeier
Paderborn University, DE
- Yun Sing Koh
The University of Auckland, NZ
- Mark Last
Ben-Gurion University of the
Negev – Beer Sheva, IL
- Loong Kuan Lee
Monash University –
Clayton, AU
- Pavlo Mozharovskyi
Télécom Paris, FR
- Eirini Ntoutsi
University of Hannover, DE
- Arno Siebes
Utrecht University, NL
- Jerzy Stefanowski
Poznan University of
Technology, PL
- Ruth Urner
York University, CA
- Geoffrey I. Webb
Monash University –
Clayton, AU
- Indrė Žliobaitė
University of Helsinki, FI

Participants

- Barbara Hammer
Universität Bielefeld, DE
- Claudia Kirch
Universität Magdeburg, DE
- Georg Krempel
Utrecht University, NL
- Mykola Pechenizkiy
TU Eindhoven, NL
- Sarah Schnackenberg
Köln, DE
- Myra Spiliopoulou
Universität Magdeburg, DE
- Dirk Tasche
FINMA – Bern, CH
- Andreas Theissler
Hochschule Aalen, DE

