ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

Contrasting Explanation of Concept Drift*

Fabian Hinder[1], André Artelt[1†], Valerie Vaquet[1] and Barbara Hammer[1]

1 - Bielefeld University - Cognitive Interaction Technology (CITEC)
Inspiration 1, 33619 Bielefeld - Germany

Abstract. The notion of concept drift refers to the phenomenon that the distribution, which is underlying the observed data, changes over time. As a consequence machine learning models may become inaccurate and need adjustment. While there do exist methods to detect concept drift or to adjust models in the presence of observed drift, the question of *explaining* drift is still widely unsolved. This problem is of importance, since it enables an understanding of the most prominent drift characteristics. In this work we propose to explain concept drift by means of contrasting explanations describing characteristic changes of spatial features. We demonstrate the usefulness of the explanation in several examples.

1 Introduction

Data from the real world such as social media entries or measurements of IoT devices are subject to continuous changes known as concept drift [1, 2]. It can be caused by seasonal changes, changed demands, ageing of sensors, etc. Since drift might induce severe problems in machine learning models, it is important to understand the nature and the characteristics of the ongoing drift. In recent years, several approaches were proposed to deal with concept drift [3, 4]. These range from non-parametric methods over gradient techniques up to ensemble techniques for dealing with streaming data [5]. In addition to model adaptation schemes, a large number of methods aims for a detection of drift, an identification of change points in given datasets, or a characterization of overarching types of drift [6, 7]. While drift detection is a first step to enable necessary human intervention, a more detailed description is desirable. Current methods are limited to drift detection, drift quantification [4], the identification of features that characterize the drift [8, 9], and first approaches to determine the location of drift in data space [10], which allows one to characterize certain forms of drift, but do not provide a condensed and easily accessible explanation.

In this work, we aim for a novel, fully automated, exemplar-based drift characterization and explanation scheme, which highlights characteristic spatial locations and their temporal behavior, which lead to the observed drift. To achieve this we combine explainable AI (XAI), which usually does not focus on the explanation of drift [11], with the drift localization method presented in [10]. We focus on *contrasting explanations* [12], and in particular, counterfactual explanations [13] which are considered to align well with explanation schemes that are used by humans [14]. However, the presented explanation scheme can easily be adapted to other explanation methods from the literature [15]. Thereby, we significantly improve existing methods, which are usually limited to featurewise inspections [16, 17, 8]. We are not aware of approaches, which investigate complex explanations of drift.

[*] Funding in the frame of the BMBF project TiM, 05M20PBA, and from the VW-Foundation for the project *IMPACT* is gratefully acknowledged.

[†] Affiliation with the University of Cyprus

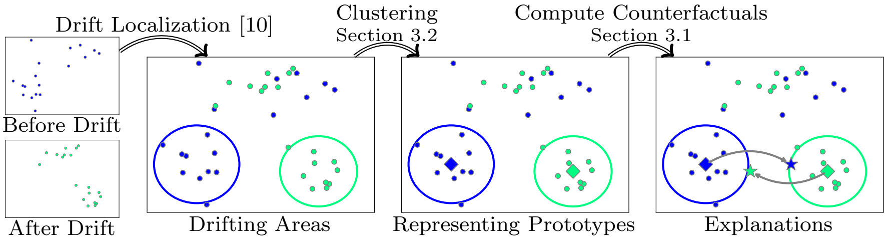


Fig. 1: Generating conterfactual explanation for concept drift (conceptual)

This paper is organized as follows: In the first part (Section 2) we recall the formal definitions of concept drift and drift localization, and provide a high level description of the provided explanation scheme. We then consider contrasting explanation in the context of drift localization in Section 3. In the second part we empirically evaluate the resulting algorithm in two experiments (Section 4).

2 Problem Setup

In classical machine learning (ML) one considers a generative process p on the sample space $\mathcal{X}$. A data point is an instance of a random variable $X \sim p$. Many processes in real-world applications are time dependent. One prominent way to take time into account, is to consider a family of probability measures p_t on $\mathcal{X}$, indexed over a set $\mathcal{T}$, representing time. The distributions p_t can change over time and concept drift takes place if $p_t \neq p_s$ for at least one pair $t \neq s$ [18].

While drift detection techniques enable automatic drift identification it is often unclear how to react to such drift. This challenge is ill-posed in general and requires expert insight. An explanation would increase an understanding of the ongoing drift and thereby enable a human to initiate an appropriate reaction. A drift characterization is particularly demanding for high dimensional data or a lack of clear semantic features. For this purpose, we rely on an example-based explanation scheme, which work by presenting pairs of samples that are similar up to contrasting features, that are particularly relevant for the drift [13, 15].

Suppose we are considering steams of pictures taken by stationary web cams. If a building is constructed or demolished within sight of one of the cams, it will cause drift. A fairly good explanation of this drift is given by presenting two picture, taken by the same web cam, one containing and one lacking the building – allowing the user to grasp the difference in an intuitive way.

As shown in [10] such features can be found by ML models that characterize regions in data space where the distribution changes. The problem of identifying those regions – referred to drift localization [4, 10] – was tackled in [10] by training a model to predict the time t based on the observed sample x. Since this results in a classical ML model, we can apply well known methods to extract explanations – like the absence of the building in the example above.

Since contrasting explanations provide local rather than global insight, we also need to identify particularly characteristic samples where the drift manifests itself in an automated fashion. We rely on weighted prototype-based clustering methods and show the validity of this simplified approach. In Figure 1 we illustrated the explanation process and provide the pseudocode in Algorithm 1.

In accordance to [10] we will focus on the case of two time points $\mathcal{T} = \{1, 2\}$ with disjoint samples sets S_1, S_2 sampled from p_1 and p_2, respectively.

2.1 Related Work

Quite a number of approaches aim for a detection and quantification of drift [4, 19], its localization in space [4, 10], or visualization [16, 8, 17]. Several approaches focus on feature-wise representations of drift [16, 19, 8, 17]. These are limited if high dimensional data or correlated features are dealt with. To the best of our knowledge no other work uses general XAI methods to explain concept drift, and the existing methods cannot be applied to domains like images.

3 Extracting Explanations

Since the task of drift localization can be reduced to a probabilistic classification problem as discussed in [10], we can rely on standard explanation schemes to explain the classification models used for the drift localization to obtain an explanation for the drift. Model explanations have been studied extensively in the past [11]. One prominent tool, which is considered to be intuitive for human users [20, 15], are counterfactual explanations, which we will recall in the next section. The main idea is to illustrate the characteristic patterns by contrasting a sample, obtained from a time point, with a generated second sample, that is as similar as possible except that it does not carry the time point specific patterns. By presenting both samples to the user they can grasp this patterns and thereby understand the drift. We exploit the detail of this idea in the next sections.

3.1 Explanations via Counterfactual

Counterfactuals explain the model's classification of a given sample by contrasting it with a similar sample, which is classified differently [13]: For a classifier h, a loss function ℓ, and dissimilarity d, a counterfactual x' for $x \in \mathcal{X}$ with class $y \neq h(x)$ is obtained by minimizing $\ell(h(x'), y) + C \cdot d(x', x)$, where $C > 0$ is a regularization constant.

This initial definition suffers from several problems as they might be implausible [21, 22], a problem that can be solved by considering samples that lie on the data manifold only [22], e.g., by enforcing a lower threshold $\alpha > 0$ for their probability. Another approach [23] restricts feasible solutions to the training data, i.e. we select the closest sample from the training dataset classified as y.

3.2 Explaining Drift by Means of Contrasting Explanations

We propose to apply counterfactuals to obtain an explanation for drift using the following pipeline (see Figure 1): In [10] the task of identifying relevant information regarding the observed drift was reduced to a probabilistic classification problem, mapping representative samples to their time of occurrence via a model h. This connection enables us to reduce understanding drift to providing contrasting explanations (i.e. counterfactuals) for h. Those can be computed using h directly or by training a second model that classifies the samples into "before drift", "after drift" and "non-drifting" – where the classes are obtained by using drift localization [10]. This can drastically decrease the computational cost of the construction of counterfactuals as the model is usually less complex.

Algorithm 1 Explain Drift

```
Input: $S \subset \mathcal{X} \times \mathcal{T}$ dated data points
$L \leftarrow$ SelectDriftingSamples($S$)                      ▷ Drift localization
$h \leftarrow$ TrainModel($L \cup \{(x,-1) | (x,t_x) \in S \setminus L\}$)   ▷ $-1$ is class "no drift"
for all $t \in \mathcal{T}$ do
    $P_t \leftarrow$ ClustProto($\{x \mid (x,t_x) \in L, t_x = t\}$)
    for all $x \in P_t$ and $t' \in \mathcal{T}$ do
        print ContrastExpl($h, x, t'$)
    end for
end for
```

The first step of this approach is to determine the samples which are used for the computation of counterfactuals. In case there is a human in the loop, they can select the most interesting samples. As we aim for a completely autonomous system, which explains the drift at each time step it is mandatory to automate this step: to do so we first filter out the non-drifting samples and then compute representations of the drifting data using prototype-based clustering algorithms like mean shift, Gaussian mixture models, or k-means. We refer to the obtained prototypes as *characteristic samples.* This approach provides a valid resampling scheme that allows a statistical interpretation in terms of reweighting functions.

Combining all these steps, we obtain Algorithm 1. The explaining routine is started if drift is detected and explanations are requested. Time and space complexity depend on the specific algorithmic instantiations, however, for many popular choices of subroutines we end up with $\mathcal{O}(n^2)$.

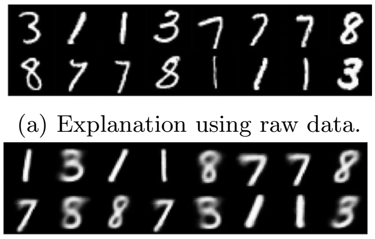
4 Experiments

In this section, we empirically evaluate our proposed method[1]. This includes a quantitative evaluation of identifying relevant features, as well as example applications to benchmark and image data streams. In the following we assume that the data stream is partitioned by a drift detection method into non-drifting segments, for which contrasting explanations will be computed.

Drifting features / Sensor faults (Nebraska Weather) We evaluate the capability to identify drifting features. We simulate sensor fault induced drift on the Nebraska Weather dataset [24]. We draw two windows from the stream (each containing 500 randomly selected samples) and induce one of the following feature perturbations (FP) to one feature: setting to zero, adding a fixed shift, or adding Gaussian noise. We then use sparse counterfactuals to identify those features, whereby we consider a feature as relevant for the drift, if the number of explanations that make use of it significantly surpasses average.

We use decision trees as model h, affinity propagation to obtain the characteristic samples, and assume that a localization is provided to evaluate the robustness of the explanation. We investigate the effect of the amount of localization errors (LE), which are simulated by marking a certain percentage of samples as non-drifting. To evaluate the method we measure how many times the perturbed features is accurately identified as drifting by the method (pre-

[1] The code and additional examples are available at `https://github.com/FabianHinder/Contrasting-Explanation-of-Concept-Drift.`



(a) Explanation using raw data.

(b) Explanation using VAE.

Fig. 2: Explanation of MNIST. In each image: Original/Prototype (Top), Counterfactual (Bottom)

FP	LE	precision	recall	F1
gaussian	0%	0.84±0.31	0.91±0.29	0.87±0.29
	10%	0.84±0.33	0.89±0.31	0.85±0.32
	20%	0.80±0.35	0.88±0.32	0.82±0.33
	40%	0.78±0.37	0.85±0.36	0.80±0.36
shift	0%	0.86±0.24	0.99±0.10	0.90±0.17
	10%	0.88±0.22	1.00±0.00	0.92±0.14
	20%	0.86±0.25	0.98±0.14	0.90±0.19
	40%	0.87±0.25	0.97±0.17	0.90±0.21
zero	0%	1.00±0.00	1.00±0.00	1.00±0.00
	10%	1.00±0.00	1.00±0.00	1.00±0.00
	20%	1.00±0.00	1.00±0.00	1.00±0.00
	40%	1.00±0.00	1.00±0.00	1.00±0.00

Table 1: Mean feature detection scores on Nebraska Weather dataset over 100 runs.

cision, recall, and F1). The results are shown in Table 1. As can be seen our method provides sufficient performance for all types of considered perturbations and even in case of a high number of wrong localizations.

Non-sematic, high dimensional data (MNIST) We specifically showcase our method on a stream of image data. In contrast to the first experiment, the drift localization is considered as a part of the explanation process. We consider a subset of the 28×28-pixel black-white MNIST images. The digits 1, 3, and 4 are present before and the digits 7, 8, and 4 after the drift. Intuitively speaking the drift replaces 1 and 3 by 7 and 8 in the stream. We performed two experiments: 1) on the raw dataset we use a decision tree as model h (as in [10]) and select characteristic samples by affinity propagation and the counterfactuals from the training data (Raw), 2) we use a variational autoencoder and perform the drift localization and generation of the counterfactuals in the latent space using decision trees and k-means (VAE). We generate four explanations in both directions, i.e. before to after drift and after to before drift, the results are presented in Figure 2, with the upper row showing the characteristic samples and the lower row the associated counterfactuals. We observe that only the digits 1, 3, 7, 8 are considered to be relevant for the drift; only the digit 4, which is non-drifting by design, is not considered to be relevant for the drift as expected. Thus, our method shows exactly the "replacement" that constitutes the drift.

5 Conclusion and Further Work

We introduced a new method for explaining drift by means of characteristic sample. We derived the explanation methodology by reducing the problem of explaining drift to the problem of explaining a model that is used to localize the drift. We demonstrated the usefulness of this methodology in two examples, and the empirical results demonstrate that this proposal constitutes a promising approach as regards drift explanation in an intuitive fashion. The technology is yet restricted to discrete time points with well defined change points. An extension to continuous drift is subject of ongoing work.

References

[1] A. Bifet and J. Gama. Iot data stream analytics. *Ann. des Télécomm.*, 75(9-10), 2020.

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

[2] S. Tabassum, F. S. F. Pereira, S. Fernandes, and J. Gama. Social network analysis: An overview. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 8(5), 2018.

[3] G. Ditzler, M. Roveri, C. Alippi, and R. Polikar. Learning in nonstationary environments: A survey. *IEEE Comp. Int. Mag.*, 10(4), 2015.

[4] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang. Learning under concept drift: A review. *IEEE TKDE*, 2018.

[5] B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, and M. Wozniak. Ensemble learning for data stream analysis: A survey. *Inf. Fusion*, 37, 2017.

[6] S. Aminikhanghahi and D. J. Cook. A survey of methods for time series change point detection. *Knowl. Inf. Syst.*, 51(2), May 2017.

[7] I. Goldenberg and G. I. Webb. Survey of distance measures for quantifying concept drift and shift in numeric data. *Knowl. Inf. Syst.*, 60(2), 2019.

[8] G. I. Webb, L. K. Lee, F. Petitjean, and B. Goethals. Understanding concept drift. *CoRR*, abs/1704.00362, 2017.

[9] Limin Yang, Wenbo Guo, Qingying Hao, Arridhana Ciptadi, Ali Ahmadzadeh, Xinyu Xing, and Gang Wang. CADE: Detecting and explaining concept drift samples for security applications. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2327–2344. USENIX Association, August 2021.

[10] Fabian Hinder, Valerie Vaquet, Johannes Brinkrolf, André Artelt, and Barbara Hammer. Localization of Concept Drift: Identifying the Drifting Datapoints. In *International Joint Conference on Neural Networks, IJCNN Padua, Italy 18-23 July, 2022*, 2022.

[11] Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. *Advances in Neural Information Processing Systems*, 32, 2019.

[12] A. Dhurandhar, P. Chen, R. Luss, C. Tu, P. Ting, K. Shanmugam, and P. Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *NeurIPS*, 2018.

[13] S. Wachter, B. D. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *CoRR*, abs/1711.00399, 2017.

[14] Ruth M. J. Byrne. Counterfactuals in explainable artificial intelligence (xai): Evidence from human reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6276–6282. International Joint Conferences on Artificial Intelligence Organization, 7 2019.

[15] Christoph Molnar. *Interpretable Machine Learning.* Lulu.com, 2019. `https://christophm.github.io/interpretable-ml-book/`.

[16] X. Wang, W. Chen, J. Xia, Z. Chen, D. Xu, X. Wu, M. Xu, and T. Schreck. Conceptexplorer: Visual analysis of concept drifts in multi-source time-series data. *2020 IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2020.

[17] Kevin B. Pratt and Gleb Tschapek. Visualizing concept drift. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '03, New York, NY, USA, 2003. Association for Computing Machinery.

[18] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Comput. Surv.*, 46(4), March 2014.

[19] G. Webb, L. Lee, B. Goethals, and F. Petitjean. Analyzing concept drift and shift from sample data. *Data Mining and Knowledge Discovery*, 32, 09 2018.

[20] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *CoRR*, abs/1706.07269, 2017.

[21] A. Artelt and B. Hammer. Convex density constraints for computing plausible counterfactual explanations. *29th International Conference on Artificial Neural Networks (ICANN)*, 2020.

[22] A. Looveren and J. Klaise. Interpretable counterfactual explanations guided by prototypes. *CoRR*, abs/1907.02584, 2019.

[23] R. Poyiadzi, K. Sokol, R. Rodriguez, T. Bie, and P. A. Flach. FACE: feasible and actionable counterfactual explanations. *CoRR*, abs/1909.09369, 2019.

[24] R. Elwell and R. Polikar. Incremental learning of concept drift in nonstationary environments. *IEEE Transactions on Neural Networks*, 22(10), Oct 2011.