ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

On the Fine Structure of Drifting Features

Fabian Hinder, Valerie Vaquet, and Barbara Hammer*

Bielefeld University – Inspiration 1, 33619 Bielefeld – Germany

Abstract. Feature selection is one of the most relevant preprocessing and analysis techniques in machine learning, allowing for increases in model performance and knowledge discovery. In online setups, both can be affected by concept drift, i.e., changes of the underlying distribution. Recently, an adaption of classical feature relevance approaches to drift detection was introduced. While the method increases detection performance significantly, there is only little discussion on the explanatory aspects. In this work, we focus on understanding the structure of the ongoing drift by transferring the concept of strongly and weakly relevant features to it. We empirically evaluate our methodology using graphical models.

1 Introduction

Feature selection and feature relevance analysis [1, 2, 3] are relevant techniques for model and data analysis in machine learning and data science. Machine assistance in and automation of data analysis is particularly relevant when facing time-critical tasks that involve potentially high dimensional data [4, 5, 6, 7]. An important instantiation of this setup is the monitoring of technical systems where human operators have to analyze and understand the changes in the underlying system to initiate appropriate action to keep the system functioning. This is particularly relevant when the underlying system is part of critical infrastructure [7]. A promising way to address this problem is to consider it through the lens of concept drift [8, 9, 10, 4, 11], i.e., a change of the underlying data generating process. Here, a core task is drift detection [10] which is closely related to analyzing and understanding the drift [4]. First works on applying ideas from feature selection to improve the performance of drift detectors indicate a close relation to classical feature selection [12]. However, those lack the explanatory aspect that is of high relevance for monitoring tasks [4, 11]. In particular, an in-depth analysis of the structural properties of drifting features, similar to weakly and strongly relevant features that are linked to graphical models [1] and computational causality [13], is still outstanding.

In this work, we study the fine structure properties of drifting features as introduced in [12]. We extend and deepen the understanding of drifting features to further categorize them into *drift-inducing* features that cause the drift and *faithfully drifting* features that follow along. This allows us to derive important information on the underlying structure of the drift by contrasting the time (in-)dependent functional sub-structures.

This paper is organized as follows: In the first part (Section 2), providing an overview of the related works, we recall the definition of concept drift and

*We gratefully acknowledge funding from the European Research Council (ERC) under the ERC Synergy Grant Water-Futures (Grant agreement No. 951424)

feature relevance. We then extend the ideas that led to the notion of drifting features to enhance the understanding of the effect of drift by analyzing the flow of drift-related information through involved features. This leads to efficient discovery algorithms (Section 3). Finally, we empirically evaluate the resulting algorithm on several datasets based on Bayes networks (Section 4).

2 Problem Setup and Related Work

In this section, we will briefly recall the main definitions of concept drift, drift intensity, and feature relevance including a short overview of the related work.

In the following, we will consider a dataspace $\mathcal{X}$ composed of multiple features $f \in \mathbf{F}$, i.e., $\mathcal{X} = \prod_{f\in\mathbf{F}} \mathcal{X}_f$ for an index set $\mathbf{F}$. We assume that each $\mathcal{X}_f$ is standard Borel, e.g., $\mathcal{X}_f = \mathbb{R}^d$. For a datapoint $X \in \mathcal{X}$ we denote the feature $f \in \mathbf{F}$ by X_f and for a subset $F \subseteq \mathbf{F}$ we write $X_F = (X_f)_{f\in F}$.

To model concept drift we consider a family of probability measures $\mathcal{D}_t$ on $\mathcal{X}$ indexed over a time-domain $\mathcal{T}$, in place of a time-invariant data distribution $\mathcal{D}$ as considered in classical machine learning. *Concept drift* takes place if $\mathcal{D}_t \neq \mathcal{D}_s$ for some $s, t \in \mathcal{T}$ [9] which can be rephrased to a statistical dependence of random variables X and T representing a data and observation time [10].

To analyze the feature-wise effect of drift, [12] suggested to consider the statistic of an idealized drift detector dubbed *(Kullback-Leibler) drift intensity* $I_{\mathcal{D}_t}$. A feature f is drifting if it can increase this quantity, i.e., $I_{\mathcal{D}_t}(F) < I_{\mathcal{D}_t}(F \cup \{f\})$ for some $F \subseteq \mathbf{F}$. It was shown that $I_{\mathcal{D}_t}$ is equivalent to the mutual information of time and the selected features $I_{\mathcal{D}_t}(F) = I(T; X_F)$ [12, Theorem 1].

Similarly, in feature selection, a feature X_f is *relevant* to a target Y if it increases the mutual information, i.e., $I(Y; X_F) < I(Y; X_F, X_f)$, or more commonly if $Y \not\perp\!\!\!\perp X_f \mid X_F$ [1]. This is then split into *strong relevance* when we may choose $F = \mathbf{F} \setminus \{f\}$ and *weak relevance* when a feature is strongly relevant only for a subset of features. Otherwise, the feature is *irrelevant*.

In [12] the similarity of drifting and relevant features was used to derive an efficient algorithmic solution to find all drifting features. In the following, we will study the interaction of drifting features. As it will turn out, this essentially relates to the notion of weakly and strongly relevant features.

3 Fine Structure of Drifting Features

The notion of drifting features describes which features are affected by the drift. Here, we aim for a more elaborated description by introducing the sub-categories of *drift-inducing* – those that introduce the drift into the system – and *faithfully drifting* features – those that follow along with the drift. The differentiation is derived from the following idea: Assume that a set of features X_F are drifting and that another feature X_f can be computed from those, i.e., $X_f = h(X_F, \varepsilon)$ with independent noise ε. Then it is very likely that X_f too is drifting but not because it is affected by the drift itself but rather because the relation to X_F is not affected by the drift. Using causality terminology [13] one may say

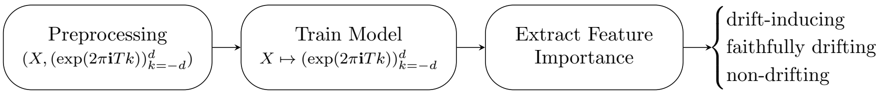
Preprocessing
$(X, (\exp(2\pi \mathbf{i} T k))_{k=-d}^{d})$
Train Model
$X \mapsto (\exp(2\pi \mathbf{i} T k))_{k=-d}^{d}$
Extract Feature Importance
drift-inducing
faithfully drifting
non-drifting

Fig. 1: Drifting Feature Analysis Algorithm.

that drift-inducing features cause the drift while faithfully drifting features are drifting as an effect, yet, a full causal analysis is beyond the scope of this paper.

To phrase this idea in terms of mutual information we start with the observation that a set $N \subseteq \mathbf{F}$ contains non-drifting features only if and only if $I_{\mathcal{D}_t}(J) = I_{\mathcal{D}_t}(J \cup N)$ for all sets $J \subseteq \mathbf{F}$. We generalize this idea to obtain the notion of faithfully drifting features:

Definition 1. A set of features $F \subseteq \mathbf{F}$ is *faithfully drifting* given $I \subseteq \mathbf{F}$, denoted $I \xrightarrow{\perp\!\!\!\perp T} F$, iff adding F cannot make the drift more prominent, i.e., $I_{\mathcal{D}_t}(J) = I_{\mathcal{D}_t}(J \cup F)$ for all $I \subseteq J \subseteq \mathbf{F}$. A set $I \subseteq \mathbf{F}$ is *drift-inducing* iff the remaining features are faithfully drifting given I, i.e., $I \xrightarrow{\perp\!\!\!\perp T} \mathbf{F} \setminus I$. A drift-inducing set of features is *minimal drift-inducing* iff it is minimal wrt. set inclusion.

Faithfully drifting and drift-inducing sets have reasonable properties including transitivity and the desired functional property:

Lemma 1. *Let $F, F_i, I, I_i \subseteq \mathbf{F}$ and $\varepsilon \perp\!\!\!\perp X, T$. It holds (i) transitivity: if $F_1 \xrightarrow{\perp\!\!\!\perp T} F_2$ and $F_2 \xrightarrow{\perp\!\!\!\perp T} F_3$ then $F_1 \xrightarrow{\perp\!\!\!\perp T} F_3$ (ii) union stability: if $I_i \xrightarrow{\perp\!\!\!\perp T} F_i$ then $\bigcup_i I_i \xrightarrow{\perp\!\!\!\perp T} \bigcup_i F_i$ (iii) functionality: it exists a map h such that $X_F = h(X_I, \varepsilon)$ if and only if $I \xrightarrow{\perp\!\!\!\perp T} F$.*

Proof. For (i) and (ii) follow directly from [12, Theorem 1]. For (iii) apply Kuratowski's theorem and then use inverse transform sampling. □

Notice that this implies that there are no synergy effects in faithful drift, i.e., $I \xrightarrow{\perp\!\!\!\perp T} F$ if and only if $I \xrightarrow{\perp\!\!\!\perp T} f \forall f \in F$. Furthermore, this motivates the notion of minimal drift-inducing sets: functionality for projections together with transitivity implies that supersets of drift-inducing sets are drift-inducing and subsets of faithfully drifting sets are faithfully drifting, i.e., if $I_2 \subseteq I_1$, $F_1 \subseteq F_2$, $I_2 \xrightarrow{\perp\!\!\!\perp T} F_2$ then $I_1 \xrightarrow{\perp\!\!\!\perp T} F_1$. In other words, minimal drift-inducing sets can be seen as minimal explanations of the drift. Notice that every drift-inducing set contains a minimal drift-inducing set and that a minimal drift-inducing set does not contain a non-drifting feature.

To determine the minimal drift-inducing sets we can make use of the same algorithmic scheme used in [12] which is shown in Fig. 1, i.e., we can address finding minimal drift-inducing sets using feature relevance for the learning task $X \mapsto T$. The difference is that we also take the distinction between weakly and strongly relevant features into account with strongly relevant features relating to drift-inducing features. This is justified by the fact that if the global time-mean $\mathcal{D}_\mathcal{T}$ is strictly positive, then the intersection of two drift-inducing sets is again drift-inducing and in this case, there is exactly one minimal drift-inducing set which is exactly the set of all strongly relevant features. More formally:

Theorem 1. *Let I_0 be a minimal drift-inducing set and S the set of all strong relevant features for $X \mapsto T$. Then it holds:*

$$S = \bigcap_{\substack{I \subseteq \mathbf{F} \\ \textit{drift-inducing}}} I \subseteq I_0. \tag{1}$$

Furthermore, if $\mathcal{D}_\mathcal{T}$ has a strictly positive density then equality holds in Eq. (1)*. In particular, in this case, $S = I_0$ is the unique minimal drift-inducing set.*

Proof. Use [12, Theorem 1] then the statement follows by weak union and intersection if positive definite (notice that $\mathcal{D}_\mathcal{T} > 0$ suffices). □

However, as already pointed out by [12] one has to be careful as most methods for feature selection are not directly applicable. Yet, as discussed in [2] this problem can be solved by using a suitable preprocessing. In this work, we make use of Fourier preprocessing. Thus, using Theorem 1 and [2, Theorem 2] discovering minimal drift-inducing sets boils down to applying standard MSE-based feature selection to the multi-output regression problem $X \mapsto (\exp(2\pi \mathbf{i} T k))_{k=-d}^{d}$ assuming $T \in [0, 1]$.

4 Experiments

To evaluate our methodology, we conduct a discovery experiment on data drawn from randomly generated linear Bayesian networks with one sudden drift event induced by adding an offset.[1] Notice that standard benchmark drift datasets cannot be used due to the lack of ground truth.

Data generating networks The distribution is given as

$$X_i = \sum_{j<i} a_{ij} X_j + b_j \operatorname{sign}(T) + \varepsilon_i$$

where $T \sim \mathcal{U}([-1, 1])$, $\varepsilon_i \sim \mathcal{N}(0, 1)$ and $T, \varepsilon_1, \ldots, \varepsilon_n$ are independent. The weights a_{ij} are sparsely sampled from a normal distribution with link probability p, i.e., of $j < i$ we have $\mathbb{P}[a_{ij} \neq 0] = p$, for b_i a predetermined number M of non-zero elements is selected.

As pointed out by [1, 14] the drift-inducing features are given by those X_i that are either directly impacted by T, i.e., $b_i \neq 0$, and their parents, i.e., $\exists k : a_{ik} \neq 0 \wedge b_k \neq 0$. The set of all drifting features are given as those that are connected to T by a path, i.e., $\exists k_1, \ldots, k_l = i : (a_{k_j k_{j+1}} \neq 0 \vee a_{k_{j+1} k_j} \neq 0) \wedge b_{k_1} \neq 0$. All other features are non-drifting.

Data and used Methods The networks we consider consist of three disconnected blocks, one with 25, and two with 5 nodes each. Drift is only induced in the first block, the link probability p is the same for all blocks. Note that

[1] Code at `https://github.com/FabianHinder/Analyses-of-Drifting-Features`

Table 1: Results of experiments. ROC-AUC of feature identification task over 100×5 runs (median: m, mean: μ, standard deviation: σ).

		I		C		D	
		m	$\mu\pm\sigma$	m	$\mu\pm\sigma$	m	$\mu\pm\sigma$
$M:1$ $p:0.05$	B	**0.99**	**0.92±0.13**	**0.99**	**0.98±0.05**	**0.73**	**0.78±0.19**
	FI	0.59	0.58±0.31	0.62	0.60±0.34	0.53	0.55±0.24
	MI	0.53	0.55±0.27	0.59	0.58±0.32	0.51	0.53±0.21
	PFI	0.81	0.74±0.26	0.93	0.79±0.28	0.68	0.68±0.22
$M:1$ $p:0.10$	B	**0.97**	**0.90±0.12**	**0.97**	**0.97±0.05**	0.60	**0.66±0.17**
	FI	0.53	0.55±0.30	0.59	0.57±0.34	0.46	0.49±0.18
	MI	0.52	0.55±0.26	0.59	0.58±0.32	0.50	0.52±0.17
	PFI	0.73	0.70±0.25	0.88	0.76±0.27	**0.67**	**0.67±0.16**
$M:1$ $p:0.20$	B	**0.80**	**0.78±0.16**	**0.96**	**0.92±0.11**	0.63	0.62±0.08
	FI	0.55	0.56±0.24	0.62	0.58±0.34	0.42	0.44±0.16
	MI	0.53	0.54±0.21	0.59	0.57±0.30	0.53	0.52±0.12
	PFI	0.65	0.64±0.22	0.82	0.73±0.29	**0.74**	**0.73±0.13**
$M:2$ $p:0.05$	B	**0.85**	**0.86±0.12**	**0.98**	**0.95±0.09**	**0.70**	**0.73±0.15**
	FI	0.56	0.57±0.24	0.58	0.59±0.27	0.50	0.52±0.18
	MI	0.55	0.57±0.21	0.56	0.59±0.26	0.53	0.54±0.16
	PFI	0.74	0.74±0.20	0.85	0.79±0.21	0.67	0.68±0.16
$M:2$ $p:0.10$	B	**0.76**	**0.77±0.13**	**0.97**	**0.93±0.09**	0.60	0.61±0.07
	FI	0.52	0.53±0.21	0.53	0.56±0.27	0.47	0.47±0.13
	MI	0.53	0.54±0.17	0.55	0.58±0.26	0.52	0.52±0.10
	PFI	0.67	0.66±0.18	0.82	0.75±0.22	**0.68**	**0.67±0.11**
$M:2$ $p:0.20$	B	**0.66**	**0.67±0.13**	**0.91**	**0.86±0.13**	0.66	0.66±0.08
	FI	0.50	0.51±0.18	0.53	0.55±0.26	0.41	0.43±0.16
	MI	0.52	0.52±0.14	0.56	0.58±0.24	0.53	0.54±0.12
	PFI	0.62	0.62±0.14	0.74	0.72±0.21	**0.76**	**0.74±0.12**

		I		C		D	
		m	$\mu\pm\sigma$	m	$\mu\pm\sigma$	m	$\mu\pm\sigma$
$M:3$ $p:0.05$	B	**0.83**	**0.84±0.11**	**0.98**	**0.96±0.06**	0.69	**0.72±0.12**
	FI	0.54	0.56±0.21	0.57	0.59±0.24	0.51	0.52±0.16
	MI	0.55	0.57±0.19	0.56	0.60±0.24	0.54	0.55±0.14
	PFI	0.74	0.73±0.16	0.83	0.81±0.17	**0.70**	0.70±0.13
$M:3$ $p:0.10$	B	**0.70**	**0.72±0.11**	**0.94**	**0.90±0.10**	0.62	0.62±0.07
	FI	0.52	0.54±0.17	0.55	0.58±0.24	0.47	0.48±0.13
	MI	0.53	0.54±0.16	0.54	0.59±0.23	0.53	0.53±0.11
	PFI	0.67	0.67±0.14	0.81	0.78±0.18	**0.70**	**0.69±0.11**
$M:3$ $p:0.20$	B	**0.64**	**0.65±0.08**	**0.91**	**0.86±0.11**	0.66	0.66±0.06
	FI	0.49	0.50±0.15	0.52	0.55±0.23	0.43	0.45±0.15
	MI	0.52	0.52±0.13	0.54	0.58±0.23	0.52	0.54±0.14
	PFI	0.63	0.63±0.12	0.78	0.75±0.18	**0.77**	**0.75±0.11**
$M:5$ $p:0.05$	B	**0.79**	**0.79±0.10**	**0.89**	**0.90±0.08**	0.68	0.69±0.07
	FI	0.51	0.54±0.17	0.51	0.56±0.21	0.49	0.51±0.14
	MI	0.53	0.56±0.17	0.53	0.59±0.22	0.54	0.55±0.14
	PFI	0.75	0.74±0.13	0.83	0.82±0.14	**0.72**	**0.71±0.11**
$M:5$ $p:0.10$	B	**0.69**	**0.70±0.09**	**0.86**	**0.85±0.10**	0.65	0.65±0.05
	FI	0.50	0.52±0.15	0.52	0.55±0.20	0.47	0.48±0.13
	MI	0.54	0.56±0.14	0.55	0.59±0.21	0.53	0.54±0.12
	PFI	**0.69**	0.69±0.12	0.79	0.79±0.13	**0.73**	**0.72±0.11**
$M:5$ $p:0.20$	B	0.60	**0.61±0.08**	**0.80**	**0.79±0.10**	0.70	0.69±0.06
	FI	0.49	0.49±0.13	0.48	0.52±0.19	0.42	0.44±0.15
	MI	0.52	0.52±0.12	0.55	0.57±0.19	0.53	0.55±0.15
	PFI	**0.62**	**0.61±0.11**	0.73	0.72±0.15	**0.79**	**0.78±0.10**

a block can by chance consist of disconnected sub-networks. We consider the link probabilities $p = 0.05, 0.1, 0.2$ and number of directly drift-affected features $M = 1, 2, 3, 5$. For each combination of p and M we generate 100 example networks. We take 5 independent samples per network, 500 data points each.

We apply the proposed feature selection algorithm based on extra trees [15] which performed best in [12]. We consider the the native feature importance (FI), permutation feature importance (PFI), Boruta [3] (B) – which builds on FI but extends it by a normalization – and feature-wise mutual information [16] (MI) – which is based on comparing distances of k-th neighbors.

Results We score the feature importances obtained by the different methods using the ROC-AUC as it has the advantage of not being affected by imbalances and does not require defining a decision threshold. Besides the drift-inducing (I) and drifting (D) features we also consider the capability to identify the directly affected features (C; $b_i \neq 0$) which play an important role from a causal perspective [13]. The overall results are presented in Table 1. As can be seen, B performs quite well in the discovery of drift-inducing features, followed by PFI. Interestingly, B performs even better if only the directly affected features are considered. FI and MI perform rather poorly. For MI this can be explained by the fact that the method works feature-wise. We observe that B and FI are negatively affected by a larger number of directly affected features (M) and links (p), for PFI and MI there is no obvious pattern. All methods have problems distinguishing drifting and non-drifting features. This is in line with the findings of [12]. Furthermore, the variance of different samples from the same network is nearly as large as the overall variance.

We thus conclude that B offers a good choice for discovering drift-inducing features. For discovering drifting features PFI seems to be the best option.

5 Conclusion and Further Work

In this work, we analyzed the fine structure of drifting features relating it to the notion of weakly and strongly relevant features. The introduced notions of drift-inducing and faithfully drifting features allow for a better understanding of the inner workings of the ongoing drift. We derived an efficient algorithm to determine drift-inducing features by extending the ideas of [12, 4] in particular answering this open question of [12] in the process. Our considerations revealed a close connection between drifting features and functional graphical models which play an important role in computational causality [13]. This might provide insights into the causal structure of drift which is of high practical relevance. Further considerations in this direction as well as performing comparable analysis on real-world data seem to be interesting and relevant future work.

References

[1] R. Nilsson, J. M. Pena, J. Björkegren, and J. Tegnér. Consistent feature selection for pattern recognition in polynomial time. *JMLR*, 8:589–612, 03 2007.

[2] F. Hinder, J. Brinkrolf, and B. Hammer. Feature selection for trustworthy regression using higher moments. In *ICANN*, pages 76–87. Springer, 2022.

[3] M. B. Kursa and W. R. Rudnicki. Feature selection with the boruta package. *Journal of Statistical Software*, 36(11):1–13, 2010.

[4] F. Hinder, V. Vaquet, J. Brinkrolf, and B. Hammer. Model-based explanations of concept drift. *Neurocomputing*, 555:126640, 2023.

[5] L. Yang, W. Guo, Q. Hao, A. Ciptadi, A. Ahmadzadeh, X. Xing, and G. Wang. {CADE}: Detecting and explaining concept drift samples for security applications. In *USENIX Security 21*, pages 2327–2344, 2021.

[6] K. B. Pratt and G. Tschapek. Visualizing concept drift. In *Proceedings of the ninth ACM SIGKDD*, KDD '03, New York, NY, USA, 2003. Association for Computing Machinery.

[7] V. Vaquet, F. Hinder, J. Vaquet, K. Lammers, L. Quakernack, and B. Hammer. Localizing of anomalies in critical infrastructure using model-based drift explanations. In *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024.

[8] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang. Learning under concept drift: A review. *IEEE TKDE*, 31(12):2346–2363, 2018.

[9] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia. A survey on concept drift adaptation. *ACM Comput. Surv.*, 46(4), March 2014.

[10] F. Hinder, V. Vaquet, and B. Hammer. One or two things we know about concept drift-a survey on monitoring in evolving environments. part b: Locating and explaining concept drift. *Frontiers in Artificial Intelligence*, 7:1330258, 2024.

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

[12] F. Hinder and B. Hammer. Feature selection for concept drift detection. *ESANN. Ed. by Michel Verleysen*, 2023.

[13] J. Pearl. *Causality*. Cambridge university press, 2009.

[14] C. Uhler, G. Raskutti, P. Bühlmann, and B. Yu. Geometry of the faithfulness assumption in causal inference. *The Annals of Statistics*, pages 436–463, 2013.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *JMLR*, 12:2825–2830, 2011.

[16] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Physical review E*, 69(6):066138, 2004.