ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Feature Selection for Concept Drift Detection

Fabian Hinder and Barbara Hammer

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

Abstract. Feature selection is one of the most relevant preprocessing and analysis techniques in machine learning. It can dramatically increase the performance of learning algorithms and also provide relevant information on the data. In online and stream learning concept drift, i.e., the change of the underlying distribution over time, can cause tremendous problems for learning models and data analysis. While there do exist feature selection methods for online learning, to the best of our knowledge there do not exist methods to perform feature selection for drift detection, i.e., to increase the performance of drift detectors and to analyze the drift itself. In this work, we study feature selection for concept drift detection and provide a formal derivation and semantic interpretation thereof. We empirically show the relevance of our considerations on several benchmarks.

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]. Such can be caused by seasonal changes, changed demands, aging of sensors, etc. Since drift might induce severe problems in machine learning models and render performed data analysis useless, it is important to detect and understand the nature and the characteristics of the ongoing drift [3, 4, 5].

Feature selection and feature relevance analysis [6, 7, 8] are relevant techniques for data preprocessing and analysis in machine learning as well as data science. If used when training machine learning models, one can dramatically increase generalization capabilities while reducing training time and the amount of training-data needed. Factors that are of particular relevance in online and stream learning where only limited resources by means of data and computation-time are available [2]. On the other hand, such techniques also provide relevant information about the problem and dataset as well the global model structure. We thus gain relevant insights into the ongoing drift. While there do exist works on explaining concept drift [5, 3, 9] and feature selection for stream learning in the presence of drift [10, 11], only a few works consider both problems at once [9]. Furthermore, performing feature selection for drift detection, e.g., to increase performance, is a relevant, non-trivial problem. Notice that drift detection does not admit a loss function so well-known feature selection methods cannot be applied directly.

In this work, we study feature selection for concept drift detection using a filter-method-like approach. We provide a semantic interpretation of the obtained feature sets. This allows us to derive important information on and about the underlying structure of the drift and potential downstream tasks. This paper is organized as follows: In the first part (Section 2) we recall the definition of concept drift and provide an overview of the related work. We then derive a notion of feature relevance for concept drift detection which we show to be natural, analyze its properties, and discuss its connection to concept drift as a whole as well as feature relevance theory (Section 3). We then empirically evaluate the resulting algorithm on several benchmarks (Section 4).

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

2 Problem Setup

In this section, we will briefly recall the main definitions of concept drift and provide a short overview of related work on feature selection in the presence of drift as well as drift explanations.

2.1 Concept Drift and Setup

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 $\mathbf{F}$. 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 will 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. Hence, $\mathcal{D}_t$ can change over time and concept drift takes place if $\mathcal{D}_t \neq \mathcal{D}_s$ for some $s, t \in \mathcal{T}$ [2]. This idea can be generalized to the statistical dependence of random variables X and T representing data and time, respectively [12]. In the supervised case the notion of real drift, i.e., drift of the conditional distribution $\mathcal{D}_t(Y \mid X) \neq \mathcal{D}_s(Y \mid X)$, becomes relevant [2]. As shown in [13] it is equivalent to conditional dependence of label Y and time T given X. Here, we will make use of this terminology a bit more loosely and apply it to any feature, not just label given data only. Furthermore, by abuse of notation we will denote the (conditional) marginal distribution on the features $F, F' \subseteq \mathbf{F}$ by $\mathcal{D}_t(X_F)$ (and $\mathcal{D}_t(X_F \mid X_{F'})$).

2.2 Related Work

Quite a number of approaches aim for the detection and quantification of drift, its localization in space, visualization, or feature-wise representation [5, 8, 9]. Also, feature selection and feature relevance analysis in the presence of drift has been heavily studied in the past, but usually through the lens of a learning model only [10, 11, 4] which is only loosely connected to drift if considered from an analyst's point of view [13]. To the best of our knowledge, there are no works that deal with the application of feature selection to drift detection algorithms directly. Recently, [5, 8] introduced a new class of drift explanations, called model-based explanations, that use learning models as surrogates in order to compute explanations and thus allow the application of feature relevance analysis to drift directly. As pointed out in [8] the relation between drift and feature importance measures is not well understood yet.

3 Drifting Features

We will now derive and analyze a notion of features that are affected by drift which we will refer to as *drifting features*. As already stated above these are not only relevant for learning with drift [11] but also drift detection [14] and analysis [3]. The naïve way to define drifting features would be to apply the common definition of drift feature-wise. However, this leads to counterintuitive statements if the drift only affects correlations, so that there is drift but no drifting feature. Thus, we have to take the joint distributions into account. Unfortunately, once the joint distribution has drift, adding another feature does not change that. Thus, we must quantify the drift. Inspired by the common use

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

of similar distances in drift detectors we consider the *drift intensity*:

$$I_{\mathcal{D}_t}(F) = \int D_{\mathrm{KL}}\left(\mathcal{D}_t(X_F) \middle\| \int \mathcal{D}_t(X_F)\mathrm{d}\mathbb{P}_T(t)\right)\mathrm{d}\mathbb{P}_T(t), \tag{1}$$

which can be interpreted as "how well does the mean distribution $\int \mathcal{D}_t(X_F)\mathrm{d}\mathbb{P}_T(t)$ approximate the time-point specific distribution $\mathcal{D}_t(X_F)$ on average?". A drifting feature is one that "contributes" to the drift of a suitable marginal distribution and thus makes it easier to detect the drift if included:

Definition 1. Let $\mathcal{D}_t$ be a drift process, i.e., a Markov kernel from $\mathcal{T}$ to $\mathcal{X}$ [12], with features $\mathbf{F}$. We say that X_f with $f \in \mathbf{F}$ is a *drifting feature* iff it can increase the drift intensity, i.e., if there exists $F \subseteq \mathbf{F}$ such that $I_{\mathcal{D}_t}(F) < I_{\mathcal{D}_t}(F \cup \{f\})$.

Some obvious questions are whether the notion of drifting feature depends on the notion of drift intensity specifically, e.g., whether replacing Kullback-Leibler divergence by another divergence measure will result in different drifting features or not, how drift and drifting features are related, and how a decrease in drift intensity can be interpreted. Those are answered by the following theorem:

Theorem 1. *Let $\mathcal{D}_t$ be a drift process and $F' \subseteq F \subseteq \mathbf{F}$ feature sets. It holds (i) $I_{\mathcal{D}_t}$ is non-negative, i.e., $I_{\mathcal{D}_t}(F) \geq 0$, (ii) monotonously increasing, i.e., $I_{\mathcal{D}_t}(F') \leq I_{\mathcal{D}_t}(F)$, and (iii) strict inequality holds if and only if the additional features have real drift given the old ones, i.e., if $\mathcal{D}_t(X_{F\setminus F'} \mid X_{F'})$ has real drift. In particular, $\mathcal{D}_t$ has drift if and only if there exists at least one drifting feature and X_f is a drifting feature if and only if there exists a set of features F such that $\mathcal{D}_t(X_f \mid X_F)$ has real drift. The notion depends on the drift process, only.*

Sketch of proof. Notice that $I_{\mathcal{D}_t}(F) = I(X_F; T)$ where the right-hand side is the mutual information. Due to space restrictions details are left to the reader. □

Notice that many discrepancy measures have similar properties, for instance, [3] showed monotony for the total variation norm for discrete data. Furthermore, many important divergence measures used in drift detection are not increased or at all affected by adding non-drifting features so they can be excluded from drift detection. This includes MMD for various kernels, total variance norm, Wasserstein metric, the feature-wise Kolmogorov-Smirnov statistic, and several classification-based approaches. Hence, the notion of drifting features captures which features are relevant for drift detection.

To determine the drifting features, we make use of the close relation to feature relevance theory [15, 16] when described in terms of conditional independence which is similar to drifting features.

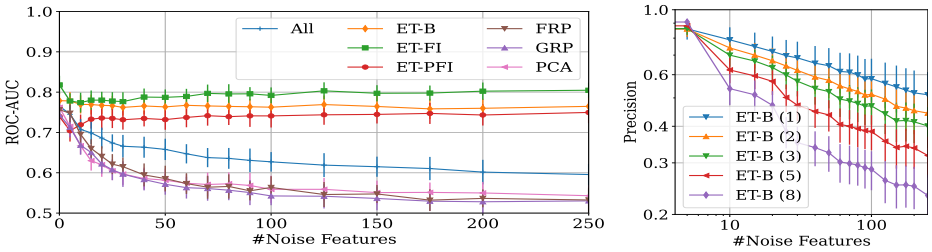
Definition 2. A feature X_f is *relevant* to Y iff X_f and Y are not independent given some set of features R, i.e., $Y \not\perp\!\!\!\perp X_f \mid X_R$, otherwise it is called *irrelevant*.

From this definition, Theorem 1, and [13] it becomes apparent that:

Corollary 1. *A feature X_f is drifting if and only if X_f is relevant to T.*

Thus, a large variety of algorithms for determining drifting features becomes available as finding drifting features can be done by performing feature selection for $X \mapsto T$. However, one has to be careful as most approaches are designed for classification tasks but T is usually a continuous random variable, i.e., a

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.



(a) ROC-AUC of MMD (5 features selected) (b) Precision of selection

Fig. 1: Performance over all datasets (with ≥ 8 features in (b)) and runs vs. number of noise features for some feature selection methods and baselines.

regression task. This can be resolved using appropriate preprocessing [7]. This shows the close connection to the ideas of [8]. Furthermore, it can be shown that a feature can only become (ir-)relevant to a prediction task over time if it is a drifting feature. This provides theoretical justification for online feature selection methods such as [11] and shows the relevance of our considerations to several other downstream tasks.

4 Experiments

We evaluate our ideas empirically by applying the proposed feature selection technique to data-streams that are enriched with additional non-drifting, noise features. We perform two evaluations using the same experimental setup: (1) the performance increase of drift detection and (2) the capability to distinguish drifting and noise features.

Setup: For feature relevance we use Feature Importance (FI), Permutation Feature Importance (PFI), and Boruta [6] (B) based on Random Forests (RF) and Extra Trees (ET) both with Fourier preprocessing of degree 5 [7]. We compare against the baselines of using all features (All), Full Random Projections [14] (FRP), Gaussian Random Projections [17] (GRP), and PCA [17] (PCA). We select a predefined number of features, considering different numbers. We use the following datasets AGRAWAL, LED, MIXED, RandomRBF, RandomTree, SEA, Sine, STAGGER [18], "Electricity market prices" [19], "Forest Covertype" [20], and "Nebraska Weather" [21]. We consider the label as a feature and samples of 250 datapoints with either one abrupt drift in the middle or without any drift, following the setup in [14]. We remove the linear feature correlation and standardize mean and variance. We add different numbers (0-250) of noise features using permuted versions of the original features, comparable to Boruta's shadow features [6]. We run each setup 1.000 times. A summary is given in Fig. 1.

(1) Drift Detection: We use the kernel-two-sample test [22, 17] (MMD; gauss-kernel) and the feature-wise Kolmogorov-Smirnov test [23] (KS) to detect drifts. We use the ROC-AUC on the obtained p-values to evaluate detection performance. To allow better reproducibility we simplify the evaluation by assuming to know the correct candidate time-point, so the problem reduces to a two-sample test [14]. The time-point is *not* available to the feature selection. Regarding detection performance, we find that for All both drift detectors

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Table 1: Results for realworld datasets, noise features, and drift detectors. Table shows ROC-AUC for ET-FI with 1/5/10 selected features and All baseline.

		KS				MMD			
		ET-FI			All	ET-FI			All
DS	#NF	1	5	10		1	5	10	
Forest	0	0.95±0.04	0.99±0.01	1.00±0.00	0.99±0.01	0.92±0.05	0.93±0.04	0.93±0.05	0.92±0.04
	25	0.92±0.06	0.98±0.02	0.99±0.01	0.99±0.01	0.90±0.06	0.88±0.07	0.88±0.05	0.89±0.06
	80	0.93±0.05	0.98±0.02	0.99±0.01	0.99±0.02	0.92±0.05	0.89±0.06	0.88±0.06	0.88±0.06
	250	0.94±0.04	0.99±0.02	0.99±0.02	0.97±0.03	0.93±0.05	0.90±0.06	0.89±0.05	0.85±0.07
Weather	0	0.94±0.04	0.96±0.03	0.96±0.03	0.96±0.03	0.91±0.05	0.92±0.06	0.68±0.12	0.68±0.09
	25	0.90±0.05	0.96±0.03	0.96±0.03	0.94±0.04	0.88±0.07	0.78±0.08	0.70±0.08	0.58±0.08
	80	0.90±0.05	0.95±0.04	0.95±0.04	0.92±0.06	0.89±0.06	0.79±0.07	0.72±0.08	0.55±0.10
	250	0.89±0.06	0.95±0.04	0.96±0.03	0.91±0.05	0.89±0.05	0.80±0.08	0.73±0.09	0.52±0.10

are negatively affected by the number of noise features, but MMD significantly more (see Table 1). This is to be expected as KS operates feature-wise. We observe similar results for FRP, GRP, and PCA which are outperformed by All. If feature selection is used, we do not observe a significant decline in detection performance (see Fig. 1a). Usually, FI and B outperform PFI, the comparison of FI and B is inconclusive. For MMD All is outperformed by all feature select methods and those with fewer selected features perform better. For KS All outperforms all PFI-based approaches but is outperformed by B and FI if an adequate number of features is used. PFI and FI with more selected features perform better, regarding B the results are not clear. The difference between RF and ET seems to be negligible in all cases.

(2) Feature Relevance: We compare the set of affected and selected features, focusing on precision rather than the F1-score to evaluate feature selection as the number of selected features is predefined and recall favors larger selections. As we only increase the negatives, we can directly compare the obtained scores. Here, B outperforms FI which outperforms PFI and ET outperforms RF. Considering the number of noise dimensions we observe a superlinear relationship in the log-log plot with a negative slope (see Fig. 1b). For larger selections performance decreases faster. All strategies perform significantly better than random chance.

5 Conclusion and Further Work

In this work, we adapted the notion of feature relevance and feature selection for drift detection and showed the relevance of this idea to the quality of drift detection as well as the understanding and explanation of concept drift. We provided a formal definition of the notion of drifting features, showed its connection to feature relevance theory, answering the question of [8] on the nature of the observed effects, and provided an efficient algorithmic solution to the problem. The technology provides convincing results for downstream tasks, while the specificity of the detection of features is yet not well understood and leaves room for improvements. Also, the effects of different change points, gradual drift, and specific properties of the dataset like the number and correlation of features are subject to further research. Furthermore, in feature relevance theory there is a further distinction between strong and weak relevant features which can be transferred to the drift setup but have not been considered so far. Further analysis in this direction seems to be interesting for future work.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

References

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

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

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

[4] P. Siirtola and J. Röning. Feature relevance analysis to explain concept drift–a case study in human activity recognition. *arXiv preprint arXiv:2301.08453*, 2023.

[5] F. Hinder, A. Artelt, V. Vaquet, and B. Hammer. Contrasting explanation of concept drift. *ESANN 2022 proceedings*, 2022.

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

[7] F. Hinder, J. Brinkrolf, and B. Hammer. Feature selection for trustworthy regression using higher moments. In *Artificial Neural Networks and Machine Learning – ICANN 2022*, Cham, 2022.

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

[9] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang. Learning under concept drift: A review. *IEEE transactions on knowledge and data engineering*, 31(12):2346–2363, 2018.

[10] J. P. Barddal, H. M. Gomes, and F. Enembreck. A survey on feature drift adaptation. In *2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 1053–1060. IEEE, 2015.

[11] M. Hammoodi, F. Stahl, and M. Tennant. Towards online concept drift detection with feature selection for data stream classification. 2016.

[12] F. Hinder, A. Artelt, and B. Hammer. Towards non-parametric drift detection via dynamic adapting window independence drift detection (dawidd). In *ICML*, 2020.

[13] F. Hinder., V. Vaquet., J. Brinkrolf., and B. Hammer. On the hardness and necessity of supervised concept drift detection. In *Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods - ICPRAM,*, pages 164–175. INSTICC, SciTePress, 2023.

[14] F. Hinder, V. Vaquet, and B. Hammer. Suitability of different metric choices for concept drift detection. In Tassadit Bouadi, Elisa Fromont, and Eyke Hüllermeier, editors, *Advances in Intelligent Data Analysis XX*, pages 157–170, Cham, 2022. Springer International Publishing.

[15] H. H. John, R. Kohavi, and K. Pfleger. Irrelevant features and the subset selection problem. In *Machine learning proceedings 1994*, pages 121–129. Elsevier, 1994.

[16] R. Nilsson, J. Peña, J. Björkegren, and J. Tegner. Consistent feature selection for pattern recognition in polynomial time. *Journal of Machine Learning Research*, 8:589–612, 03 2007.

[17] S. Rabanser, S. Günnemann, and Z. Lipton. Failing loudly: An empirical study of methods for detecting dataset shift. *Advances in Neural Information Processing Systems*, 32, 2019.

[18] J. Montiel, J. Read, A. Bifet, and T. Abdessalem. Scikit-multiflow: A multi-output streaming framework. *Journal of Machine Learning Research*, 19(72):1–5, 2018.

[19] M. Harries, U Nsw cse tr, and New South Wales. Splice-2 comparative evaluation: Electricity pricing. Technical report, 1999.

[20] J. A. Blackard, D. J. Dean, and C. W. Anderson. Covertype data set, 1998.

[21] R. Elwell and R. Polikar. Incremental learning of concept drift in nonstationary environments. *IEEE Transactions on Neural Networks*, 22(10):1517–1531, 10 2011.

[22] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[23] D. M. dos Reis, P. Flach, S. Matwin, and G. Batista. Fast unsupervised online drift detection using incremental kolmogorov-smirnov test. KDD '16, page 1545–1554, New York, NY, USA, 2016. Association for Computing Machinery.