Concept Drift Detection by Tracking Weighted Prediction Confidence of Incremental Learning

PINGFAN WANG

Computer Information and Sciences, Northumbria University Newcastle, United Kingdom

pingfan.wang@northumbria.ac.uk

WAI LOK WOO

Computer Information and Sciences, Northumbria University Newcastle, United Kingdom

wailok.woo@northumbria.ac.uk

NANLIN JIN

Department of computing, Xi'an Jiaotong-liverpool University, China

Nanlin.Jin@xjtlu.edu.cn

DUNCAN DAVIES

Notify Technology Ltd, Newcastle Tyne Upon, United Kingdom

duncan.davies@notifytechnology.com

Data stream mining is great significance in many real-world scenarios, especially in the big data area. However, conventional machine learning algorithms are incapable to process because of its two characteristics (1) potential unlimited number of data is generated in real-time way, it is impossible to store all the data (2) evolving over time, namely, concept drift, will influence the performance of predictor trained on previous data. Concept drift detection method could detect and locate the concept drift in data stream. However, existing methods only utilize the prediction result as indicator. In this article, we propose a weighted concept drift indicator based on incremental ensemble learning to detect the concept. The indicator not only considers the prediction result, but the change of prediction stability of predictor with occurs of concept drift. Also, an incremental ensemble learning based on vote mechanism is especially used to get constantly updated value of indicator. Based on the experiment result on both benchmark and real-world dataset, our method could effectively detect concept drift and outperform other existing methods.

CCS CONCEPTS • Computing methodologies~Artificial intelligence•Information systems~Information systems applications~Data mining~Data stream mining

Additional Keywords and Phrases: concept drift detection, data stream mining, incremental learning, ensemble learning, prediction stability

1 INTRODUCTION

The data stream is now very widely distributed in many real-world scenarios, especially in big data era. Such areas include IoT sensors network, social network and traffic information stream [1]. There are two common characteristics in these areas. The first is unlimited amount of data in the stream and limited memory cannot store all of them. Therefore, data stream mining must be operated online, process the newly input and discard the old data. The second is dynamic evolving environment of data stream, it will bring the concept drift problem.

Concept drift means machine learning model trained on past data will be inaccurate, since the decision boundaries of the model has been changed [2]. For example, a naïve bayes model applied on electricity price to do future price prediction. When new renewable energies (concept drift) added to the grid, the accuracy of the original prediction model will obviously reduce. Conventional machine learning algorithm always suffer from concept drift.

There are two main strategies to deal with concept drift, active and passive [3]. The active strategy will monitor the performance of model trained on data stream and report a concept drift while pre-defined threshold is reached. On the other hand, passive strategy only self-update to adapt the concept drift rather than report it actively.

In this article, we focus on active concept drift detection strategy, aim to detect the drift more accurate and less delay. Similar to concept drift process, there are two kinds of concept drift in data stream, real and virtual[4]. Given a data stream of length n, denotes as $D_{1,n} = [d_1, \cdots, d_n]$, where $d_n = (X_n, y_n)$ represents a single data sample in data stream $D_{1,n}$. X_n and y_n are the feature and class label correspondingly. If there is no concept drift in data stream, then its joint distribution $JD_{1,n}(X, y)$ will keep stable. However, when concept drift happens in time $t+1$, $JD_{1,t}(X,y) \neq JD_{t,\infty}(X,y)$, denotes as $\exists t, P_t(X,y) \neq P_{t+1}(X,y)$ while P_t represents the joint distribution of (X, y) at time t[5].

Based on bayes theory, $P(X,y) = P(X) \times P(y|X)$, real concept drift refers to $P_t(y|X) \neq P_{t+1}(y|X)$ when $P_t(X) = P_{t+1}(X)$, indicating the change of conditional probability [6]. Virtual concept drift is the opposite, $P_t(X) \neq P_{t+1}(X)$ when $P_t(y|X) = P_{t+1}(y|X)$, which only indicates the change of marginal probability. We focus on real concept drift.

In this paper, we propose an incremental weighted performance drift detection method (IWPDDM) by monitoring proposed weighted and combined indicator. The principal idea is first to construct the indicator, which chooses not only the accuracy as one of the indicators, but also considers the prediction stability. This is because the latter contains attributes related to precision than accuracy. Statistics is used to define two thresholds for concept drift detection and validation. Furthermore, an incremental vote-based ensemble predictor is developed on data stream to make prediction and it is capable to self-update with each data input. Finally, the update indicator will be triggered when concept drift occurs, and it resets the predictor.

The main contributions of this study are summarised as follows:

- To the best of our knowledge, this paper is the first to introduce prediction stability to concept drift
- Defining the weighted prediction stability and its calculation rule under vote-based ensemble learning
- Developing statistics to configure the threshold of concept drift detection and validation
- Developing an incremental vote-based ensemble learning strategy on data stream mining to deal with concept drift detection

2 LITERATURE REVIEW

This section mainly includes two aspects: the first is concept drift detection with different strategies, the other is research works about prediction stability.

2.1 Concept drift

Drift detection refers to the methods for determining whether drifts has occurred. The changes in the environment typically cause concept drift, making the previously fitted machine learning model prone to inaccuracy. The performance of the model will be influenced before and after the concept drift.

There has been some works done for concept drift detection. In terms of the statistical methods used, these works can be spited into three categories: (1) performance-based (2) data distribution-based and (3) hierarchical structure [7].

The performance-based methods mainly detect the drift by monitoring the performance change of the machine learning model. Drift is detected when these is a statistically significant rise or reduction. The Drift Detection method (DDM) [8] is the most well-known work which tracking the overall error rate of the machine learning model. The Early Drift Detection Method (EDDM)[9], which is slightly different from DDM, identifies drift by measuring the distance between two successive prediction errors. It won't work in some circumstances until there are 30 prediction mistakes, which is impractical because the drift happens quickly. Similar methods have been used to drift detection using Hefting's inequality-based Drift Detection Method(HDDM) [10], which uses Hoeffding's inequality to determine the crucial area of a drift. Furthermore, Adaptive WINdowing (ADWIN) [11], a moving window-based approach, identifies drift by measuring the difference between two adaptive size windows. By comparing the recorded values and their mean up to the current instant, Page-Hinkley(PH) [12] finds the drift.

The dissimilarity of new data and previous data is measured with distance function in data distribution-based method. The statistically significantly difference reflect the occurs of concept drift. Its advantage is to process both labelled and unlabelled data. The Kolmogorov–Smirnov (KS) test is used to identify static drift in unlabelled data [13]. Aside from the KS test, the clustering approach is also widely employed, with K-nearest neighbours being used to create data clusters for density estimation[14]. Hierarchical structure methods are increasing recent years. They normally have two layers or more, these layers with increasing stricter rules to identify the concept drift. The Hierarchical Change-Detection Tests (HCDTs) [15] applies two layers structure to detect concept drift. Hierarchical Linear Four Rates (HLFR) following similar structure to detect drift with more detection indicators.

2.2 Prediction stability

Quality of the machine learning models is measured in terms of outcome-oriented predictive monitoring techniques [16]. Such as precision, recall and Area Under the ROC Curve (AUC). However, [17] argues that these are not sufficient to measure predictive performance. For example, the predictor is easy to suffer from new seeing data, which will bring unacceptable consequences in areas with high accuracy requirements, such as healthcare and public transport. Therefore, the stability of a classifier is very important when used to make successive prediction. In this article, we use the distribution of number of votes on each class, the degree of the distribution reflects the prediction stability.

3 WEIGHTED PREDICTION CONFIDENCE

We propose an incremental weighted performance drift detection method (IWPDDM) to address the concept drift problem in data stream. The proposed IWPDDM has three steps i.e., detection and validation.

3.1 Design of IWPDDM

The proposed IWPDDM has three steps, Incremental learning, Detection and Validation. The first incremental learning step is executed in real time, incrementally to make prediction and update itself with new data sample input. With the prediction result and confidence from first step, weighted incremental indicator is constructed and updated. Then the detection step and validation step will be successively triggered when weighted incremental indicator trigger corresponding thresholds. The IWPDDM will report the drift and reset the incremental classifier.

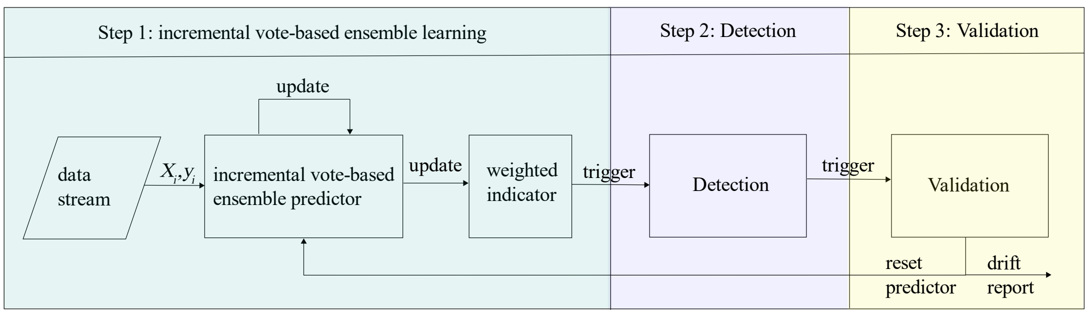

Figure 1: Design of IWPDDM, include three steps, incremental vote-based ensemble learning, detection and validation respectively. In step 1, weighted indicator will be updated and monitored. Once corresponding statistical thresholds are reached, step 2 and 3 will be triggered successively. When validation step validate a concept drift, signal for predictor reset will be sent, and drift will be reported synchronous.

3.2 Incremental vote-based ensemble learning step

The aim of incremental vote-based ensemble learning step is to dynamical train the model to make prediction with upcoming data sample. The performance of incremental ensemble learning model will change when a concept drift occurs. The prediction accuracy of the model is considered as first indicator, because of its general representation of the model prediction performance. On the other hand, accuracy is inaccuracy since it can only generate binary decision, true or false.

Motivated by [17], we define term called "prediction stability (PS)" of incremental learning algorithm for data stream mining. Normally, a learning algorithm is considered unstable if small changes the training set can cause significant changes in the predictor. In this paper, we use incremental voted-based ensemble learning algorithm as the predictor. Therefore, change of PS can reflect the impact of concept drift on incremental learning model.

We consider the weighted sum ($S_{1,t}$) of accuracy value and prediction stability as the indicator of the performance over time period from 1to t, which shows in equation (1):

$$S_{1,t} = W_1 \times AC_{1,t} + W_2 \times PS_{1,t} \quad (1)$$

Where $AC_{1,t}$ is average value of accuracy of incremental learning model from starting time point 1 to t. $PS_{1,t}$ is the average value in same way. W_1 and W_2 is the weight value for $AC_{1,t}$ and $PS_{1,t}$ respectively. To make our contribution more focus, strategy of weight value allocation always use average weight value. $AC_{1,t}$ is the rate of number of correct prediction to time period from start to time point t. $PS_{1,t}$ is the average value of PS the from staring time to t, shows in equation (2):

$$PS_{1,i} = \frac{\sum_{i=1}^{t} PS_i}{t} \tag{2}$$

Where PS_i represends the degree of the prediction stablitiy of the predictor at time point I, which is calculated by equation (3):

$$PS = \frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\left|NC_i \times H_i - NC_j \times H_j\right|}{\sum_{i=1}^{N-1} H_i + \sum_{j=i+1}^{N} H_j} \times N \tag{3}$$

Where NC_i represents the stability of the predictor on class i and H_i is corresponding weight value. As mentioned before, the weight strategy always use average weight, so $H_i = H_j = 1/N$. N is the number of the class in data stream, e.g. N equal to 2 in binary classification task. NC_i is calculated by following equation (4):

$$NC_i = \frac{Number\ of\ \ classifiers\ vote\ for\ class\ k}{Number\ of\ all\ classifiers} \tag{4}$$

Since the predictor used in this paper is vote-based ensemble learning model, distribution of the vote number on each class reflects prediction result. For example, in binary classification task, $A_1 = A_2 = 0.5$ represent the worst PS, since it is almost the random classification PS value equal to 0 based on equation (3). Then the highest PS value 1 will reach when A_1 equal to 1or 0. Therefore, the value interval of PS is [0, 1], represents the degree of prediction stability from highest to random.

Based on equations (1)-(4), our proposed indicator for drift detection is the weighed sum ($S_{1,t}$) of accuracy($AC_{1,t}$) and prediction stability ($PS_{1,t}$). The generation of the $S_{1,t}$ is shown in Algorithm 1.

ALGORITHM 1: incremental vote-based ensemble predictor learning and generation of the $S_{1,t}$

Input: data stream(X_i, y_i), initial incremental vote-based ensemble predictor M
Output: $S_{1,t}$
$AC_{1,t} = 0$, $PS_{1,t} = 0$,
while data stream continues, do
    M partial training with coming data
    M update and make prediction
    for prediction result on each class, do,
        calculate overall AC_t and PS_t (based on equation 3 and 4) at time t
    end
    calculate $AC_{1,t}$ and $PS_{1,t}$ based on equation 2
    calculate $S_{1,t}$ based on equation 1
end

In Algorithm 1, incremental vote-based ensemble learner is initialled and dynamic trained with upcoming data, at the same time to make prediction on input data. Then the value of AC_t and PS_t is calculated on time point t. In the end, indicator $S_{1,t}$ is calculated with the available of $AC_{1,t}$ and $PS_{1,t}$.

3.3 Detection and validation step

Change of the indicator $S_{1,t}$ reflects the performance change of predictor. If there is no concept drift occurs, it will gradually increase until certain amount of data and then keep stable. However, accuracy and prediction stability, two components in $S_{1,t}$, will both drop when there is a concept drift.

To define the degree of threshold of concept drift detection. Three-sigma rules is used, 2(95%) and 3(99.7%) standard deviations corresponding to detection threshold and validation threshold respectively. Therefore the standard deviation of the indicator $S_{1,t}$ is $sd(S_{i,t}) = sqrt(S_{1,t} \times (1 - S_{1,t}))$.

ALGORITHM 2: detection and validation step

```
Input: indicator value S1,t
Output: drift point
drift flag = False
Smax = 0, sd(S)max = 0
while S1,t continues, do
        calculate its standard deviation sd(S1,t)
        if S1,t + sd(S1,t) > Smax + sd(S)max, do
                Smax = S1,t
                Sd(S)max = sd(S1,t)
        end
        If S1,t + sd(S1,t) > Smax + 2 × sd(S)max, do
                drift_flag = Detection
        end
        If S1,t + sd(S1,t) > Smax + 3 × sd(S)max, do
                drift_flag = Validation
        end
return drift_flag
end
```

In Algorithm 2, the max value of the $S_{1,t}$ and its standard deviation will update incrementally. At the same time, the drift_flag will be updated when corresponding threshold is reached. When the drift_flag is set to Validation, it means concept drift is validated and predictor should be reset for retraining from scratch.

4 EXPERIMENT AND EVALUATION

This section compares the proposed IWPDDM with the six existing drift detection methods, over four benchmark datasets and one real-world dataset. Section 4.1 describes the datasets and drift detection methods used in experiment. Section 4.2 analyzes the experimental results and compares with other 6 detection methods.

4.1 Datasets and drift detection methods

Four benchmark data streams widely used in drift detection literature are selected for experiment and evaluation, namely SEA, SINE, AGRAWAL, and LED. They are available at public stream data mining platform[18]. Descriptions of the datasets are in Table 1. Each data stream has 1000 time points, the concept drift occurs at time point (CDT) 500. To compare the performance of drift detection method, we test the methods on each dataset for 100 trials which are generated with different random seeds.

Table 1: description of the datasets

Dataset	No. of attributes	No. of classes
SEA	3	2
SINE	2	2
AGRAWAL	9	2
STAGGER	3	2

The real-world dataset used in this paper is Posture dataset, which is collected at 0.1s intervals(*u*), from sensors on ankle, belt and chest. Its class label is made by a human observer. The patterns of each data stream of posture dataset are clearly different, before and after 500 *u*. So, the concept drift point in Posture is time point 500.

Six drift detection methods are accessible in skmultiflow [18]. We use the parameters recommended by skmultiflow for each drift detection method. The detailed parameters are shown in Table 2.

Table 2: methods parameters

Algorithms	Parameters
IWPDDM	$\alpha = 2, \beta = 3$
DDM	$\alpha = 0.05, \beta = 0.01$
Adwin	$\delta = 0.05$
EDDM	$\alpha = 0.9, \beta = 0.95$
PH	$threshold = 30, \delta = 0.005$
HDDM_A	$\alpha = 0.005, \beta = 0.001$
HDDM_W	$\alpha = 0.05, \beta = 0.01, \lambda = 0.05$

4.2 Experiment result and evaluation

We use E_i to represent an independent experiment, which will run 100 times on both benchmark datasets and real-world dataset. Concept drift point will be reported if drift detection successfully detects the drift. To compare the performance of IWPDDM with other 6 drift detection methods, we use drift detection accuracy and delay.

The drift detection accuracy measures how the drift detection method works with concept drift. To compare it with other 6 drift detection methods, in terms of the ratio of the number of drifts detected data stream to the total 100 trails. The higher the accuracy, the better the performance.

Table 3: drift detection accuracy (the unit is %), our proposed IWPDDM method reach highest average value over all the drift methods

Methods	IWPDDM	DDM	ADWIN	EDDM	PH	HDDM_A	HDDM_W
SEA	89	75	13	96	5	18	71
SINE	99	93	51	100	55	71	96
AGRAWAL	95	50	14	96	37	47	94
STAGGER	94	86	84	80	76	86	91
POSTURE	96	94	94	94	82	95	91
Average	**94.6**	79.6	51.2	93.2	51	63.4	88.6

From the result of drift detection accuracy (Table 3), our proposed IWPDDM has the highest value among all the methods.

On the other hand, we also compare the performance of drift detection methods based on drift detection delay. drift detection delay refers to the distance between detected drift point and CDT. The smaller value of delay indicates the better drift detection performance.

Then the figure 1 describe the result of the drift detection delay. From the boxplot, it is clear that our proposed IWPDDM has the lowest value of drift detection delay among all the methods.

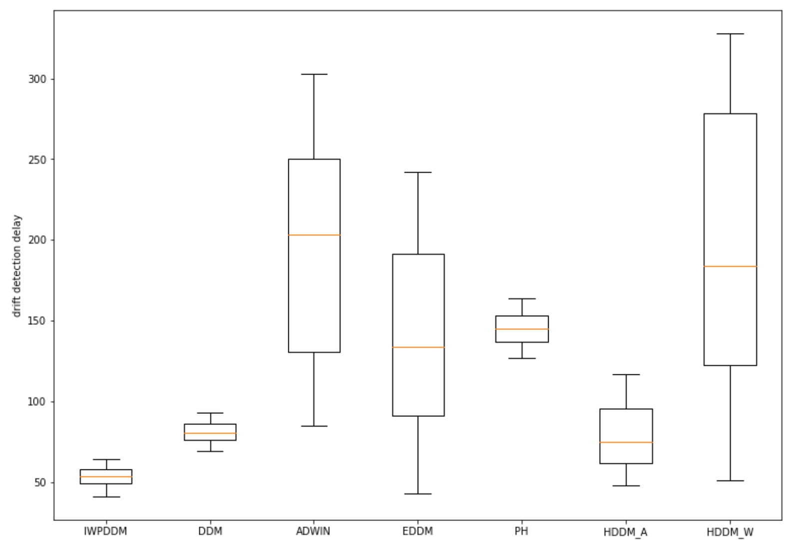

Figure 1: Drift detection delay result, x-axis represents the methods used for comparison and y-axis is the drift detection delay. From the boxplot, our proposed IWPDDM method has the lowest delay among all the drift detection methods

Based on the obtained result of drift detection accuracy and delay, it is evidenced that our proposed IWPDDM has led to the best performance of drift detection among other methods by highest average drift detection accuracy, 1.4% improvement compare to second method EDDM, and IWPDDM also have lowest drift detection delay, which is much smaller than others.

5 CONCLUSION

This paper presents an incremental weighted performance drift detection method (IWPDDM) for data stream mining. It addresses the concept drift issue by monitoring the performance of predictor incrementally trained on data stream. The prediction accuracy and stability are both used to construct the weighted indicator for the concept drift. The weighted indicator will then also update with incremental learning model. Once the indicator reaches a pre-defined statistical threshold, the concept drift will be reported immediately. The comparative studies have shown that the proposed IWPDDM performs best among the other six established detection methods, with better detection accuracy and delay. For future work, redesigning the weight of the indicator and prediction stability would be an interesting direction. Another valuable further work may to consider the scenarios with more classes in data stream.

ACKNOWLEDGMENTS

This work was supported by the European Regional Development Fund (ERDF) [25R17P01847]; Northumbria University Newcastle, Newcastle Tyne Upon, UK; Notify Technology Ltd, Newcastle, UK.

REFERENCES

[1] Paweł Zyblewski, Robert Sabourin, Michał Woźniak, Preprocessed dynamic classifier ensemble selection for highly imbalanced drifted data streams, Information Fusion, Volume 66, 2021, Pages 138-154, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2020.09.004.

[2] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama and G. Zhang, "Learning under Concept Drift: A Review," in IEEE Transactions on Knowledge and Data Engineering, vol. 31, no. 12, pp. 2346-2363, 1 Dec. 2019, doi: 10.1109/TKDE.2018.2876857.

[3] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. 2014. A survey on concept drift adaptation. ACM Comput. Surv. 46, 4, Article 44 (April 2014), 37 pages. DOI:https://doi.org/10.1145/2523813.

[4] Ning Lu, Jie Lu, Guangquan Zhang, Ramon Lopez de Mantaras, A concept drift-tolerant case-base editing technique, Artificial Intelligence, Volume 230, 2016, Pages 108-133, ISSN 0004-3702, https://doi.org/10.1016/j.artint.2015.09.009.

[5] Fang Liu, Yanwei Yu, Peng Song, Yangyang Fan, Xiangrong Tong, Scalable KDE-based top-n local outlier detection over large-scale data streams, Knowledge-Based Systems, Volume 204, 2020, 106186, ISSN 0950-7051, https://doi.org/10.1016/j.knosys.2020.106186.

[6] S. Wang, L. L. Minku and X. Yao, "A Systematic Study of Online Class Imbalance Learning with Concept Drift," in IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 10, pp. 4802-4821, Oct. 2018, doi: 10.1109/TNNLS.2017.2771290.

[7] Bartosz Krawczyk, Leandro L. Minku, João Gama, Jerzy Stefanowski, Michał Woźniak, Ensemble learning for data stream analysis: A survey, Information Fusion, Volume 37, 2017, Pages 132-156, ISSN 1566-2535, https://doi.org/10.1016/j.inffus.2017.02.004.

[8] J. Gama, P. Medas, G. Castillo, P. Rodrigues, Learning with drift detection, Brazilian symposium on artificial intelligence, Springer, 2004, pp. 286–295.

[9] M. Baena-Garcıa, J. del Campo-´Avila, R. Fidalgo, A. Bifet, R. Gavalda,R. Morales-Bueno, Early drift detection method, Fourth international workshop on knowledge discovery from data streams, Vol. 6, 2006, pp. 77–86.

[10] I. Frías-Blanco, J. d. Campo-Ávila, G. Ramos-Jiménez, R. Morales-Bueno, A. Ortiz-Díaz and Y. Caballero-Mota, "Online and Non-Parametric Drift Detection Methods Based on Hoeffding's Bounds," in IEEE Transactions on Knowledge and Data Engineering, vol. 27, no. 3, pp. 810-823, 1 March 2015, doi: 10.1109/TKDE.2014.2345382.

[11] A. Bifet, R. Gavalda, Learning from time-changing data with adaptive windowing, in: Proceedings of the 2007 SIAM international conference on data mining, SIAM, 2007, pp. 443–448.

[12] Jiarong Li, Haijun Jiang, Xuehui Mei, Cheng Hu, Guoliang Zhang, Dynamical analysis of rumor spreading model in multi-lingual environment and heterogeneous complex networks, Information Sciences, Volume 536, 2020, Pages 391-408, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2020.05.037.

[13] Sobolewski, Piotr, and Michal Wozniak. "Concept Drift Detection and Model Selection with Simulated Recurrence and Ensembles of Statistical Detectors." J. Univers. Comput. Sci. 19.4 (2013): 462-483.

[14] Denis Moreira dos Reis, Peter Flach, Stan Matwin, and Gustavo Batista. 2016. Fast Unsupervised Online Drift Detection Using Incremental Kolmogorov-Smirnov Test. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 1545–1554. DOI:https://doi.org/10.1145/2939672.2939836.

[15] C. Alippi, G. Boracchi and M. Roveri, "Hierarchical Change-Detection Tests," in IEEE Transactions on Neural Networks and Learning Systems, vol. 28, no. 2, pp. 246-258, Feb. 2017, doi: 10.1109/TNNLS.2015.2512714.

[16] Irene Teinemaa, Marlon Dumas, Marcello La Rosa, and Fabrizio Maria Maggi. 2019. Outcome-Oriented Predictive Process Monitoring: Review and Benchmark. ACM Trans. Knowl. Discov. Data 13, 2, Article 17 (June 2019), 57 pages. DOI:https://doi.org/10.1145/3301300.

[17] Irene Teinemaa, Marlon Dumas, Anna Leontjeva, and Fabrizio Maria Maggi. 2018. Temporal stability in predictive process monitoring. Data Min. Knowl. Discov. 32, 5 (September 2018), 1306–1338. DOI:https://doi.org/10.1007/s10618-018-0575-9.

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

Authors' background

Your Name	Title*	Research Field	Personal website
Pingfan Wang	PhD candidate	Machine learning, data stream mining	https://researchportal.northumbria.ac.uk/en/persons/pingfan-wang
Wai Lok Woo	Full Professor	Machine Learning, Signal Processing, Artificial Intelligence	https://www.northumbria.ac.uk/about-us/our-staff/w/wai-lok-woo/
Nanlin Jin	Associate professor	Data mining, Artificial Intelligence of Things (AIoT), Evolutionary Game Theory	https://www.xjtlu.edu.cn/en/departments/academic-departments/computer-science-and-software-engineering/staff/nanlin-jin
Duncan Davies	Co-founder and CEO of Notify Technology Ltd	Smart Health Safety software	https://www.linkedin.com/in/duncan-davies/

*This form helps us to understand your paper better, the form itself will not be published.

*Title can be chosen from: master student, Phd candidate, assistant professor, lecture, senior lecture, associate professor, full professor