An Impact Study of Concept Drift in Federated Learning

Yang, Guanhui; Chen, Xiaoting; Zhang, Tengsen; Wang, Shuo; Yang, Yun

Document Version
Peer reviewed version

Citation for published version (Harvard):
Yang, G, Chen, X, Zhang, T, Wang, S & Yang, Y 2023, An Impact Study of Concept Drift in Federated Learning. in *The IEEE International Conference on Data Mining (ICDM).*

An Impact Study of Concept Drift in Federated Learning

Guanhui Yang, Xiaoting Chen, Tengsen Zhang
National Pilot School of Software
Yunnan University
Yunnan, China

Shuo Wang
School of Computer Science
The University of Birmingham
Birmingham, UK
s.wang.2@bham.ac.uk

Yun Yang
National Pilot School of Software
Yunnan University
Yunnan, China
yangyun@ynu.edu.cn

***Abstract*—Federated learning (FL) is a rising distributed machine learning area, which aims to train a high-performing global model with data collected from a number of local clients. Many FL applications receive data over time in the form of data streams. Streaming data are likely to suffer concept drift. It can significantly harm a model's predictive ability. However, no study has characterized concept drift in FL or investigated how it can affect the global and local models' performance. This paper aims to provide such understanding by 1) categorizing concept drift in temporal and spatial dimensions with ten features and 2) investigating the impact of the features in depth. We find that: the temporal features degrade FL models to a different extend and do not affect model convergence after the new data concept becomes stable; the spatial features cause data heterogeneity and affect both accuracy and convergence speed.**

***Index Terms*—federated learning, concept drift, data stream learning**

I. Introduction

Federated learning (FL) is a recent and fast-developing machine learning (ML) area, designed to learn from distributed data from local clients without centralising them. It has raised great interest from industry, in particular the edge computing and healthcare fields, where large-scale data are generated within a modern distributed network, coupled with concerns over data privacy and transmission security. Data generated from these types of applications are mostly in the form of data streams, and tend to suffer from data non-stationarity. For example, environment sensor data are affected by human behaviours and seasons; traffic data vary with the time of the day, road conditions, etc. [1], [2].

Data non-stationarity is hardly touched upon in FL. Existing FL algorithms assume that client data has a stable distribution, so that the global model trained from the past remains valid for future data. Nevertheless, this assumption does not often hold in real-world applications. A non-adaptive model trained under a false stationarity assumption is bound to become obsolete over time. In addition, distribution changes can cause data heterogeneity among clients over time, which has been proved to be a performance hindrance to FL algorithms [3].

In classification tasks, concept drift is the main type of changes in data distributions. It is said to occur when the joint probability $P\left(\mathbf{x}, y\right)$ of data changes, where $\mathbf{x}$ is the input feature vector and y is the target label/class [4]. Concept drift has been discussed extensively in the traditional ML. There exist active drift detectors (e.g. DDM [5], ADWIN [6]) and passive approaches that evolve a learning model continuously [1]. However, none of these approaches can be directly used in FL. Clients in FL do not share data with the global model. Moreover, concept drift manifests three statistical forms at various speeds, severities, frequencies, etc. Concept drifts with different features require corresponding drift detection and adaptation strategies for the best performance [7].

Some very recent papers have realised the importance of overcoming concept drift in FL systems, detailed in the next section. Unfortunately, none of them discriminate concept drift systematically. These solutions may thus become less effective. In addition to the temporal features of concept drift that pose performance risks to FL models, new challenges arise in the spatial dimension. Across multiple clients, concept drift can affect them differently. So far, it is unclear *whether, when and how the temporal and spatial features of concept drift affect the performance of the global and local FL models*. It is impossible to develop effective solutions without such fundamental understanding.

With this question in mind, this paper aims to find out the impact of concept drift in FL. Firstly, we propose the very first categorization of concept drift in FL to characterise concept drift in temporal and spatial dimensions. Secondly, we investigate three temporal features (i.e. drifting form, speed, severity) and two spatial features (i.e. drift coverage and synchronism) in depth. It provides insights into how concept drift should be treated appropriately and guidance on developing targeted solutions. For the wide application and simplicity, horizontal FL is the focus of this work, where client data share the same feature space.

II. Related Work

In this section, we review the research progress on concept drift in traditional data stream learning and FL.

A. Data Stream Learning and Concept Drift

Data stream learning is a ML area aiming at real-time learning and prediction with time varying data streams. Con-

The source codes and the datasets in this paper are available at: https://github.com/TreeWIA/FLStream cept drift is a key challenge, referred to as a change in data distributions [2]. There exist three fundamental forms of concept drift corresponding to the three major variables in Bayes' theorem: $P(\mathbf{x})$ virtual drift, $P(y)$ prior probability drift and $P(y|\mathbf{x})$ posterior probability drift or real drift. Only the posterior probability drift causes true classification boundary changes, where the model must be adapted to the new concept. Although the other two forms do not affect the true boundary, they may cause learning bias in the model. Because the underlying reasons for performance degradation are different, each form of drift requires targeted solutions. Concept drift has also been characterised by changing speed, severity, recurrence, frequency and predictability [8], bringing in additional challenges to be considered when developing solutions. These factors form the temporal features of a concept drift, which can be used to describe any distribution change in a single data stream.

Existing concept drift work focuses on drift detection and drift adaptation. Drift detection algorithms report the timing when a drift occurs. For example, the widely used DDM [5] defines warning and drift levels by monitoring the model's error rate. Drift adaptation methods adapts the predictive model to the new concept and maintain its performance. Key techniques include simple retraining, ensemble retraining, and model adjustment.

B. Concept Drift in FL

There is limited research on the concept drift problem in FL. The existing work includes passive adaption approaches (CFL [9], FedDC [10] and Adaptive-FedAvg [11]) and approaches with active drift detection (FedConD [12], CDA-FedAvg [13], the drifting-node isolation method [14], and the most recent FedDrift-Eager and FedDrift algorithms [15]).

The common strategies of these approaches can be summarized as 1) comparing local data distributions or model parameters, 2) isolating drifted clients via regularization or the learning rate or 3) training more than one global model. None of the above work explicitly differentiate between drift in time and space. For example, CDA-FedAvg assumes that the same virtual concept drift occurs to all the clients at the same time. FedConD only considers sudden and gradual two types of drift differing in speed. This paper will fill in this gap by providing a systematic and in-depth study.

III. Categorization of Concept Drift in FL

In this section, we propose the first taxonomy that clearly categorizes and describes concept drift in FL in temporal and spatial dimensions. We propose a set of features that fall into two categories – the temporal and spatial features (see Fig. 1). The temporal features are those that exist in a single data stream, as described in Section II-A, including changing form, speed, severity, recurrence, frequency and predictability [4] [8]. The spatial features are unique to FL. They describe how concept drift occurs among the clients. We propose the following four new spatial features: coverage, synchronism, direction and correlation. A concept drift may not affect all the clients at the same time; for example, the monitored traffic jam eventually spreads to other sensors. Thus, we need coverage and synchronism to describe the scenarios. Clients can be affected by different drifts; for example, one type of disease becomes more frequent in one region, which however becomes less frequent in another region. These drifts could be the same form of change but in different directions. They can be either independent or correlated. A detailed explanation for each feature is given below.

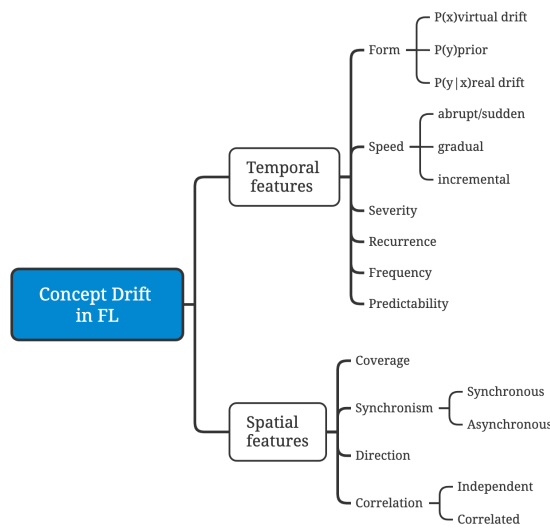


Fig. 1: Categorization of concept drift in FL.

The temporal features use form, speed and severity to characterise a single concept drift (i.e. isolated drift) in any data stream. If there are more than one (i.e. drift sequence), recurrence, frequency and predictability are used to describe the changing pattern. All the features (except "form") are further illustrated in Fig. 2, where the horizontal direction is the time steps and the vertical direction shows how the drift occurs in 9 cases. Case 1 is a data stream without any drift.

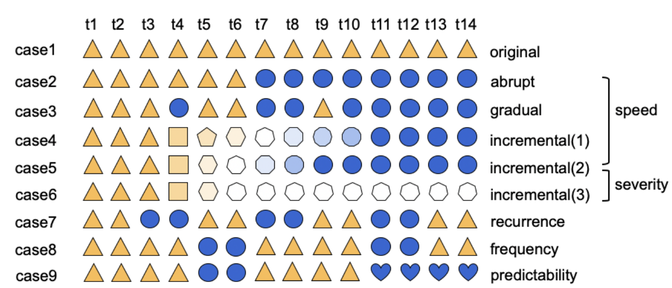
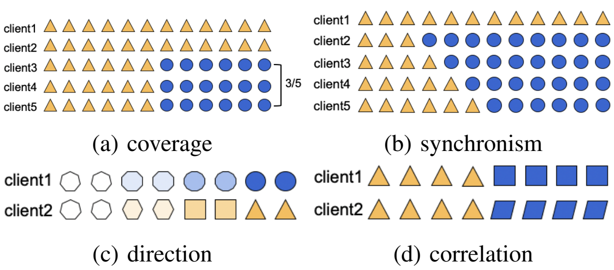


Fig. 2: Temporal features of concept drift.

1) Form: describes how data instances move within or across classification boundaries, including $P(\mathbf{x})$, $P(y)$, $P(y|\mathbf{x})$.
2) Speed: describes how fast the current data concept changes to a new concept. An abrupt/sudden drift occurs very quickly without having any intermediate or recurrent concept (case 2). If the new concept gradually takes over, this can be a

'gradual' drift or an 'incremental' drift. A 'gradual' drift is a probabilistic change, where the old concept becomes less and less frequent until it disappears (case 3). An 'incremental' drift is the scenario where the old concept moves towards the new concept, with intermediate concepts [4] generated along the way. Cases 4 and 5 illustrate two incremental changes at a different speed.
3) Severity: describes the degree of concept drift or the distance between the old and new concept. Case 5 is a more severe change than case 6.
4) Recurrence: the cases with returns to previous concepts are called "recurrent" drift (case 7).
5) Frequency: describes how often a drift occurs in the data stream. Case 8 is less frequent than case 7.
6) Predictability: describes whether a drift is predictable (i.e. randomness). Case 9 has a random change from triangle to heart shapes.

The spatial features describe concept drift across clients. They trigger another widely discussed topic in FL – data heterogeneity, a key factor of global performance of FL models. They are illustrated in Fig. 3.



(a) coverage (b) synchronism
(c) direction (d) correlation

Fig. 3: Spatial features of concept drift.

7) Coverage: describes how many clients are affected by concept drift around the same time. In Fig. 3a, 3 out of 5 clients suffers from drift (changing from triangle to circle).
8) Synchronism: describes whether concept drift occurs at the same time. In Fig. 3a, clients 3-5 suffer synchronous drift; in Fig. 3b, the drift affects clients 2-5 asynchronously.
9) Direction: describes the changing directions of concept drift among clients. Between two clients (e.g. Fig. 3c), they may have exactly the same temporal features, but their changing direction can be different (client 1 is changing from heptagon to circle and client 2 is changing to triangle).
10) Correlation: describes whether and how the drift among clients are correlated. The correlated case is illustrated in Fig. 3d, where both clients change from triangle to quadrilateral, but one becomes square and the other becomes rhombus.

This paper focuses on three temporal features (form, speed and severity) and two spatial features (coverage and synchronism). The others will be included in our future work.

IV. Scenario Analysis on Artificial Data

In this section, we conduct experiments on a set of simulated scenarios with different temporal and spatial features settings. The objective is to understand how they affect the performance of the global and local models.

A. Data Description

A hyperplane data generator [16] is used to generate artificial scenarios with $P(y|\mathbf{x})$ concept drifts. A $P(y|\mathbf{x})$ drift occurs when the hyperplane turns around. The drifting speed and severity can be easily manipulated. However, the linear separability is too easy to observe the impact of $P(\mathbf{x})$ and $P(y)$ drifts. To study these two drifting forms, we adopt the Sine data generator [5]. The generated inputs from both generators are 2-dimensional. The output label is binary, decided by the hyperplane and the Sine function.

B. Scenario Design

We have established 15 experimental scenarios, encompassing three forms of drift, speed, severity, coverage, and synchronism. Their feature settings are summarized in Table I. All of the cases generate 1000 batches (i.e. time steps) of data for 10 clients. Every batch contains 100 samples at each client.

TABLE I: Artificial Scenarios and their feature settings.

No.	Generator	Form	Speed	Severity	Cov	Syn
1	Sine	$P(\mathbf{x})$	A	H	H	S
2	Sine	$P(y)$	A	H	H	S
3	Sine	$P(y\|\mathbf{x})$	A	H	H	S
4	Hyperplane	$P(y\|\mathbf{x})$	A	H	H	S
5	Hyperplane	$P(y\|\mathbf{x})$	G	H	H	S
6	Hyperplane	$P(y\|\mathbf{x})$	I	H	H	S
7	Hyperplane	$P(y\|\mathbf{x})$	A	M	H	S
8	Hyperplane	$P(y\|\mathbf{x})$	G	M	H	S
9	Hyperplane	$P(y\|\mathbf{x})$	I	M	H	S
10	Hyperplane	$P(y\|\mathbf{x})$	A	L	H	S
11	Hyperplane	$P(y\|\mathbf{x})$	G	L	H	S
12	Hyperplane	$P(y\|\mathbf{x})$	I	L	H	S
13	Hyperplane	$P(y\|\mathbf{x})$	A	H	M	S
14	Hyperplane	$P(y\|\mathbf{x})$	A	H	L	S
15	Hyperplane	$P(y\|\mathbf{x})$	A	H	H	AS

Cov: coverage; Syn: synchronism; A: abrupt; G: gradual; I: incremental; H: high; M: medium; L: low; S: synchronous; AS: asynchronous.

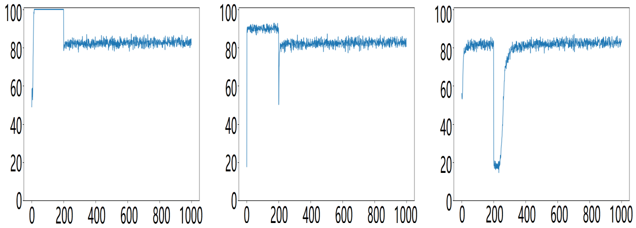
The detailed settings inside each scenario can be found on our github project page.

C. Experimental Settings

Each client trains and maintains a Multilayer Perceptron (MLP) model. Stochastic gradient descent (SGD) was chosen as the optimizer with a learning rate of 0.005. The global model is also a MLP that aggregates SGD updates through FedAvg. The experiments adopt the test-then-train method for performance evaluation at both clients and the server sides. Every client model is always tested with the next-batch local dataset. The global model is tested at each time step, with a dataset formed by 10% randomly sampled local data at each of the 10 clients from the next batch. It allows the global model to be tested on the same size of dataset as the clients. The test-then-train process is repeated 20 times. The average client and global accuracy is compared in the next section.

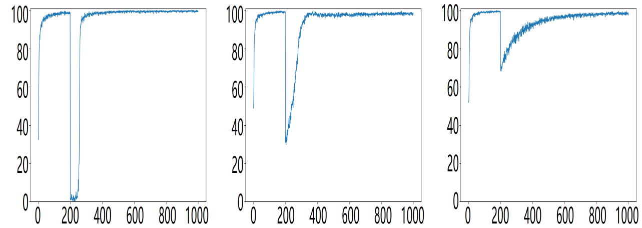
D. Experimental Analysis

We firstly discuss the three temporal features – form, speed, severity, while assuming the full coverage of synchronous



(a) $P(\mathbf{x})$. (b) $P(y)$. (c) $P(y|\mathbf{x})$.

Fig. 4: Global model accuracy in scenarios 1,2,3.



(a) Abrupt-H. (b) Abrupt-M. (c) Abrupt-L.

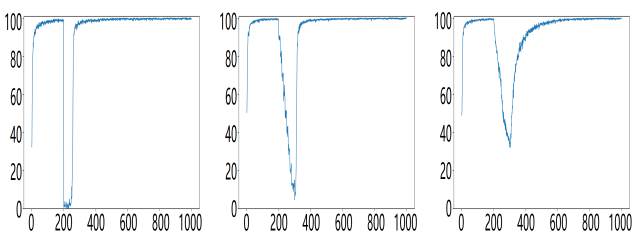
Fig. 6: Global model accuracy in scenarios 4,7,10.

concept drift on the clients. Because all the clients suffer the same drift at the same time in these cases, the global model's performance behavior is very similar to the locals', and Fig. 4 - Fig. 6 present the global performance only to save space.

Fig. 4 compares the impact of the 3 forms of concept drift. The $P(y|\mathbf{x})$ drift has been shown to be a more severe drift form than the other two, because it causes real boundary changes. The $P(y|\mathbf{x})$ plots in Fig. 4 presents an over 70% accuracy drop right after time step 200, when the real drift occurs. The $P(\mathbf{x})$ and $P(y)$ cases suffer approximately 10% accuracy drop, implying $P(\mathbf{x})$ and $P(y)$ changes affect model performance. The $P(y)$ drift causes class imbalance problems, which links to another learning challenge in FL [17] [18].

Fig. 5 compares the abrupt/gradual/incremental $P(y|\mathbf{x})$ drift cases. The abrupt case presents more severe and sharper accuracy reduction, as all the affected data undergo the change suddenly at the same time. The gradual and incremental cases produce intermediate concepts between time steps 200 and 300, so they show lesser but still significant reduction.

Fig. 6 compares the H/M/L levels of $P(y|\mathbf{x})$ drift. The more severe the concept drift at the clients is, the greater decrease in accuracy the local and global models have.

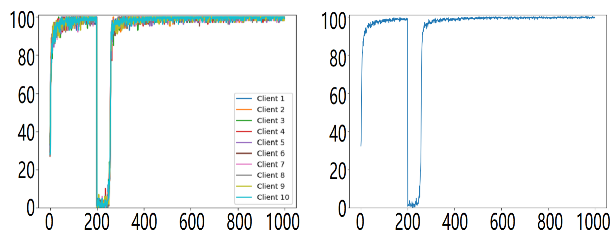


(a) Abrupt-H. (b) Gradual-H. (c) Incremental-H.

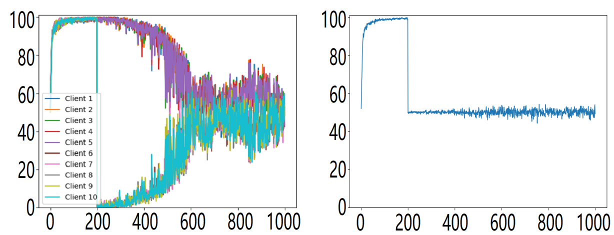
Fig. 5: Global model accuracy in scenarios 4,5,6.

The speed and severity features do not appear to affect the global model convergence. They all recover back to the similar performance level compared to that before the drift.

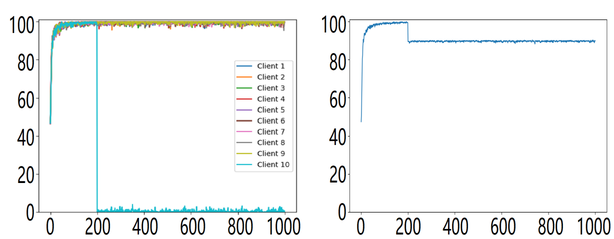
Fig. 7 compares 3 levels of drifting coverage among the 10 clients, with the same $P(y|\mathbf{x})$ drift: H – all clients suffered the drift; M – half clients (6-10) suffered the drift; L – one client (10) suffered the drift. All the local models fed with the time-drifting data show a significant accuracy reduction. It is interesting to see that the global model in the case with a full coverage of change quickly converges back to the previous performance level, because all the local models are trained and converge towards the same direction. In other words, the local data streams remain homogeneous. The M and L coverage cases, however, lead to heterogeneous data after the change, which causes convergence difficulty. In addition, the global model's performance tends to bias towards the majority data concepts. These observations suggest that, training and maintaining one global model may be insufficient to guarantee good performance on all the clients with time varying data.



(a) Coverage-H-Client. (b) Coverage-H-Global.



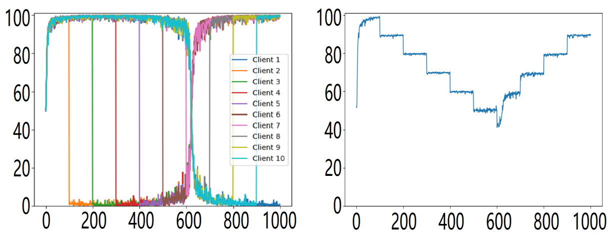
(c) Coverage-M-Client. (d) Coverage-M-Global.



(e) Coverage-L-Client. (f) Coverage-L-Global.

Fig. 7: Local and global model accuracy in scenarios 4,13,14.

Fig. 8 illustrates the asynchronous concept drift among the clients, where the same drift affect clients 2 to 10 sequentially at interval of 100 time steps from time step 100. During the first half of scenario 15 when less than half of the clients are affected by the concept drift, clients 2-6 suffer an accuracy drop because the global model is still dominated by the old concept; from time step 500 when the global model starts to be dominated by the new concept, the local models (clients 7 - 10) in the old concept start to suffer the accuracy loss until the corresponding client switches to the new concept. As the data heterogeneity level becomes less, the global performance is also rising up. This scenario further supports the claim that the global model biases towards the majority concept.



(a) Asy-Client. (b) Asy-Global.

Fig. 8: Local and global model accuracy in scenario 15.

V. Experiments on Real-World Data

In this section, we aim to find out the impact of concept drift on two real-world datasets – Electricity [19] and Forest Covertype (abbr. CoverType) [20]. Since we do not know the ground-truth of concept drift, we apply DDM [5] to detect the occurrence of drift at the clients and the central server and track the global model's accuracy over time, in order to examine if there is any correlation between the local drifts and the global accuracy. How to quantitatively measure the proposed features on real-world data is still an open question, which will be further studied. The number of training epochs at each time step was set to 500 for performance convergence. The other FL settings follow Section IV.

A. Data Description

Electricity contains 45,312 instances, collected at a 5-minute interval. The class label identifies the change of the price (2 classes) relative to a moving average of the last 24 hours. CoverType aims to predict the forest cover type (7 class labels) based on 581,012 instances and 54 attributes. To fit the datasets into our purpose, we distribute the instances one by one to each of the 10 clients without shuffling the data. Every 100 instances form a batch at one time step for each client. Electricity has 45 batches and CoverType has 581 batches in total. By doing so, the temporal order is kept, and the local datasets are likely to be homogeneous over time.

B. Experimental Analysis

We count the number of drifts detected by DDM at the local and the global models within a certain time period (every 10 batches for Electricity and every 100 batches for CoverType). They reflect how often an obvious accuracy drop can be observed during each period. Both datasets have the same DDM sensitivity threshold setting at 1.88. Table II and Table III show the numbers of the reported concept drift by DDM at the clients and the global models, in Electricity and CoverType respectively. Their global accuracy curves are shown in Fig. 9 and Fig. 10.

TABLE II: The number of detected drift in Electricity.

Model	0-9	10-19	20-29	30-39	40-46
Client 1	4.0	2.65	1.95	1.7	1.0
Client 2	3.0	2.75	0.95	0.0	1.1
Client 3	3.7	2.35	1.15	0.35	1.75
Client 4	3.05	2.8	2.35	0.4	1.05
Client 5	2.85	2.5	1.95	0.65	1.0
Client 6	3.75	1.9	1.15	1.6	1.1
Client 7	3.0	1.65	1.35	1.4	1.0
Client 8	3.05	2.25	1.85	1.55	1.1
Client 9	3.0	3.4	1.3	0.55	1.75
Client 10	2.85	3.45	1.85	1.65	2.0
Global	1.7	1.8	1.7	0.4	1.8

TABLE III: The number of detected drift in CoverType.

Model	0-99	100-199	200-299	300-399	400-499	500-581
Client 1	12.5	2.0	2.0	5.7	6.05	10.2
Client 2	11.1	2.25	6.35	4.95	5.5	9.7
Client 3	11.7	6.25	7.4	5.1	4.05	10.45
Client 4	11.3	3.3	6.5	3.0	5.6	3.25
Client 5	11.8	3.2	4.85	3.15	5.65	4.3
Client 6	13.95	3.4	5.25	3.45	6.5	10.2
Client 7	11.1	3.35	3.4	4.35	6.05	10.25
Client 8	11.95	2.3	2.4	5.0	6.2	14.05
Client 9	10.3	3.2	3.35	5.0	4.15	11.7
Client 10	14.3	2.1	4.45	5.15	3.15	9.3
Global	16.35	17.8	8.9	3.0	5.65	12.4

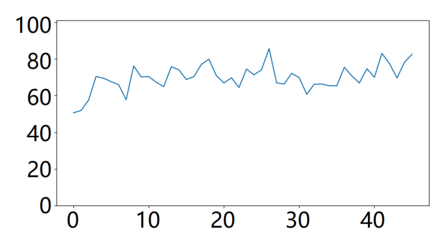


Fig. 9: Global testing accuracy in Electricity.

Comparing Electricity and CoverType, the latter has many more data batches collected over time and presents more fluctuations in accuracy. In Electricity, the number of the detected drift is relatively stable over time for both of the client models and the global model. The corresponding global accuracy is also relatively stable with a few fluctuations in the range of [60%, 80%]. In CoverType, the global model in the first two periods [0, 199] and the last period [500-581] has a higher number of drift. Correspondingly, the accuracy presents more significant drops during this time. This is consistent with our observations in the artificial scenarios. An unexpected observation is that the number of drift at the clients can be

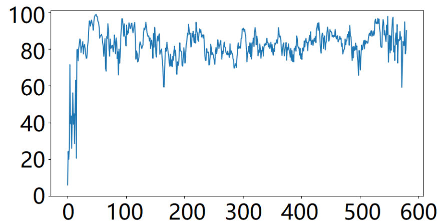


Fig. 10: Global testing accuracy in Forest CoverType.

significantly different from that at the global model. A local model significantly affected by concept drift does not mean that the global model is significantly affected, and vice versa. This finding is significant, because it sheds light on how to detect concept drift (on clients, server or both) and how to adapt the models to concept drift.

VI. Conclusions

In this paper, we studied concept drift in FL by categorizing it with ten features (including six temporal features and four spatial features) and investigating how five of the features (i.e. form, speed, severity, coverage and synchronism) affect model accuracy and convergence globally and locally.

By simulating fifteen artificial data stream scenarios with different feature settings and experimenting on two real-world datasets, we find that all the discussed features have a significant and negative impact on the global and local models' accuracy. The impact of the temporal features are similar to that in the traditional single data stream learning. Once the new concept becomes stable at the local clients, the performance recovers back to the previous level. The spatial features affect both accuracy and model convergence. The global model biases towards the data concept in the majority of the clients, and thus performs poorly on the minority concept. The accuracy does not converge back due to data heterogeneity caused by concept drift. These findings suggest us: 1) concept drift must be treated in FL. 2) Different types of drift under our categorization need targeted treatments. 3) Solely looking at the global or local performance is insufficient to tackle concept drift because the global performance does not necessarily reflect local accuracy. 4) Whether to detect concept drift locally or globally is unclear, because local and global performance can behave differently.

The next follow-up work is the five features in our categorization that are not studied in this paper. In addition, we will develop metrics to quantify the features and statistically evaluate their impact on more real-world datasets. Vertical FL will also be considered.

Acknowledgment

This work is supported by the RAEng Leverhulme Trust Research Fellowship [LTRF2122-18-106], the National Natural Science Foundation of China (NSFC) for Young Scientists [62206239] and NSFC project [62366055].

References

[1] Bartosz Krawczyk, Leandro L Minku, João Gama, Jerzy Stefanowski, and Michał Woźniak. Ensemble learning for data stream analysis: A survey. *Information Fusion*, 37:132–156, 2017.

[2] Jie Lu, Anjin Liu, Yiliao Song, and Guangquan Zhang. Data-driven decision support under concept drift in streamed big data. *Complex & intelligent systems*, 6(1):157–163, 2020.

[3] Wenke Huang, Mang Ye, and Bo Du. Learn from others and be yourself in heterogeneous federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10143–10153, June 2022.

[4] João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37, 2014.

[5] João Gama, Pedro Medas, Gladys Castillo, and Pedro Rodrigues. Learning with drift detection. In *Advances in Artificial Intelligence–SBIA 2004: 17th Brazilian Symposium on Artificial Intelligence, Sao Luis, Maranhao, Brazil, September 29-Ocotber 1, 2004. Proceedings 17*, pages 286–295. Springer, 2004.

[6] Albert Bifet and Ricard Gavalda. Learning from time-changing data with adaptive windowing. In *Proceedings of the 2007 SIAM international conference on data mining*, pages 443–448. SIAM, 2007.

[7] Shuo Wang, Leandro L Minku, and Xin Yao. A systematic study of online class imbalance learning with concept drift. *IEEE transactions on neural networks and learning systems*, 29(10):4802–4821, 2018.

[8] Leandro L Minku, Allan P White, and Xin Yao. The impact of diversity on online ensemble learning in the presence of concept drift. *IEEE Transactions on knowledge and Data Engineering*, 22(5):730–742, 2009.

[9] Yongxin Guo, Tao Lin, and Xiaoying Tang. Towards federated learning on time-evolving heterogeneous data. *arXiv preprint arXiv:2112.13246*, 2021.

[10] Liang Gao, Huazhu Fu, Li Li, Yingwen Chen, Ming Xu, and Cheng-Zhong Xu. Feddc: Federated learning with non-iid data via local drift decoupling and correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10112–10121, 2022.

[11] Giuseppe Canonaco, Alex Bergamasco, Alessio Mongelluzzo, and Manuel Roveri. Adaptive federated learning in presence of concept drift. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2021.

[12] Yujing Chen, Zheng Chai, Yue Cheng, and Huzefa Rangwala. Asynchronous federated learning for sensor data with concept drift. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 4822–4831. IEEE, 2021.

[13] Fernando E Casado, Dylan Lema, Marcos F Criado, Roberto Iglesias, Carlos V Regueiro, and Senén Barro. Concept drift detection and adaptation for federated and continual learning. *Multimedia Tools and Applications*, pages 1–23, 2022.

[14] Dimitrios Michael Manias, Ibrahim Shaer, Li Yang, and Abdallah Shami. Concept drift detection in federated networked systems. In *2021 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE, 2021.

[15] Ellango Jothimurugesan, Kevin Hsieh, Jianyu Wang, Gauri Joshi, and Phillip B Gibbons. Federated learning under distributed concept drift. In *International Conference on Artificial Intelligence and Statistics*, pages 5834–5853. PMLR, 2023.

[16] Geoff Hulten, Laurie Spencer, and Pedro Domingos. Mining time-changing data streams. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 97–106, 2001.

[17] Chenguang Xiao and Shuo Wang. An experimental study of class imbalance in federated learning. In *IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–7, 2021.

[18] Chenguang Xiao and Shuo Wang. Triplets oversampling for class imbalanced federated datasets. In *Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, page (In press), 2023.

[19] Michael Harries, New South Wales, et al. Splice-2 comparative evaluation: Electricity pricing. 1999.

[20] Jock A. Blackard and Colorado State University. "uci machine learning repository," [online]. available: http://archive.ics.uci.edu/ml/datasets/covertype.