Master's Thesis of Data Science

A Time-Machine Learning Framework: Learning from others

다자간 환경에서의 효율적 시계열 예측 프레임워크 – 선구자-추격자 학습

February 2023

Graduate School of Data Science
Seoul National University
Data Science Major

Chang Sub Chang

A Time-Machine Learning Framework: Learning from others

Advisor : Wen-Syan Li

Submitting a master's thesis of Data Science

December 2022

Graduate School of Data Science
Seoul National University
Data Science Major

Chang Sub Chang

Confirming the master's thesis written by
Chang Sub Chang
January 2023

Chair	Hyungsin Kim	(Seal)
Vice Chair	Wen-syan Li	(Seal)
Examiner	Min-hwan Oh	(Seal)

Abstract

As the concept of AI and data science is gaining popularity in solving real-world problems, many application areas are being discussed. Among those, time-series data is found readily in the real-world – sales data, electric vehicle (EV) battery data, sensor data from various appliances, stock market data, etc., and is used for anomaly detection and forecasting to name a few. This research focused on solving the time-series forecasting problem, where multiple pieces of equipment's share their results real-time and help each other forecast one's own future referring to others' past behaviors. In the novel concept of Frontier-Follower Learning, the players are divided into either Frontiers – whose past behaviors (results) be reference points for learning by others - , or Followers – who mainly refer to the past behaviors of Frontiers. Frontiers and Followers are not static but are reassigned dynamically by the degree of similarity among past data points. Frontiers' past records are evaluated by the means of similarity index, which in this paper used dynamic time warping (DTW), and the information of the Frontiers' reference data points is fed into the model only to the degree of its similarity to the Followers' model. Several scenarios with cases have been experimented with to validate the concept : base cases with 10 pieces of equipment with different usage behaviors, by increasing the number of equipment, increasing the time gap among equipment, comparison with teacher-student network model, and even validation using the real-world data of BXB corporation. The results proved that the novel concept of evaluating the value of the information and dynamically updating the model referring to its Frontiers has better performance. The concept can be further applied to real-world settings where multiple players respectively have a limited number of past records, but a collectively meaningful amount for training.

Keyword : Time-series learning, teacher-student network, dynamic time warping, similarity index, frontier-follower learning

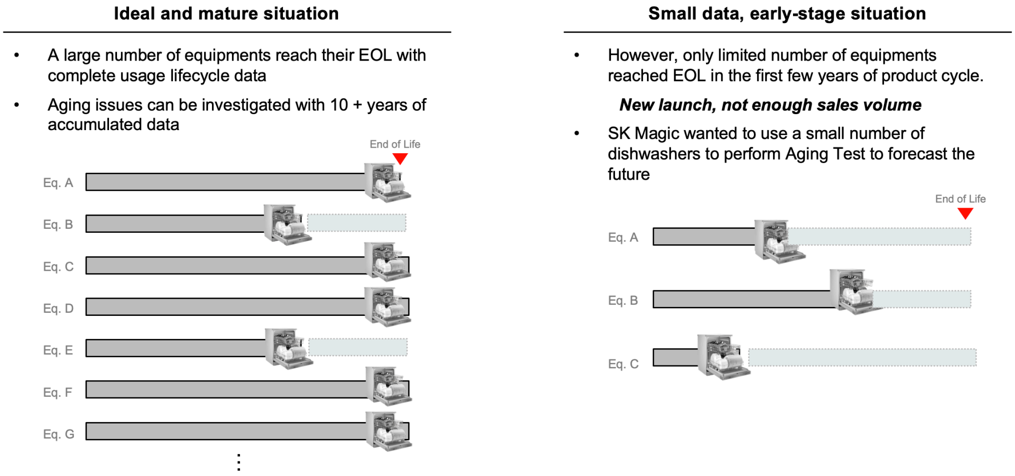
Student Number : 2020-24711

Table of Contents

Chapter 1. Introduction

Time series forecasting plays a crucial role in solving real-world problems. Financial institutions want to use it for stock market pricing forecasts, manufacturing companies to estimate market demand, and supply chain companies to optimize the use of their supply chain fleet. Nowadays, thanks to the fast introduction of electric vehicles, time series forecasting is now even used for estimating remaining battery values as well [5]. In real-world problems, there are environments where multiple similar equipment data are collected. Energy Consumption [4], EV Batteries [5], and even household appliances such as fridges and dishwashers could be examples.

Here, I would like to bring up two real-world examples that use time-series forecasting extensively. One of the dishwasher companies wanted to study the aging issues with their dishwashers. [Fig 1] In the ideal and mature situation, the company can collect big enough data from devices of their own. However, even though they wanted to collect data, since the machine requires a lot of electricity to run, and its wear-down period or end-of-life after the complete usage lifecycle is too long, the company could not run the tests thus resulting in the limited number of the data. Even though, the company considers collecting the data from the devices that were sold, in the new launch period, not enough sales volumes could be attained thus resulting in the limited number of available data again. The minimum requirement of the dataset was to have a 10+ year-long dataset with several events that lead to the wear-down of the device. However, they wanted to use a small number of dishwashers to forecast the future.



Ideal and mature situation

- A large number of equipments reach their EOL with complete usage lifecycle data
- Aging issues can be investigated with 10 + years of accumulated data

Small data, early-stage situation

- However, only limited number of equipments reached EOL in the first few years of product cycle.

New launch, not enough sales volume

- SK Magic wanted to use a small number of dishwashers to perform Aging Test to forecast the future

Fig 1. A dishwasher aging problem. LHS shows the ideal and mature situation, and RHS illustrates the opposite under the limited setting.

Similarly, BXB, a technological subsidiary of Brambles Group in Australia specialized in data management and analytics for its parent company's logistics data, wanted to monitor containers and know when to change the communication module batteries. [Fig2] The communication module, which is attached to the containers, transmits critical information on the location of the containers, the external environment, and most importantly remaining values of the batteries. Since the module transmits critical information, the company wanted to know when exactly the module would need to change batteries. However, as it was for the dishwasher cases, the module uses Bluetooth technology which requires low battery use, and to save the electricity the information is transmitted in a very limited manner, thus hard to collect the full cycled battery use information for the model training.

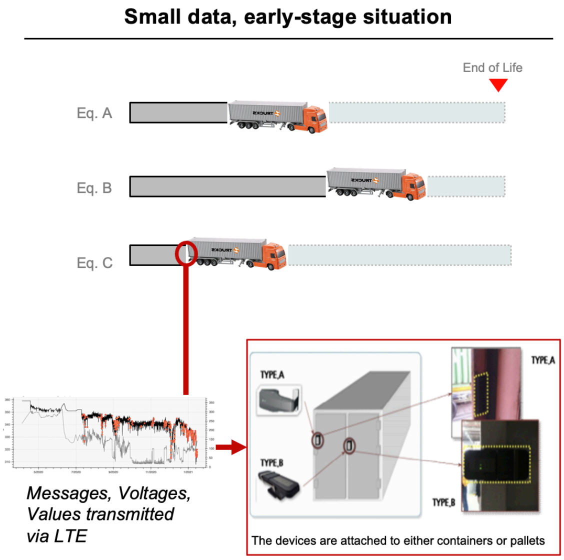

Small data, early-stage situation
End of Life
Eq. A
Eq. B
Eq. C
TYPE A
TYPE B
TYPE A
TYPE B
Messages, Voltages, Values transmitted via LTE
The devices are attached to either containers or pallets

Fig 2. Forecasting the future remaining values of the battery

Thus, I suggest a FRONTIER-FOLLOWER learning – where participants in the learning are grouped as FRONTIERS and FOLLOWERS. The FRONTIERS are the participants who have longer timestamps with extensive usage information, while the FOLLOWERS are participants who follow the trajectories of the FRONTIERS. One of the analogies could be found in the investment philosophy of Mr. Son of Softbank company. Mr. Son invests in companies in East Asia whose business models are similar to their counterparts in North America. For example, the business model of Alibaba group of China is similar to Amazon of the USA, and again, the business model of Coupang is two predecessors. These predecessors can be considered as Frontiers, and Coupang, a follower can benefit from learning from its Frontiers' past experiences. As long as the business models are similar to each other, ranging from the demographics, competitive landscapes, and customers' preferences, similar conditions that Frontiers has already experienced can give valuable information to the companies that follow years later even in different countries.

Chapter 2. Related Works

Time-series forecasting is not a new domain. There have been conventional statistical approaches such as ARIMA (Autoregressive Integrated Moving Average) [8], exponential smoothing and to name a few. Nowadays, many time-series forecasting technics are based on Machine Learning techniques, and one of the notable well-known approaches is RNN (Recurrent Neural Networks) which makes a chained network where the memory of the past is passed along to the nearby models. Not to mention LSTM and GRU which are based on the RNN-based approach. Even some Deep Learning based approaches, such as N-Beats were introduced as well. [9]

With the advent of Transformer-based models [10], even in the domain of Time-Series, many studies have been conducted to develop based on the Transformers. [2] However, Transformers were considerably computationally heavy – since it is also used for large language models and even Zeng. A. et al. proved in their paper, that even a simple linear model can beat the Transformer-based time-series forecasting model's performance. [1] (Detailed explanation of the model architecture is further described in section 5)

The transformer-based approach in the Time-series model assumes that there is a semantic embedding among points, however, when it comes to the serial number where the order of the numbers is important, it is very hard to assume that the embedding among points exists – which leads to better performance. The linear model, on the other hand, which uses two components: Decomposition and Linear components guarantees high efficiency and interpretability, and it is easy to use. [1]

When it comes to the study of similarities between time-series data, DTW (Dynamic Time Warping) techniques are widely used. [12] DTW lets you calculate between different time horizons regardless of the actual length difference there might be. The cosine similarity approach - which is conventionally used for similarity calculation in word embeddings [13] – is now being considered for time-series learning as well, and is even tested in sales forecasting. [11] However, time-series learning is prone to learning the noises leading to bad model performance.

Instead of using the conventional concept of DTW, Soft-DTW will be used here. It proposes the use of a soft minimum in replacement of the real minimum value. This enables differentiation so that the gradient can then be used as a gradient to update the model by backpropagation. [15]

The concept of learning from others was studied in the form of referring to other similar yet bigger models, or a teacher network. [14] Usually, a teacher network model is the biggest model that is pre-trained with all available data. Several student networks, which are presumed to behave similarly to the teacher network are then trained with their data, while expecting knowledge from the teacher network be integrated into the training period by 'knowledge distillation'.

Chapter 3. Approach

For example, if a certain electric vehicle is extensively driven, its time for wearing down would be faster than the rest of the ordinary electric vehicles. This vehicle can be considered as one of the FRONTIERS. [Fig 4] The FOLLOWERS, then are expected to follow the steps of the FRONTIERS. However, there could be some FOLLOWERS who share similar usage behaviors or preset characteristics to the certain FRONTIER, while others do not. Based on this intuition, thus, I would like to introduce a novel approach to incorporate and re-evaluate the value of the respective data based on the similarities within peers and how much experience the FRONTIERS are. The FRONTIERS can be constantly changed as time goes by - since the behavior of the current FRONTIER can easily be replaced with its CONTENDERS. A frequent update of the relationship among peers makes the model update every day. The FOLLOWERS learn from FRONTIERS, especially FOLLOWERS try to follow the trajectories of the FRONTIERS. Like Time-Machine, the FRONTIERS' prior experience and knowledge are integrated into the framework.

However, one thing to note is that, there could be multiple FRONTIERS for a single FOLLOWER, while there could be only one or no FRONTIER for some of the FOLLOWERS, the positions (roles) for each of the participants in the learning can change depending on the focus of the models.

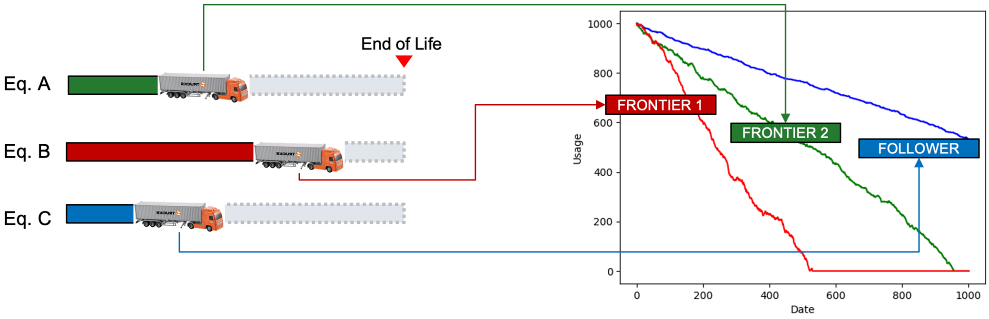
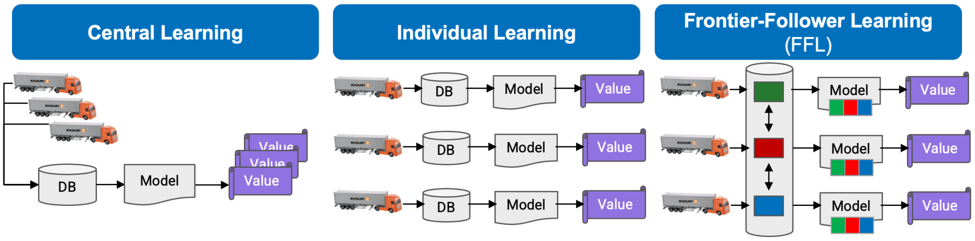


Fig 3. Proposed Approach under three equipment (trucks) situation

Chapter 4. Scenarios

In order to validate the proposed approach, three scenarios including the proposed approach, FFL (FRONTIER-FOLLOWER LEARNING) can be studied. [Fig 5]

Central Learning is where all data is transmitted and collated in the server, to build one uniform model. The model is player-agnostic in training and inference. Individual Learning is where multiple models are built using only respective data of each of the equipment. The data is not collated together and not shared with other players in the learning. FFL (FRONTIER – FOLLOWER LEARNING) is where the data is gathered, and compared simultaneously with each other to calculate the similarities and differences.



Central Learning is where all data is transmitted and collated, to build one uniform model,

Individual Learning is where all data is transmitted but not collated, thus models do not interfere with each other

FFL is where the data is gathered, and compared simultaneously with each other to calculate the similarities and differences

	Description
Central Learning	All data is gathered, and collated
Individual Learning	Data is gathered, but not collated / Siloed
FFL	Data is gathered, and compared. Selectively used for the respective model build

Fig 4. Central Learning, Individual Learning, and FFL (FRONTIER-FOLLOWER LEARNING) in regards to its use of individual data and training

Chapter 5. Frontier-Follower Learning (FFL) Algorithm

5-1. How it works

FFL Algorithm is based on the DLinear model. [1] The DLinear model is a Decomposition Linear model which showed simple yet powerful performance in forecasting the future. The reason for choosing this basic model is that a single-layer linear network is the simplest model that can compress information from the past to predict the future. Plus, in the previous studies, the time series decomposition was proven to improve the performance of Transformer-based models – which is also applicable to linear models in that it is model-agnostic.

The model is composed of a decomposition component and a linear network component. In the decomposition component, it decomposes the data into the trend part and the remainder, which are trained independently and then merged. [Fig 5]

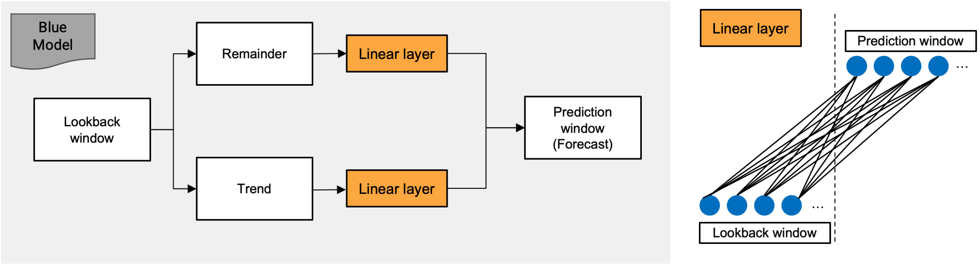
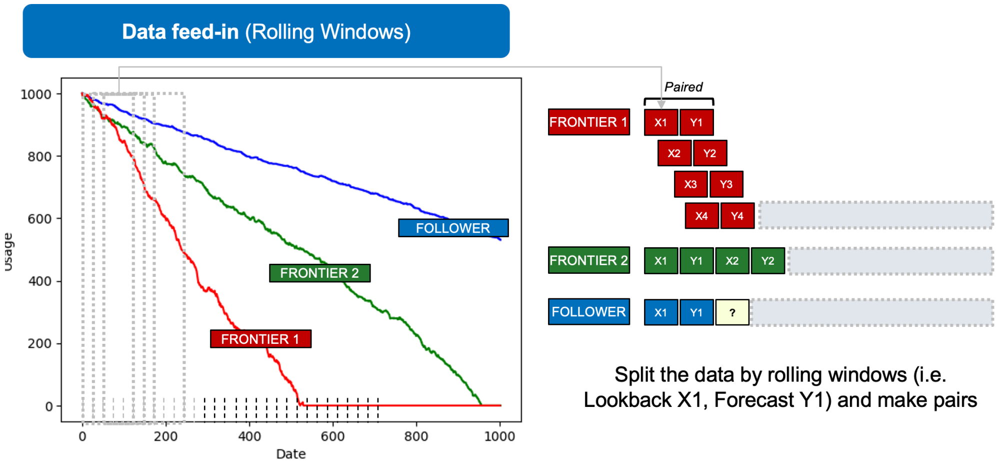


Fig 5. Illustration of DLinear Model

Imagine, there are three players in the model. [Fig 6] FRONTIER 1 (Red), FRONTIER 2 (Green), and FOLLOWER (Blue). The objective of the example is to forecast the FOLLOWER's future based on the data from FRONTIERs. The historical timestamps – denoted as X values, and prediction timestamps – denoted as Y are paired to be fed into the model on the rolling window basis. Each set of historical and prediction timestamps are paired up to be fed into the model.



Split the data by rolling windows (i.e. Lookback X1, Forecast Y1) and make pairs

Fig 6. How data is fed into the model

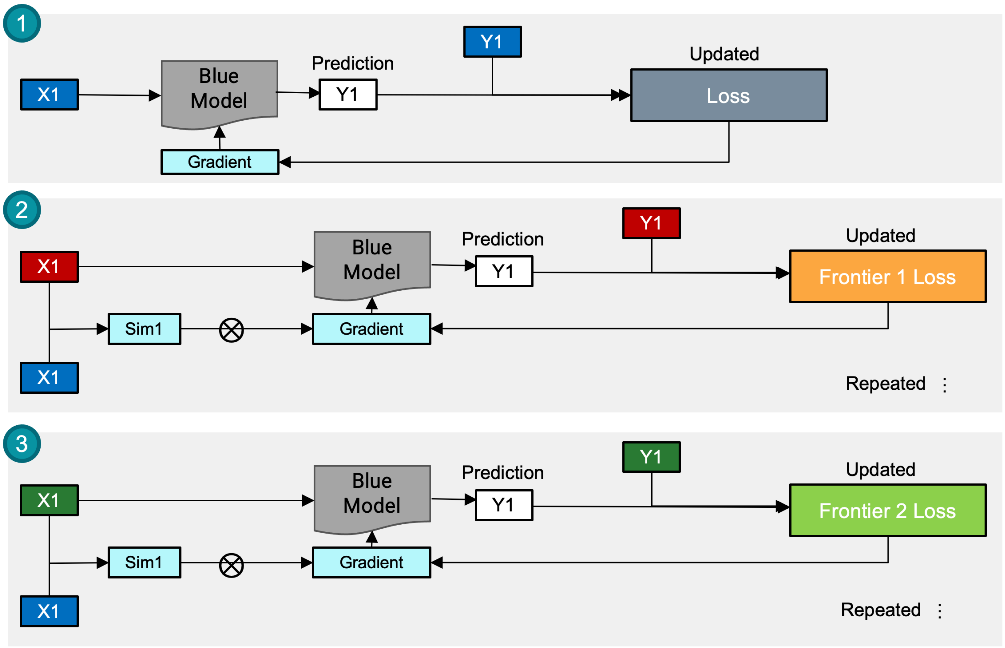


Fig 7. FFL Model Training

When training the model in FFL, the training takes five steps [Fig 7]:

(1) The FOLLOWER's data (Blue) is fed into the model, the Loss is calculated for the backpropagation, and the model is trained

(2) The FOLLOWER's data is then compared with one of the FRONTIER's data (Red / Green) for Similarity Calculation [Fig 8]. Each of the data passes through the encoder and decoder network to calculate the cosine similarity. [3]

(3) The FRONTIER's data (Red / Green) is fed into the model, and the Frontier Loss is calculated for the backpropagation.

(4) The Gradient from (3) is then multiplied with the "Sim" value from (2) and the model is trained

(5) Repeat for all the relevant data points

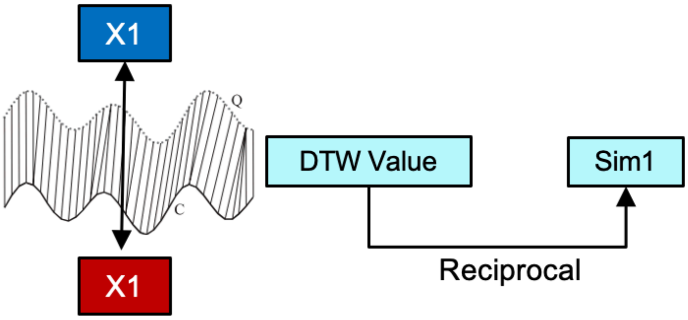


Fig 8. Similarity Calculation

Here are the highlights of the algorithm. Multiple time-series can be compared by single value – dynamic time warping (DTW) similarity. DTW is a metric used to calculate the warping distance – based on the Euclidean distance between two time series points. DTW is appropriate for the analysis since it can compare two time series values in different timestamps. The smaller the value is, the more similar two timeseries are : thus, reciprocal values are used to define the similarity. The similarity index is then multiplied by the gradient which feeds in the information to the model on a relative scale.

Cosine similarity, whose calculation considers the latent values, is a metric for measuring distance when the vector size does not matter. Since the magnitude of the time-series values is not of our interest, cosine similarity was also considered as one of the similarity metrics for the model training. However, the cosine similarity is well known for being prone to training noises – which is not suitable for this case where the number of the data points goes well beyond 1,500 points. Thus, DTW was chosen as a metric for the similarity index.

Plus, the Frontier-Follower Learning framework is model agnostic. The similarity index can be plug-and-played for every model of interest. Here, as discussed earlier, the linear model is used for its computation efficiency and proven track record, but the framework is independent of the types of the models – which gives the freedom of model selection.

5-2. Comparison with teacher-student network

As covered in Chapter 2, FFL concept can be considered similar to that of the teacher-student network. A teacher network model, which is trained with other bigger datasets, can give relevant information to the student model. [14]. The difference comes from that in the teacher-student network, the teacher model is trained with its own data, and the data itself is not considered in training the student model – the model of interest. The research focused on whether the differences from the input of data and calculating the differences among data make differences in the results as well. Instead of comparing the single datasets, the module of training a student model is composed of calculating the distillation loss. Referring to the original concept of the teacher-student network where the knowledge distillation captures the information and pass to the student model, the teacher-student network was built in two steps. [Fig 10] First, build a teacher model (individual model) based on the data of that particular model (here, blue model). Second, train a student model following the steps below:

(1) Feed in the data of the Blue model, X1, into the Red model and return the output value, Y1'

(2) Feed in the data of the Blue model, X1, into the Blue model and return the output value, Y1"

(3) Calculate the loss between Y1' and Y1", which is a distillation loss

(4) Calculate the loss between Y1" and Y1 (the real data), which is a student loss

(5) Add two losses, distillation loss, and student loss, and update the model based on the gradient from the loss

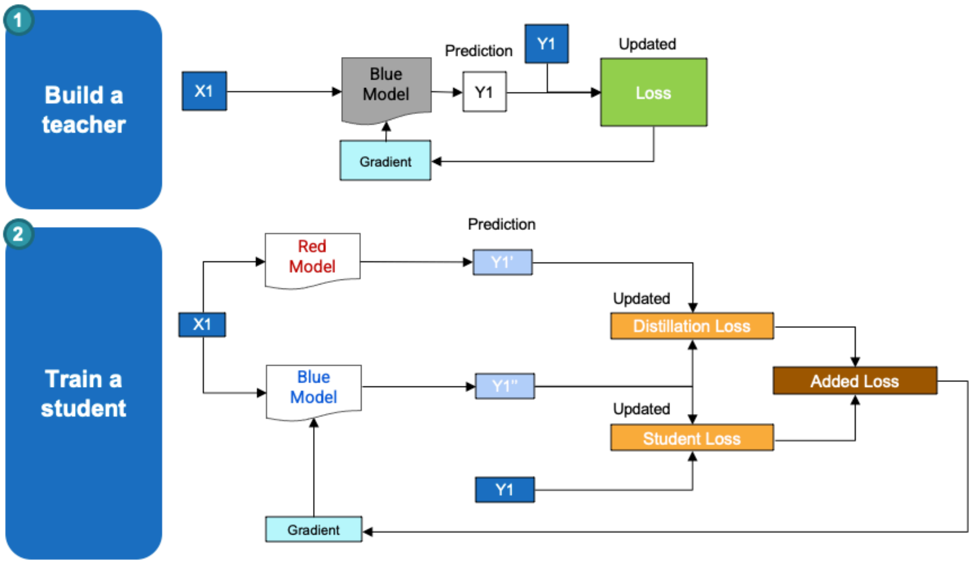


Fig 10. Build a teacher student network

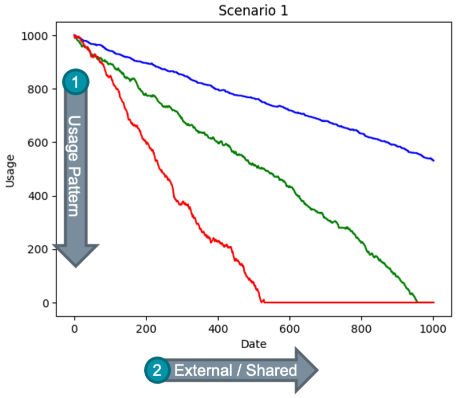
In FFL, all datasets are compared and similarities are calculated with their peers (frontiers). Thus, the relevant information from the other datasets is fed into the model directly via gradient, resulting in better performance. The magnitude of the information from the main model, and the rest are collectively combined.

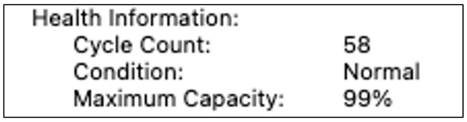
In the teacher-student network model, the information on each of the datapoints is collected in the format of model, and its distillation loss which contains the info is calculated. The model is updated via the loss fed from the frontiers' models. Efficient only when the big overarching model exists.

Chapter 6. Experiment

6-1. Overview

Three scenarios including FFL have been preliminarily experimented with. For the experiment to validate the FFL approach, the dataset has been synthesized with 1,000 timestamps. Each of the pieces of equipment were given with respective behaviors. Two major parameters that influence virtual situation was (1) Usage Patterns, and (2) External / Shared environment. [Fig 11]




Health Information:
Cycle Count: 58
Condition: Normal
Maximum Capacity: 99%

Macbook's battery information

Fig 11. Experiment data synthesis rationale

Dual-angular approach

Consider both the usage pattern and the external factored data together for the similarity calculation and training:

(1) Usage Pattern

Depending on the equipment usage behavior, the health of the battery and dropping rate is different. i.e., the more cycles the battery went through, the less capacity it has, and also the dropping rate and behavior. [Fig 10] The accelerated deteriorating behavior was considered in building the data-generating module.

(2) External / Shared

At a certain period, the external/shared environment has an impact on the values. i.e., at a certain date, if there were severe outside weather conditions, the behaviors might have been affected altogether.

6-2. Cases and datasets

The three scenarios (Central learning, Individual learning, and FFL) were experimented with in five major cases using synthetic data that resembles the real-world dataset. Those five cases consist of (1) Base cases, where the same/different time and starting values were tested, (2) Long and short time cases, where different time gaps among equipment were tested, (3) More number of pieces of equipment cases, where the number of equipment increased by two times, and four times, (4) Different usage behaviors cases, where the different intensity of the usage behaviors was tested, and (5) Teacher-student network. The experiment is then further expanded to validate the idea using the real-world data of BXB technologies.

As a default setting, 1,000 hypothetical days of data are synthesized for 10 equipment

(1) Base case (with four sub-cases)

- Case 1-1. Same starting time, same starting value
- Case 1-2. Different starting time, same starting value
- Case 1-3. Same starting time, different starting value

- Case 1-4. Different starting time, Different starting value

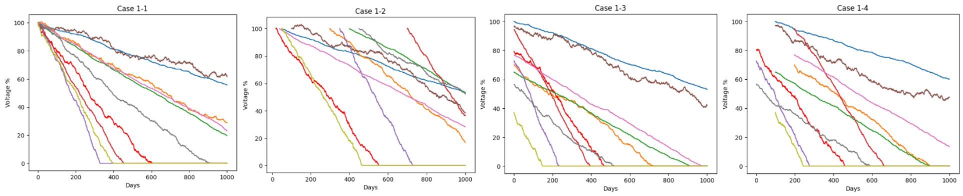


Fig 12. Conditions and plots for case 1

(2) Long and short time periods

Different time gaps among equipment, as the gap between each of the equipment are wider, the more room for learning from past experience with more data points.

- Case 2-0 : Same as Case 1-2 for reference
- Case 2-1 : -25 for low 5 equipment / +200 for top 5 equipment to widen 125 timestamps
- Case 2-2 : -50 for low 5 equipment / +400 for top 5 equipment to widen 125 timestamps

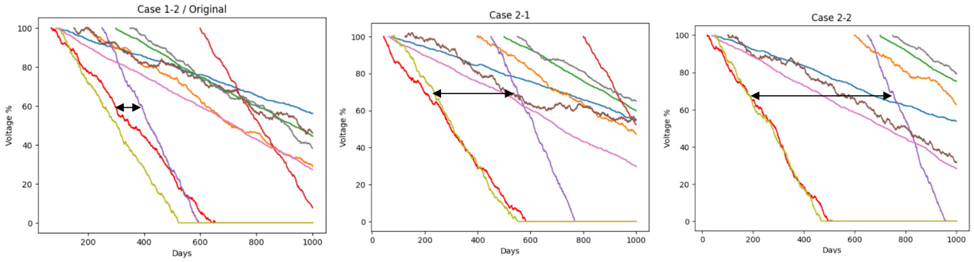


Fig 13. Conditions and plots for case 2

(3) Various number of equipment

More number of equipment, as the number of equipment increases, the more data points to learn from. However, the time for learning might get longer.

- Case 3-0 : Same as Case 1-2 for reference, 10 equipment

- Case 3-1 : 20 equipment
- Case 3-2 : 40 equipment

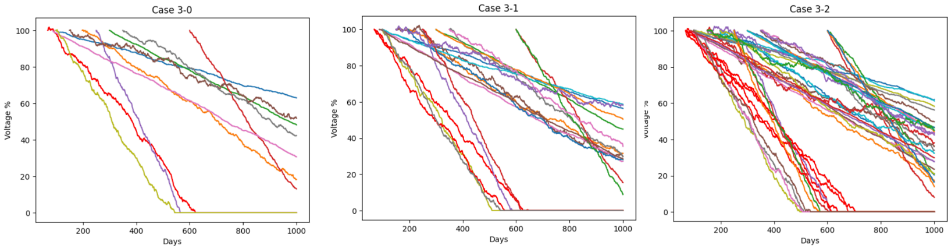


Fig 14. Conditions and plots for case 3

(4) Different usage behaviors

Under the imaginary case where equipments are from different manufacturers, hypothetically, each piece of equipment might behave differently

- Case 4-0 : Homogenous equipment, all the equipment behaves identically
- Case 4-1 : Heterogenous equipment with mild usage behavior
- Case 4-2 : Heterogenous equipment with heavy usage behavior

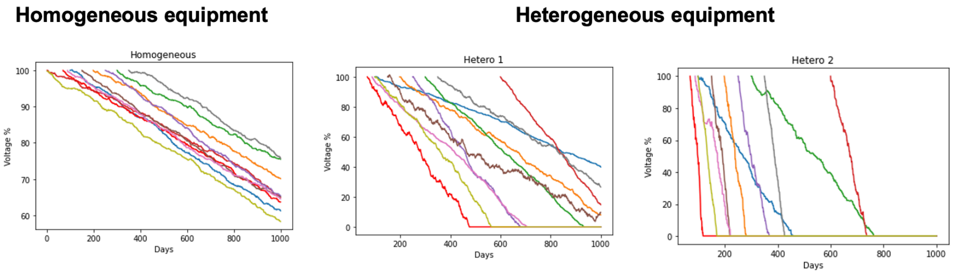


Fig 15. Conditions and plots for case 4

(5) Teacher-student network

The data and the settings are the same as its reference case, Case 1-2.

(6) Real-world data (BXB)

BXB shared 12,668 records (data points) for 10 devices of the dataset which were collected from the device attached to the containers around the world [Fig 2]. The dataset is composed of 26 columns which included server time, device time, link, message type, types, counter, temperatures, accelerator, voltage, timestamp, etc. The voltage, which is a proxy for the battery life, drops when the energy is consumed and the information kept (e.g., messages, and message types) could have used energy to drop the voltage levels. [Fig 16] Raw data is pre-processed in terms of noise handling and timestamps adjustment so that it resulted in the 160-day-long dataset with 10 devices.

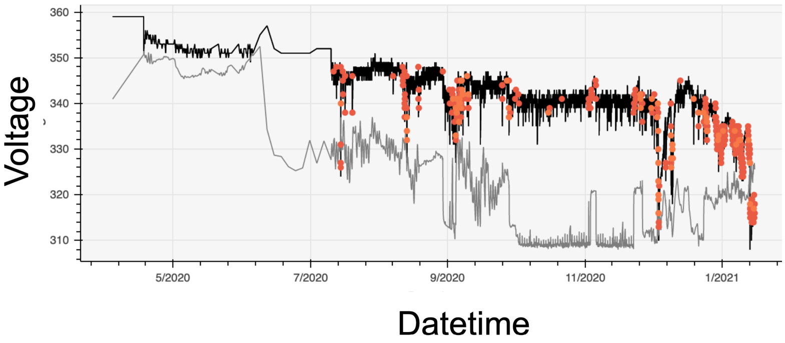


Fig 16. Voltage level movement for one of the devices

The research here is to focus on the univariate time-series forecasting model, thus, the relationship between time and voltage is investigated. [Fig 17] Unlike the synthetic data used for the hypothetical cases discussed above, BXB did not have enough data that span the entire lifecycle for each of the devices which resulted in limited experiments and the results will be discussed in the following Chapter 7.

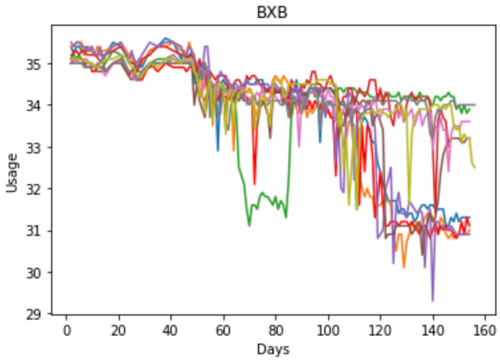


Fig 17. BXB dataset

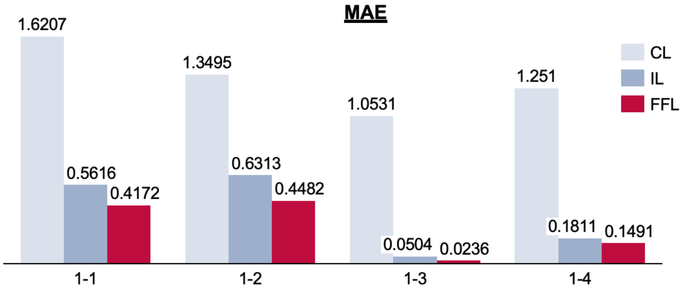
6-3. Results

As used in the DLinear work and other previous works, following previous works [1, 7], Mean Squared Error (MSE) and Mean Absolute Error (MAE) are calculated as metrics. However, to minimize the impact/influence of the different distributions each of the data is coming from, the entire dataset has been normalized between 0 and 1 for model training. Thus, MSE results are getting too small for the value being between 0 and 1, so that MAE is primarily investigated for the result interpretation.

[Case 1] Base case – same/different starting time, same/different starting value

Description			MSE			MAE		
			CL	IL	FFL	CL	IL	FFL
Case 1	1-1	Same starting time, starting value	0.3982	0.0656	0.0407	1.6207	0.5616	0.4172
	1-2	Different starting time, same starting value	0.2830	0.0757	0.0593	1.3495	0.6313	0.4482
	1-3	Same starting time, different starting value	0.2023	0.0066	0.0016	1.0531	0.0504	0.0236
	1-4	Different starting time, starting value	0.2586	0.0068	0.0047	1.2510	0.1811	0.1491

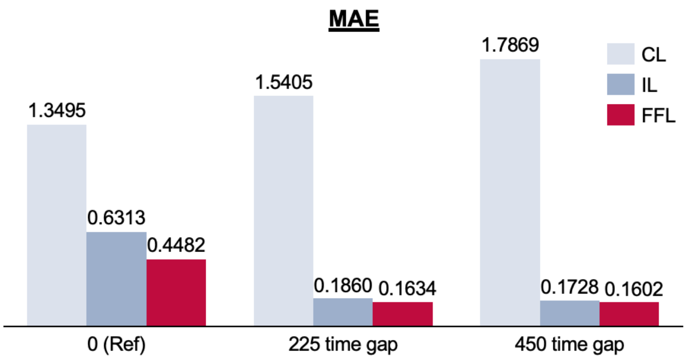
For all cases, FFL showed better performance than CL and IL. Case 1-3, generally showed better performance than other cases since it all starts at the same starting time, which implies that there were more data points to refer to, and different starting values give them a clear distinction between frontiers and followers compared to case 1-1.



[Case 2] Long- and short-time horizons

Description			MSE			MAE		
			CL	IL	FFL	CL	IL	FFL
Case 2 (1,500 Timestamps)	1-2	Ref.	0.2830	0.0757	0.0593	1.3495	0.6313	0.4482
	2-1	225 time gap	0.3156	0.0086	0.0069	1.5405	0.1860	0.1634
	2-2	450 time gap	0.4124	0.0082	0.0064	1.7869	0.1728	0.1602

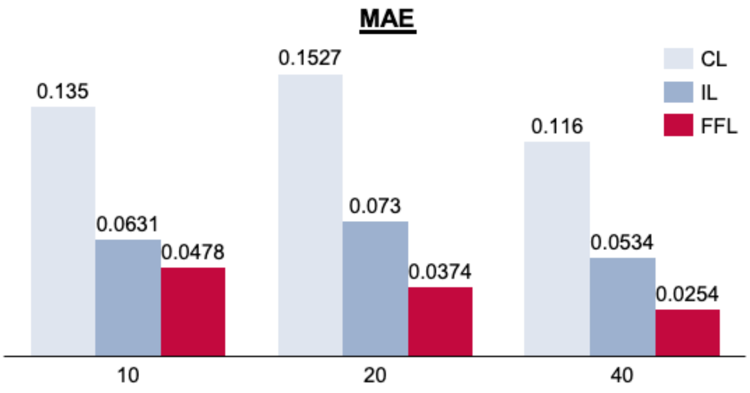
For CL, as the time gap widens the accuracy went down since it does not take into consideration the value of the different information. However, for FFL, wider time gap among agents gave explicit distinction between frontier and followers, leading to better performance than CL and IL scenarios. Also, there was a slight better gain in the performance between 225 timestamps case, and the 450 timestamps case. It implies that the greater number of points to compare to, either by increasing the number of equipment or the number of Frontiers, the better performance a model can expect.



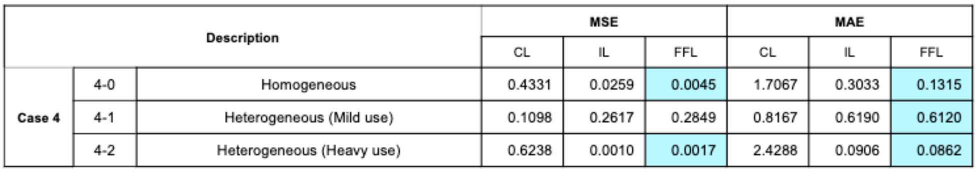
[Case 3] Various number of equipment

Description			MSE			MAE		
			CL	IL	FFL	CL	IL	FFL
Case 3	1-2	Ref (10 devices)	0.0283	0.0076	0.0059	0.1350	0.0631	0.0478
	3-1	20 devices	0.0308	0.0103	0.0061	0.1527	0.0730	0.0374
	3-2	40 devices	0.0203	0.0132	0.0119	0.1160	0.0534	0.0254

The overall MAE result is downward trend. FFL went down from equipment # 10 to 20, while CL and IL's accuracy went worse. However, from equipment # 20 to 40, FFL's performance did not increase drastically, and is since there were many followers not frontiers, thus the information gained from similarity values did not increase to sufficient #.

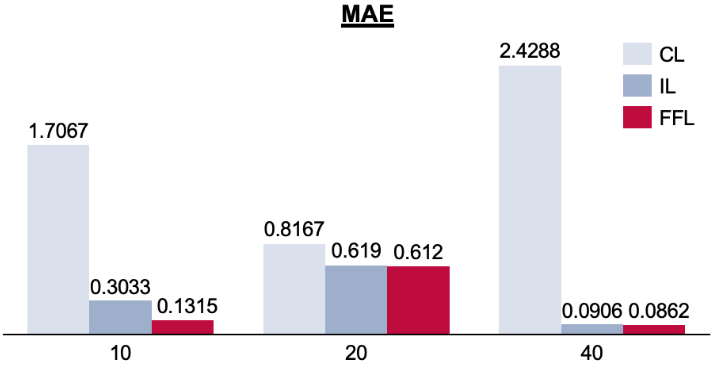


[Case 4] Usage behaviors

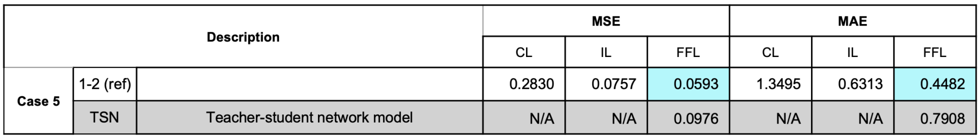


Description			MSE			MAE		
			CL	IL	FFL	CL	IL	FFL
Case 4	4-0	Homogeneous	0.4331	0.0259	0.0045	1.7067	0.3033	0.1315
	4-1	Heterogeneous (Mild use)	0.1098	0.2617	0.2849	0.8167	0.6190	0.6120
	4-2	Heterogeneous (Heavy use)	0.6238	0.0010	0.0017	2.4288	0.0906	0.0862

Regardless of the usage behaviors or homogeneity of the equipment, FFL showed better performance than CL and IL. However, in the case of heavy usage (where the voltage value dropped faster than mile usage case), the accuracy gap between CL and FFL is more drastic than that of other cases. It can be comprehended as since all the voltage values drop fast, there is limited room for noise to take part in thus resulting in, so bigger odds for FFL to find similarities from the Frontiers.

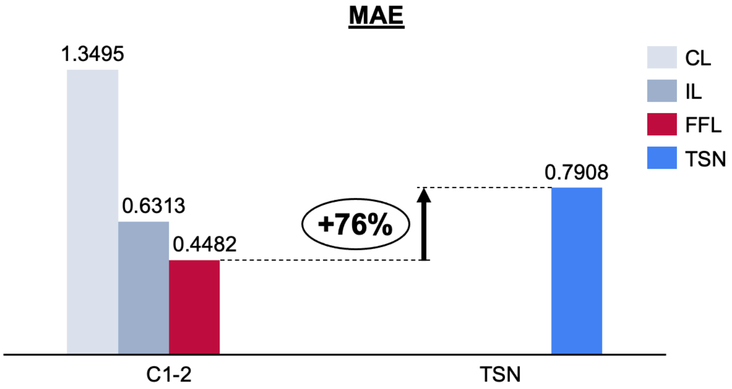


[Case 5] Teacher-Student Network

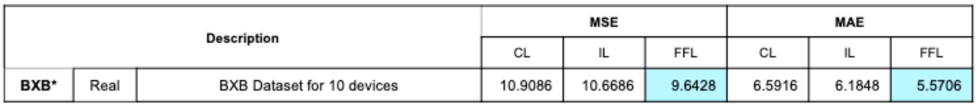


Description			MSE			MAE		
			CL	IL	FFL	CL	IL	FFL
Case 5	1-2 (ref)		0.2830	0.0757	0.0593	1.3495	0.6313	0.4482
	TSN	Teacher-student network model	N/A	N/A	0.0976	N/A	N/A	0.7908

Compared to the teacher-student network (TSN), FFL showed ~76% better performance. Even though TSN had lower MAE than CL, it could not beat either IL or FFL. Training the model based on the second-hand information by the means of knowledge distillation did not show sufficient performance.

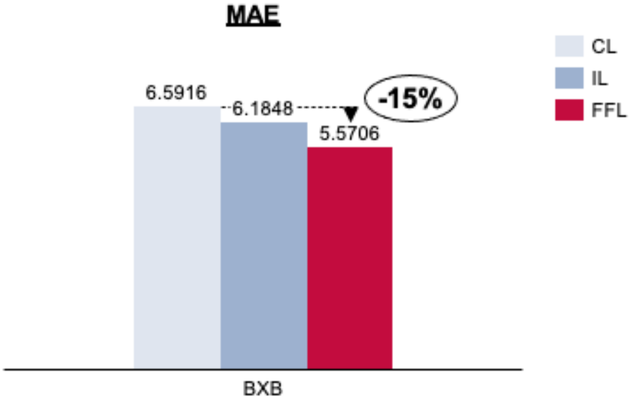


[Case 6] Real data (BXB)



Description			MSE			MAE		
			CL	IL	FFL	CL	IL	FFL
BXB*	Real	BXB Dataset for 10 devices	10.9086	10.6686	9.6428	6.5916	6.1848	5.5706

FFL showed better performance compared to CL and IL by 15%. However, due to the limited number of data points, only limited experiment could be conducted, and resulting in the limited enhance performance of FFL compared to CL and IL



Chapter 7. Interpretation

In most cases, FFL resulted in better performance compared to Central Learning(CL) and Individual Learning(IL). Even though the degrees of better performance vary by cases investigated, in general, FFL proved its value. In the study to compare with already existing concepts in order to validate the novelty, the Teacher-student network showed less accuracy compared to FFL. It implies that the direct comparison of the data points has better performance. Teacher student network is more suitable for cases where soft labels exist (i.e., deep learning image classification), but not for time-series data training where the soft labels do not exist and the size of the data for training is not too computationally heavy.

In the cases of testing robustness, even when the number of equipment increases, FFL still showed better performance compared to CL and IL. The higher the time gap between equipment is, the higher the performance is. It is due to more data points collected from the past frontiers helping followers make more reference points. The types of equipment – whether the equipment is completely identical or not - have limited influence on the results. It implies that the concept of FFL can be further utilized for the cases even when the origin of the data is slightly different (i.e., utilizing the data streamed from the older version of an electronics model for the newer version).

Due to a limited number of samples, BXB data could not be fully studied - no full cycle data was provided, however, still in the BXB case, FFL showed better performance compared to CL and IL.

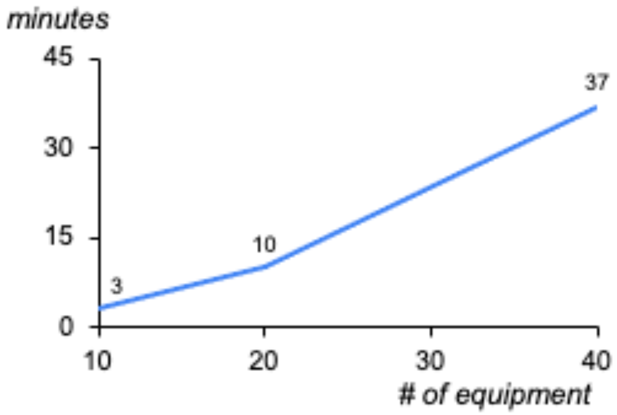
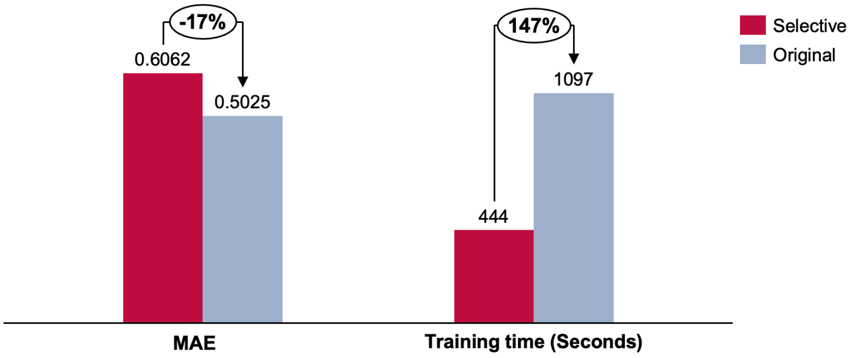


Fig 18. Training time by # of equipment

When it comes to real-world utilization, the time used for training matters. Even though the accuracy of the novel idea is higher than the conventional approach, if it takes longer training, then the business value of the novel idea cannot be comprehensible. In the empirical study, the elapsed time for training has increased by O(NlogN). Although the number of equipment increased, its training time did not grow exponentially.

However, if the number of timestamps to refer to increases, the time needed for training and inference might go beyond the expected limit boundary. In order to cater to this, one can consider using only selective datasets that contain the most information. Only vertical usage patterns from [Fig 11] could be considered, and the results showed roughly much faster training time of 147% reduction with a 17% of performance trade-off. (Lower performance). When external conditions are similar, thus its influence on the model training is limited, the approach of smaller training can be considered as well.

Description			MSE	MAE	Training time (sec.)
			FFL	FFL	
Selective		Model only using selective usage patterns	0.0739	0.6062	1,097
Ref.	1-1	Original model	0.0531	0.5025	444



There are mainly two values for using FFL. First, Learn-as-you-go. The model does not request to wait until all the data is gathered – until the end of life. However, the model lets you have the best possible result, by evaluating the value of the information gathered as of now and approximate based on the similarities among followers. Second, the model lets you multiply unusual rare cases. When collecting data from containers worldwide, rare cases might happen – certain behaviors are seldom captured. To investigate the issues associated with the rare cases, one must gather very long-time horizon data – which is inefficient and time-consuming. If we can use the data of other similar frontiers, it will eventually make us use multiplied rare cases with minimal efforts.

Chapter 8. Conclusion / Future works

The study started by solving the real-world problem: how to efficiently build and train the models in the low data regime – especially where identical players (equipment) are in the different stages of respective life cycles. The novel idea of evaluating the value of the information fed into the model and selectively training the model was examined. Under the controlled environments of 10,000 ~ 15,000 data records for each of the cases, the novel idea, Frontier-Follower-Learning (FFL) proved its potential by showing better performance than the conventional approaches. Even when the real data was tested, FFL proved its better performance than the conventional approaches. The learnings from this algorithm can be further tested and utilized in settings where distributed systems exist: such as an electric vehicle battery management system to better forecast the state and health of each battery, airplane/vessel parts that deteriorate over multiple times all around the world, etc. Further study on the additional datasets to validate solving real-world problems is needed.

Bibliography

1. Zeng, A., Chen, M., Zhang, L., & Xu, Q. (2022). Are Transformers Effective for Time Series Forecasting?. *arXiv preprint arXiv:2205.13504*.
2. Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., & Sun, L. (2022). Transformers in time series: A survey. *arXiv preprint arXiv:2202.07125*.
3. Nakamura, T., Taki, K., Nomiya, H., Seki, K., & Uehara, K. (2013). A shape-based similarity measure for time series data with ensemble learning. *Pattern Analysis and Applications*, *16*(4), 535-548.
4. Chou, J. S., & Tran, D. S. (2018). Forecasting energy consumption time series using machine learning techniques based on usage patterns of residential householders. *Energy*, *165*, 709-726
5. Al-Ogaili, A. S., Hashim, T. J. T., Rahmat, N. A., Ramasamy, A. K., Marsadek, M. B., Faisal, M., & Hannan, M. A. (2019). Review on scheduling, clustering, and forecasting strategies for controlling electric vehicle charging: Challenges and recommendations. *Ieee Access*, *7*, 128353-128371.
6. Caillault, É. P., Lefebvre, A., & Bigand, A. (2020). Dynamic time warping-based imputation for univariate time series data. *Pattern Recognition Letters*, *139*, 139-147.
7. Wu, H., Xu, J., Wang, J., & Long, M. (2021). Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems*, *34*, 22419-22430.
8. Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, *50*, 159-175.
9. Oreshkin, B. N., Carpov, D., Chapados, N., & Bengio, Y. (2019). N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. *arXiv preprint arXiv:1905.10437*.
10. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.
11. Ekambaram, V., Manglik, K., Mukherjee, S., Sajja, S. S. K., Dwivedi, S., & Raykar, V. (2020, August). Attention based multi-modal new product sales time-series forecasting. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 3110-3118).
12. Senin, P. (2008). Dynamic time warping algorithm review. *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, *855*(1-23), 40.

13. Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., ... & Kurzweil, R. (2018). Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.
14. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, *2*(7).
15. Cuturi, M., & Blondel, M. (2017, July). Soft-dtw: a differentiable loss function for time-series. In International conference on machine learning (pp. 894-903). PMLR.

Abstract

실제 문제를 해결하는 데 AI와 데이터 과학의 개념이 인기를 끌면서 많은 응용 분야가 논의되고 있다. 그 중에서도 시계열 데이터는 실제 판매 데이터, 전기 자동차(EV) 배터리 데이터, 다양한 가전제품의 센서 데이터, 주식 시장 데이터 등에서 쉽게 찾아볼 수 있으며, 시계열 이상 징후 감지 및 향후 동향 예측 등에 광범위하게 사용된다. 본 연구는 여러 장비가 실시간으로 결과를 공유하고 서로의 과거 행동을 참고하여 자신의 미래를 예측하는 시계열 예측 문제를 해결하는 데 초점을 맞췄다. 플레이어는 Frontier (과거의 행동(결과)이 다른 사람들에 의해 학습되는 기준점이 되는 플레이어) 또는 주로 Frontier의 과거 행동을 기반으로 하는 하는 Followers로 나뉜다. Frontier와 Follower는 일회적으로 정해지지 않으며, 식시간 데이터 포인트 간의 유사성 정도에 따라 동적으로 재할당된다. Frontier의 과거 데이터 포인트는 Dynamic Time Warping (DTW)을 사용한 유사성 지수를 통해 평가되며 Frontier의 참조 데이터 포인트 정보는 Follower 모델과 유사한 정도로만 모델에 입력된다. 개념을 검증하기 위해 사례가 포함된 여러 시나리오가 실험되었다. 장비 수를 늘림으로써, 장비 간 시간 간격을 늘림으로써, 장비 간의 기본 사례, 교사-학생 네트워크 모델 (Teacher-student Network) 과의 비교, 심지어 BXB 기업의 실제 데이터를 사용한 검증까지. 결과는 정보의 가치를 평가하고 Frontier를 참조하여 모델을 동적으로 업데이트하는 새로운 개념이 더 나은 성능을 가지고 있음을 보였다. 이 개념은 여러 플레이어가 각각 과거

기록의 수가 제한되어 있지만 집합적으로 의미 있는 수준의 모델을 만들고 훈련하는 다양한 실제 현장의 문제를 푸는데 활용 될 수 있을 것이다.