American Journal of Education and Information Technology
2020; 4(2): 56-65
http://www.sciencepublishinggroup.com/j/ajeit
doi: 10.11648/j.ajeit.20200402.13

Pedagogical Data Analysis Via Federated Learning Toward Education 4.0

Song Guo[1, 2], Deze Zeng[3, *], Shifu Dong[3]

[1]Department of Computing, The Hong Kong Polytechnic University, Hong Kong SAR, China

[2]The Hong Kong Polytechnic University Shenzhen Research Institute, Shenzhen, China

[3]School of Computer Science, China University of Geosciences, Wuhan, China

Email address:

cssongguo@comp.polyu.edu.hk (Song Guo), deze@cug.edu.cn (Deze Zeng)

[*]Corresponding author

To cite this article:

Song Guo, Deze Zeng, Shifu Dong. Pedagogical Data Analysis Via Federated Learning Toward Education 4.0. *American Journal of Education and Information Technology*. Vol. 4, No. 2, 2020, pp. 56-65. doi: 10.11648/j.ajeit.20200402.13

Received: June 23, 2020; **Accepted**: July 15, 2020; **Published**: August 4, 2020

Abstract: Pedagogical data analysis has been recognized as one of the most important features in pursuing Education 4.0. The recent rapid development of ICT technologies benefits and revolutionizes pedagogical data analysis via the provisioning of many advanced technologies such as big data analysis and machine learning. Meanwhile, the privacy of the students become another concern and this makes the educational institutions reluctant to share their students' data, forming isolated data islands and hindering the realization of big educational data analysis. To tackle such challenge, in this paper, we propose a federated learning based education data analysis framework FEEDAN, via which education data analysis federations can be formed by a number of institutions. None of them needs to direct exchange their students' data with each other and they always keep the data in their own place to guarantee their students' privacy. We apply our framework to analyze two real education datasets via two different federated learning paradigms. The experiment results show that it not only guarantees the students' privacy but also indeed breaks the borders of data island by achieving a higher analysis quality. Our framework can much approach the performance of centralized analysis which needs to collect the data in a common place with the risk of privacy exposure.

Keywords: Pedagogical Data Analytics, Federated Learning, Education 4.0

1. Introduction

We are now on the cusp of education revolution to Education 4.0 where various technologies, especially the IT technologies, blend with each other to support modern and future education. Thanks to the mass infiltration of IT technologies in education, not only the students are able to learn anytime and anywhere, but also more students' data are in digital format. This potentially provides the teachers and education institution administrators the possibility of applying IT technologies to unlock the insights for more efficient education, e.g., the students' performance assessment, learning process refactoring, course design, learning process monitoring and evaluation, etc. Big education data analysis recently has attracted impressively hot concern together with the development of big data analysis related technologies, especially machine learning [1].

Along with the development of big education data analysis, the combustible mix of data analysis needs and student privacy become a crucially concerned issue. Guaranteeing the students' privacy is always of utmost importance. A well-known story is on the exposure of former US president George W. Bush's college transcript as a C student. Actually, in 1974, US issued the Family Educational Rights and Privacy Act (FERPA) as a federal law that confines the access to educational data from publicly funded educational institutions, employers, and foreign governments. In order to guarantee the students' privacy, the education institutions therefore become reluctant to share their data, and only keep the students' data locally, forming isolated data islands. Such hard “isolation” although indeed highly ensures the privacy, but makes contradiction to the big education data analysis, as which asks for large volume of data in high variety. To address such problem, pioneering researchers have already proposed various privacy-guaranteeing technologies, such as differential privacy, harmonic encryption, multi-party

computation, etc. However, they more or less have the problem like limited application scope, low performance efficiency, low analysis accuracy. How to explore the big education data to well support the education process with students' privacy guarantee thus becomes a critical challenge in Education 4.0.

To tackle such challenge, federated learning becomes a promising enabling technology. Federated learning is a machine learning technology that collaboratively learn a common model by a number of participatory servers holding data locally, without exchanging data between each other. Obviously, in contrast to traditional centralized machine learning, federated learning does not require the data owners to upload their data onto one centralized server. Such feature naturally fits for the privacy guaranteeing needs of education data, and motivates us to apply federated learning to pedagogical data analysis for Education 4.0 in this paper. The main contributions of this paper are as follows.

1) We propose a FEderated Education Data ANalysis (FEEDAN) framework by applying the advanced federated learning technology. To our best knowledge, we are the first to detail a federated learning based education data analysis framework in the literature.
2) We conduct a set of real trace-driven experiments to verify the feasibility and efficiency of our proposed framework. The experiments show that our framework can achieve the privacy guaranteeing analysis, with even higher performance in certain cases, in comparison with traditional centralized machine learning paradigm.
3) We discuss and outline several future challenges to implement and apply our proposed framework in practice, from a joint perspective of pedagogy and IT technologies.

The rest of this paper is organized as follows. We next present some related work on pedagogical data analysis in Section 2. Then, we detail the design of our federated education data analysis framework in Section 3. We also conduct a case study based on our framework and report the experiment results in Section 4. Finally, Section 5 concludes this work and outlines some future challenges.

2. Related Work

Pedagogical data analysis has been widely investigated recently, especially with the hotness of big data and machine learning technologies. For example, Masood et al. evaluate 11 representative machine learning technologies on the public students' academic performance dataset [2] and student grade prediction dataset [3] in their student's performance prediction work [4]. Their investigations show that "Decision Tree" and "Random Forest" can achieve the best accuracy in the student performance prediction. Ciolacu et al. [1] use a pivot table and accumulate all log entries for each user in each month of the semester to estimate student's performance at examination based on neural networks, SVM, decision trees and cluster analysis. The results show that non-linear kernel methods and neural networks are superior in terms of prediction accuracy. Actually, the application of various machine learning technologies in students' performance analysis has been widely discussed [5-7]. Recently, Xu et al. develop a bi-layered structure comprising multiple base predictors and a cascade of ensemble predictors to discover course relevance [8]. Different from the above studies, some works focus on the feature selection of education data. Arunrerk et al. propose three feature selection methods named Chi-square, Pearson correlation coefficient, and mutual information to identify the most significant and intrinsic features [9]. Masood et al. use a hybrid feature selection algorithm with different machine learning classifiers [4]. Both the experiment results indicate the essential of pre-processing on feature selection.

Besides performance prediction, dropout prediction also attracts researchers' interests. Kloft et al. propose a machine learning framework based on SVM for the prediction of dropout in Massive Open On-line Courses (MOOC) solely from click stream data [10] in their student dropout prediction work [11]. The random forest model and actual data set of Learning Management System (LMS) is studied by Kondo et al. to detect students at high drop-out risk early so as to intervene them effectively [12]. A computational approach using educational data mining and different supervised learning techniques to evaluate the behaviour of different prediction models in order to identify the profile of at-risk university students in a Brazilian university environment is also studied by Santos et al. [13].

By literature survey, we observe the hotness and high potential of applying various machine learning technologies in pedagogical data analysis. But obviously they all conduct locally on their own data, failing to explore the benefit of big data technology from the consideration of volume and variety. It is without doubt that more students' data imply higher accuracy in students' performance assessment and education support. Nonetheless, the privacy concern hinders the sharing of students' data between different education institutions, impeding the development of big education data analysis. To address this problem, we are motivated to apply federated learning and propose a federated education data analysis framework in [14]. In this article, we further detail the FEEDAN framework and provide more case studies via exploring two different federated learning paradigms, i.e., horizontal and vertical federated learning, to show the benefits of our framework.

3. Federated Education Data Analysis Framework

The main philosophy of federated learning is that many servers collaboratively train a common model with their own local training data under the orchestration of an aggregation server. As there is no data exchange or sharing between these servers, it migrates the risk of privacy exposure. In this section, we explore such advantage to build a FEderated EDucation Data ANalysis (FEEDAN) framework.

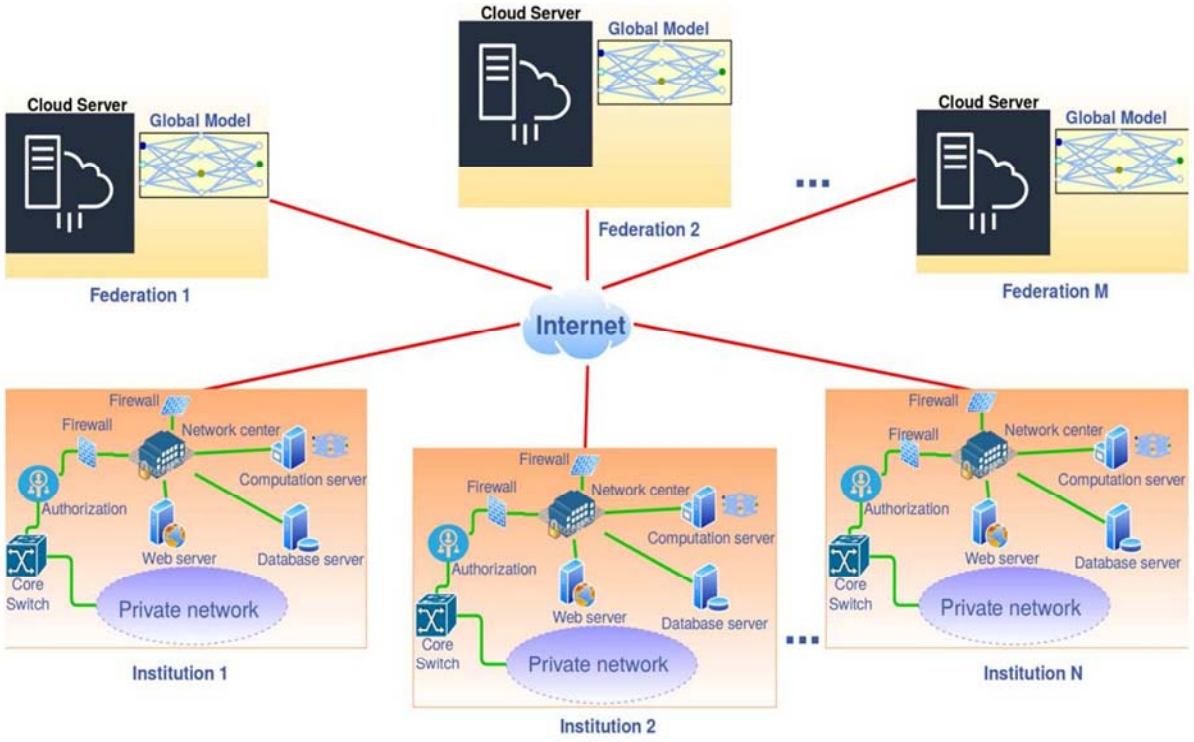


Figure 1. *Architecture Overview of FEEDAN.*

3.1. Architecture

Figure 1 presents an architecture overview of our FEEDAN framework. It can be obviously seen that there is no destructive changing to existing information system already used in some education institutions, where each only need to build a local *federated learning server* and communicate with a newly added *aggregation server*. The rests, e.g., local private network, Web server, local database server, etc., are all kept intact. Therefore, we would rather regard FEEDAN as an evolution of current education information management system. It is natural and beneficial.

As shown in Figure 1, there may coexist many different federations in FEEDAN. A federation is a group of institutions with a common goal. For example, many universities want to understand how various factors affect the students' performance. They may form a federation by contributing their own students' data for one common model training. Therefore, each education institution in the federation can be defined as a participant. A participant may participate in different federations.

Next, we first briefly introduce the key components related to privacy protection and the application of federated learning in FEEDAN.

3.1.1. Computation Server

Computation server may not be one server, but usually a cluster of servers or a mini datacenter as private cloud. Computation server already exists in existing education information systems and bears many different computation tasks, e.g., student performance analysis, office automation, financial accounting, etc. To facilitate federated learning, some computation power needs to be allocated to the federated learning task. Thanks to the wide adoption of various virtualization technology, it is not difficult to deploy a federated learning environment and make corresponding computation power allocation for federated learning training and derivation tasks.

Besides the provision of computation power, each education institution in FEEDAN also requires to deploy a federated learning environment. There are many different options available for building federated learning environment. For example, FATE is an industry-level federated learning framework developed by the Webank AI team that enables organizations to collaborate on AI while protecting data security and data privacy [15]. TensorFlow Federated (TFF) is an open source framework developed by Google for machine learning and other computing on decentralized data [16]. PaddleFL mainly focuses on the deep learning design, and provides numerous algorithms in computer vision, natural language processing, and recommendation areas [17]. Clara Federated Learning developed by NVIDIA for distributed collaborative federated learning training is also available [18]. To enable an education institution to participate in different federations, multiple federated learning environment may need to deploy on its computation server as different federations may adopt different environment to train their model. It is desirable that all participants in a federation use one uniform environment.

3.1.2. Aggregation Server

Aggregation server is used to maintain and update a global model for an education federation. Each computation server

run the federated learning training algorithm locally and send the trained results to the aggregation server, which is responsible for aggregating the trained results from different education institutions in FEEDAN. Actually, this is also the only thing that each education institution needs to communicate outside for a federated learning task. Upon the reception of trained results from a number of institutions, the aggregation server shall apply an aggregation algorithm (e.g., federated average [19]) and update the global common model. After that, the aggregation server disseminates the newly obtained global model to the participated education institutions such that they can proceed to train their local models based on the new global model. As only weight and loss values are exchanged between the aggregation server and local computation server, the privacy of the students is guaranteed.

Therefore, for each federation, there shall be an aggregation server. There is also no specific requirement that an aggregation server must be in hardware form. An aggregation server could also be a virtual machine located in public cloud. Throughout the federated learning training process, the aggregation server never obtains the actual data from the participants, implying no risk of privacy exposure. One may concern on the security of the global model as it may reside in public cloud. But thanks to the separation of model and data, even a malicious attacker obtains the global model, the model is useless without data. Besides, the network traffics between the computation servers and the aggregation servers are all encrypted on wire. The privacy is still guaranteed.

3.1.3. Database

Data are the key to any big data application or machine learning. Actually, the first revolution taken by IT technologies to education is the digitalization of students' records. Nowadays, many education institutions ranging from elementary school to universities have already built their own information systems with local database to reserve various students' data such as personal information and performance records. As discussed above, many have already applied various machine learning technologies to process these data.

On applying federated learning, there is no requirement on how the students' data are reserved, provided that they are accessible. Of course, for operability, the data are usually stored in database systems, like Microsoft SQL Server, MySQL, Oracle, PostgreSQL. FEEDAN does not require the institutions in a federation to use the same type of database system. As it does not need, or even we would rather say not allow, direct access to the other institution's student data, whether the institutions use the same type of database system or not does not matter.

3.1.4. Authentication and Authorization

The data value mining is always a contradictory issue of data privacy protection. No matter whether federated learning based data analysis is introduced or not, uncontrolled access to the students' data without suffers from privacy exposure. Like in any information system, there are mainly two aspects related to the data access control, i.e., authentication and authorization, in FEEDAN. Authentication refers to the verification of a legal user accessing to FEEDAN while authorization refers to the grant of access privileges to different data. Only an authenticated user can access certain, note that not all, data. The data that can be accessed are determined by the authorization results.

Authentication and authorization have already been built in any education information system and therefore FEEDAN can also inherit from existing system. The main things to do for applying FEEDAN is the changing of authentication and authorization policy. If a new user is decided to add specially for the management of federated learning task, a new user shall be added in the authentication policy, but this is not a must. However, the access privilege to the data needed for federated learning must be granted to the user running the federated learning task.

3.1.5. Firewall

In an information system, firewall, as an essential network security component, is used to monitor and control both the incoming and outgoing network traffic based on the predetermined security rules. A firewall could be in hardware, software, or both. FEEDAN therefore can also inherits firewall from existing system. It establishes a barrier between the local private network and the external Internet, where the former is usually regarded as trusted but the later as untrusted. Thus, firewall is essential to guarantee the security of the students' data, no matter whether federated learning is applied or not. With federated learning, all the training tasks are performed behind the firewall and only the trained results, e.g., neural network weights and loss value, will go through the firewall to the external aggregation server. The communication between the computation server and the aggregation server is also under the monitoring and management of the firewall to guarantee that no students' data are sent out.

3.1.6. Web Server

Many information systems now usually adopt the “MVC” (i.e., Model, View, and Controller) design structure where Web server provides a cross-platform user interface (i.e., View). Web server is also widely used in existing education information systems. Via Web browser, authorized users can access their wanted and access granted data in different visual formats. After the introduction of federated learning, Web server also plays an important role as it can offer the users with both the training process visualization as well as the derivation results. Actually, many federated learning solutions (e.g., TFF) as discussed above already provides the access portal in Web browser for training process monitoring. Depends on the system development needs, the derivation results can be visualized in Web browser freely.

3.2. Working Process

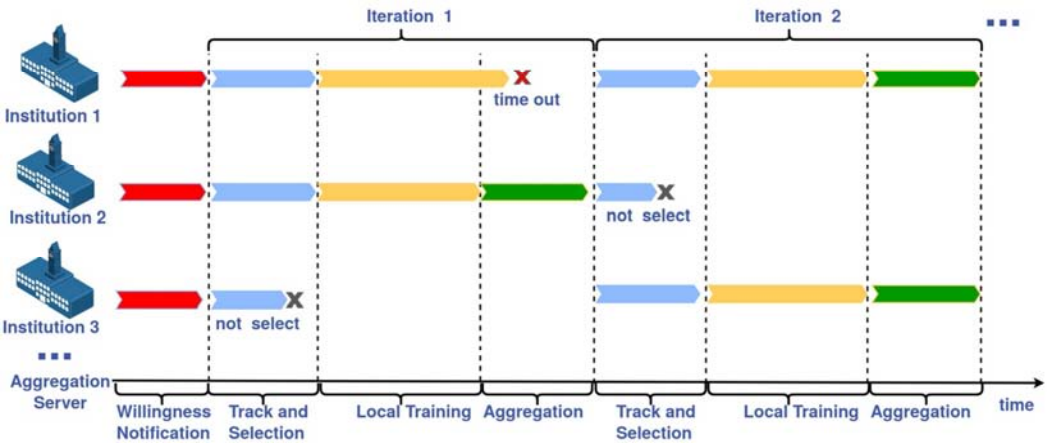


Figure 2. *The Federated Training Procedures in FEEDAN.*

After the introduction of the architecture, now let us proceed to the federated learning based data analysis process in FEEDAN. The main difference between federated learning and traditional centralized machine learning is on how the model is trained. There is no difference on how the model is used for derivation. Therefore, we focus on the training procedures in FEEDAN, whose main phases are illustrated in Figure 2. After the invention of federated learning, there are many different training paradigms differing on how the participants collaborate with each other. Here, we discuss a representative classical one, i.e., horizontal and synchronous, which is easy for understanding.

Once a federation is formed, the participants may start to train their common model collaboratively with the following phases. At first, the participants notify their willingness in participating the model training to the aggregation server, which will then start the training and manage the participants throughout the training process. Similar to centralized training, the federated training is also conducted in an iterative manner until the predetermined criterion is achieved, i.e., converged. At the beginning of each iteration, the aggregation server tracks the statuses (e.g., liveness, available computation capacity) of the participants and selects a number of participants for training in current iteration. Note that not all participants need to participate the training in each iteration. A participant not participating the training in one iteration may also be able to be involved in future training iteration. Then, the selected participants will be disseminated with the global model for local training.

Upon the reception of the global model, the computation server in each participated education institution then starts to train the model using its local data. Our FEEDAN framework does not impose any restriction on the local training methods, provided that all the participants adopt the same method. That is, different federation may have different training algorithm but one federation shall use the same one. For example, a federation may require the participants to train the model using SGD-style algorithm. After certain predetermined time or when some local criteria are met, the local training results (e.g., weight and loss values) shall be sent to the aggregation server.

Upon the reception of the training results from the participants, the aggregation server can then aggregate them into the global model using various aggregation methods such as federated average. FEEDAN does not impose any restriction on the aggregation algorithm either. A federation can freely choose the desired aggregation algorithm according to their needs. After aggregating the results from the selected participants, the aggregation server will check whether the predetermined convergence criterion is met or not. If met, the training process can be stopped; otherwise, the above procedures will continue to proceed.

4. Case Study and Analysis

In order to show the feasibility and the efficiency of FEEDAN framework in education data mining, we apply it to analyze two real education datasets using horizontal and vertical federated learning, respectively. The two federated learning paradigms mainly differ in whether the data in different clients share the same structure (i.e., with the same features) or not [20]. In this section, we will investigate how the two types of federated learning can be applied in FEEDAN to perform pedagogical data analysis. The datasets we used are KDDcup2010 [21] and KDDcup2015 [22]. With KDDcup2010 dataset, we apply horizontal federated learning to build model to predict the student performance. By using KDDcup2015 dataset, we use both the horizontal and vertical federated learning to predict the dropout of students.

4.1. Dataset Introduction

KDDcup2010 dataset comes from Intelligent Tutoring Systems (ITS) used by thousands of students over the course of the 2008-2009 school year. There are 30 million training rows and 1.2 million test rows in total. For this pedagogical federated training task, we use the Algebra-2008-2009 dataset with information recorded by the tutoring system.

The main feature information before preprocessing about the KDDcup2010 dataset is shown in Table 1. Overall, this dataset contains 17 features and one binomial label (correctFirstAttempt), indicating whether the student correctly solve the problem at the first attempt or not. Our goal is to apply machine learning to predict whether students can solve a problem in their first attempt.

Table 1. *Main Feature Information about KDDcup2010.*

Main feature name	Description
knowledgeComponent	The knowledge contained in the problem
stepDuration (sec)	Time used to solve each step of the problem
correctStepDuration (sec)	Time used on correct step
errorStepDuration (sec)	Time used on error step
hints	Number of hints to solve the problem
corrects	Correct number to solve the problem
incorrects	Incorrect number to solve the problem
correctFirstAttempt	Binomial label

KDDcup2015 dataset provides partial period information of 39 courses during the half year of MOOC, and can be used to study students' dropout behaviors. This dataset mainly includes five CSV files and the information about these files is shown in Table 2. Each line in *object.csv* file describes a module in one course, including its category, submodules, and the time of publication. These modules represent different parts of the course, such as chapters, sections, online video materials, problem sets, etc. In *enrollment_train.csv*, each line is about the information on a student attending a course, and contains the enrollment_id, username and course_id. In *log_train.csv*, each line is a behavior record of an "event", which mainly contains the enrollment_id, time (event source) and event (problem, video, access, wiki, discussion, navigate, page_close and object). In *true_train.csv*, each line records whether a student with enrollment_id dropout or not.

Table 2. *File Information About KDDcup2015.*

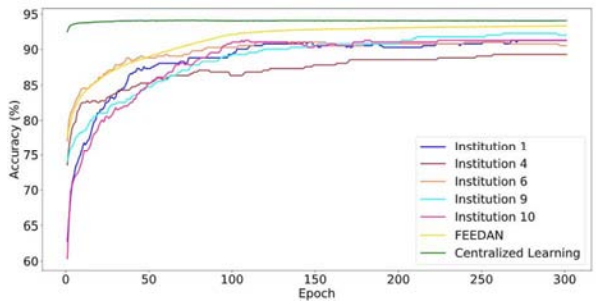
File name	Size (KB)	Description
enrollment_train.csv	8646	Registration number (training set)
log_train.csv	603782	Learning log (training set)
truth_train.csv	995	Dropout label (training set)
enrollment_test.csv	5765	Registration number (testing set)
log_test.csv	398863	Learning log (testing set)
object.csv	3062	Course and module information
date.csv	3	The earliest and latest log data for the courses
sampleSubmission.csv	663	Template for submission

Due to the irregularity of education data, it is hard to directly apply them for machine learning based training. We therefore first preprocess these data by removing some useless data and normalizing the data values. After preprocessing, KDD2010 dataset contains 60,000 records, each record contains 5 features. KDD2015 dataset includes 100,000 records and each record has 116 features.

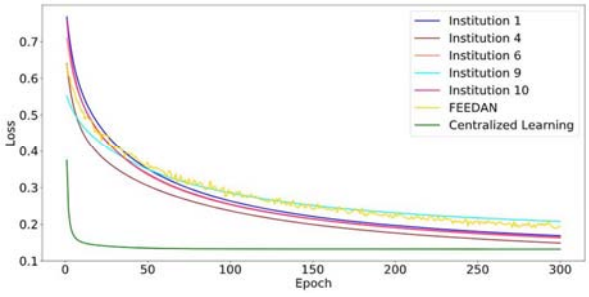
4.2. Horizontal Education Federated Learning Experiment

Horizontal federated learning is mainly introduced in the scenarios in which datasets share the same features but with different samples. Different education institutions may have different students, but they record the same information about the students. To analyze such dataset, we can apply the horizontal federated learning. According to the horizontal federated learning process, we emulate a federation consisting of a number of institutions by dividing the preprocessed data uniformly among these institutions. For example, we divided datasets into 100 education institutions randomly. For student performance prediction, each institution has 600 records. For student dropout prediction, each institution has 1000 records. The preprocessed data are trained with Logistic Regression and a two-layer Neural Network, respectively.

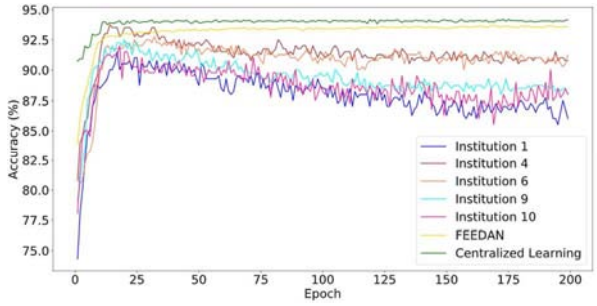
In FEEDAN, we apply FedAvg [23] for federated learning. The default settings of federated learning in the experiments are as follows. The institution participation probability in each training round is set as 10%, indicating that averagely there are 10 institutions in each training round. The learning rate, the local epoch, and batch size are set as 0.001, 1, 50, respectively. To show the efficiency of FEEDAN, we also compare the performance efficiency of our framework against two competitors. One is aggregating all the data into a central server and another one is conducting the model training locally with own data only.



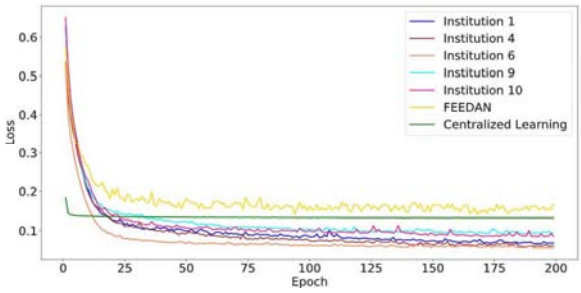
(a) On the accuracy comparison using logistic regression



(b) On the loss comparison using logistic regression

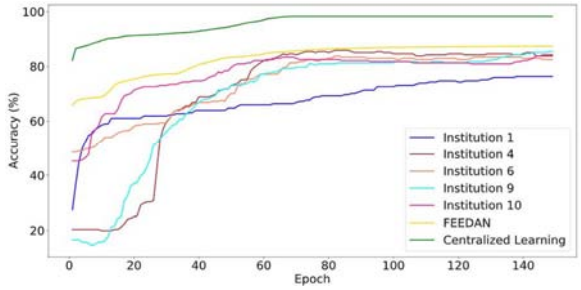


(c) On the accuracy comparison using two layer Neural Network

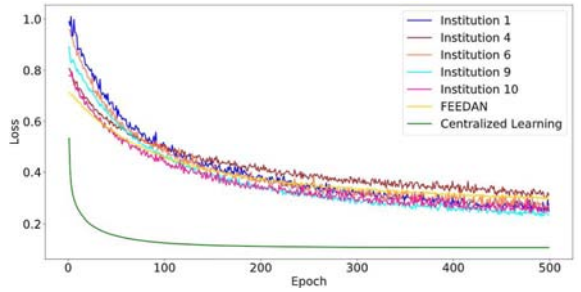


(d) On the loss comparison using two layer Neural Network

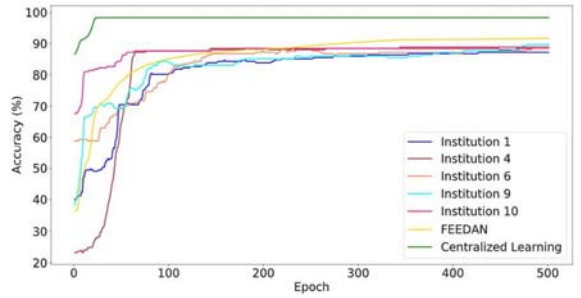
Figure 3. *On the student drop out prediction performance comparison during horizontal federated learning training.*



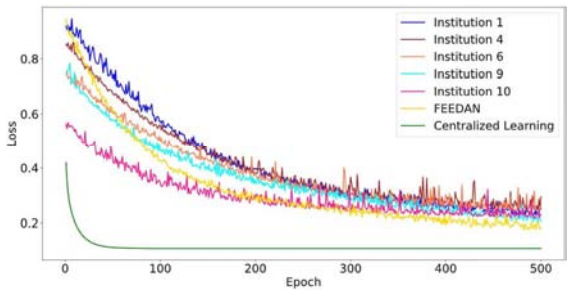
(a) On the accuracy comparison using logistic regression



(b) On the loss comparison using logistic regression



(c) On the accuracy comparison using two layer Neural Network

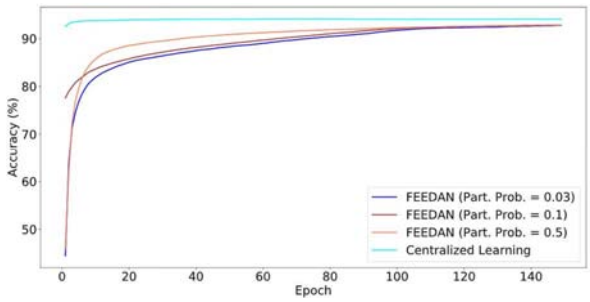


(d) On the loss comparison using two layer Neural Network

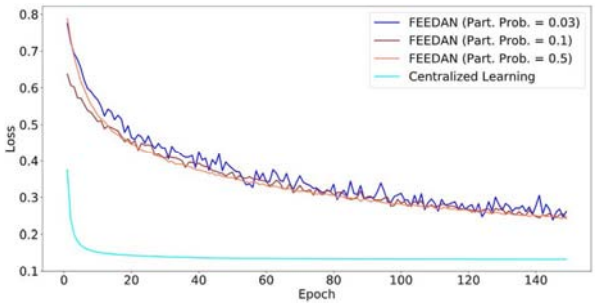
Figure 4. *On the student performance prediction result comparison during horizontal federated learning training.*

We run all the algorithms for 500 epochs and evaluate their accuracy and loss values. The training results are reported in Figure 3 and Figure 4. For tractability, we randomly select some institutions and report their local training results. Figures 3 (a), 3 (c) and Figures 4 (a), 4 (c) give the accuracy achieved after each epoch. We can see that the accuracy gradually increases with the training epoch, for any training method. This first verifies the feasibility of FEEDAN as it indeed can provide accurate students' dropout prediction. Furthermore, we shall notice that centralized training by aggregating all the data onto one server achieves the best performance as the accuracy converges fast. While, FEEDAN exhibits better performance than any other case with local data only training. FEEDAN gradually approaches the performance of centralized training after certain training epochs. This is attribute to the larger volume of data with higher variety used during the training. With more training epochs, the data used also gradually approach the one used in centralized training case.

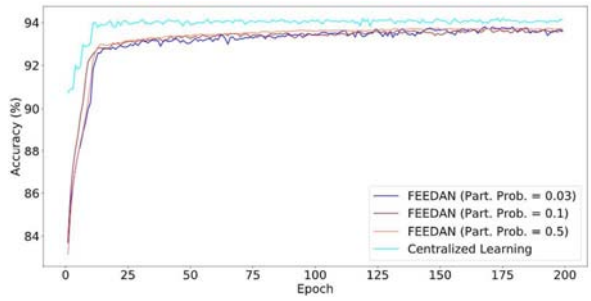
We envision that more participants in a federation implies higher model training quality. To verify it, we also conduct a series of experiments with different participation probabilities (i.e., different federation scales). Figures 5 (a), 5 (c) and Figures 6 (a), 6 (c) report the evaluation results on the accuracy and loss during the training of FEEDAN. As shown in Figure 5 and Figure 6, obviously faster convergence speed can be observed with higher participation probability. Among the three probabilities, i.e., 0.03, 0.1 and 0.5, the best performance is observed when the probability is 0.5. This verify that recruiting more participants in a federation is beneficial to the global model training, and therefore is more benefits to all the participants in the federation.



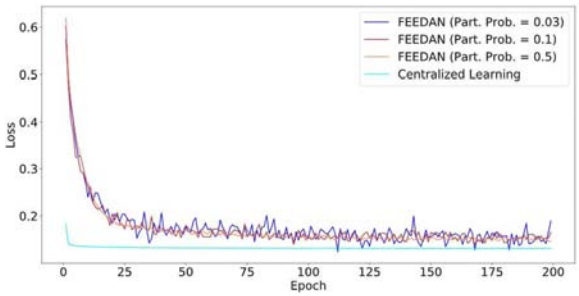
(a) On the accuracy comparison using logistic regression



(b) On the loss comparison using logistic regression

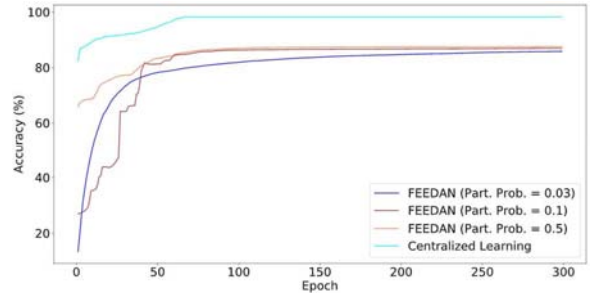


(c) On the accuracy comparison using two layer Neural Network

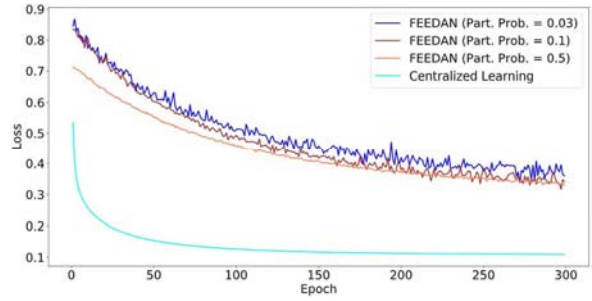


(d) On the loss comparison using two layer Neural Network

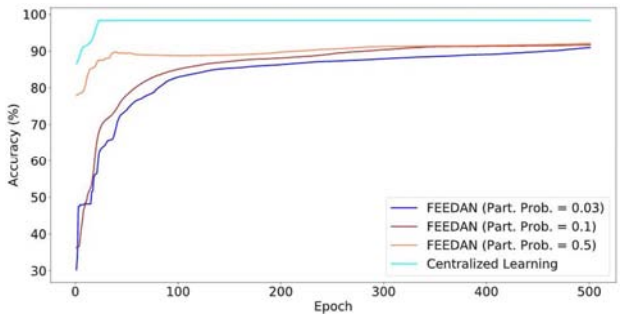
Figure 5. *On the effect of federation scale to student drop out prediction.*



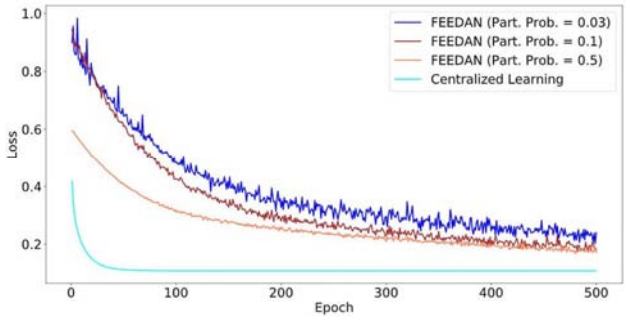
(a) On the accuracy comparison using logistic regression



(b) On the loss comparison using logistic regression



(c) On the accuracy comparison using two layer Neural Network



(d) On the loss comparison using two layer Neural Network

Figure 6. *On the effect of federation scale to student performance prediction.*

4.3. Vertical Education Federated Learning Experiment

Vertical federated learning is applicable to the cases in which two datasets share the same sample ID but differ in features. For example, a student may be recruited in different education institutions, e.g., school and on-line learning platform. The school shall have some pedagogical records of the students, and the on-line education platform also have some other different records. A lot of students may register in both the school and on-line education platform and therefore the two education institutions may have different records of the same student. In order to use both records to analyze the pedagogical data while protecting students' privacy, we can apply vertical federated learning to collaboratively analyze the data in different institutions. In the training process, we use the KDDcup2015 datasets and apply a multi-layer neural network to assess the students' dropout probabilities. We divide the 116 features of each sample into two institutions, each institution has 58 features.

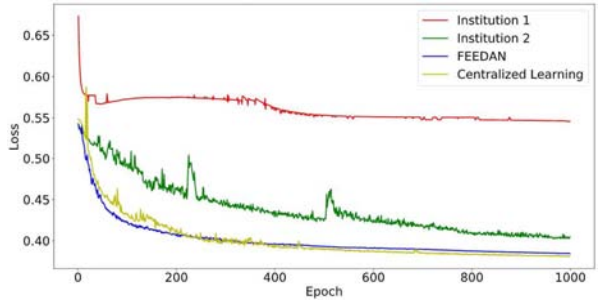


Figure 7. *On the loss comparison of vertical education federated learning.*

Similarly, to show the efficiency of FEEDAN using

vertical federated learning, we compare the performance efficiency of our framework against two competitors. One is aggregating all the data into a central server and another one is conducting the model training locally with own data only. We run these algorithms for 1000 epochs and evaluated their loss values. The training results are reported in Figure 7. We can see that the loss value gradually decreases with the training epochs for any training method except training locally at institution 1. Meanwhile, we can also notice that FEEDAN's training loss decreasing behaves like the centralized training method. This verifies the feasibility of applying vertical federated learning as it indeed can deal with the case that the same student's records involved in different institutions' dataset, without exposing students' privacy. We can also observe that vertical FEEDAN exhibits better performance than any other case with local data only training. With the training epoch increasing, FEEDAN gradually approaches the performance of centralized training.

5. Conclusion and Future Work

In this paper, to address the isolated education data island problem, we propose a federated education data analysis framework FEEDAN via the adoption of federated learning. Based on the proposed framework, we also conduct a case study to assess the students' performance using public education dataset. The results verify that it is feasible to apply FEEDAN to analyze students' data collaboratively, without exposing students' data privacy at the same time. In addition, the benefits of breaking the borders between the data island is also noted by the fact that our framework achieves higher performance prediction accuracy than individual training on only own local data. We believe that FEEDAN provisions a promising potential pedagogical data analysis solution to pave the way for forming education federation in Education 4.0. It shall be helpful in improving educational practice and improve the student performance.

While, at the early stage of federated learning and intelligent education data analysis, there are still remaining challenges to be addressed in the future work such as data structure standard, public learning model design, data alliance forming, which are need to studied further from a joint perspective of pedagogy and IT technologies.

1) Data structure standard: For an education federation, more participants imply larger volume of data with higher diversity, and thus higher model quality. However, as the education institutions reserve their data locally, different institutions may use different data formats. Obviously, it is desirable to use a unified data format to reserve the data such that they can be directly applied for training in a federation. Even when the data cannot be directly applied, using a unified data format enables a federation to easily determine the data fields used in the model. Without such standard, the missing of a data field may make an education institution fail to participate in a federation, hindering the proliferation of an education federation.
2) Public learning model: Different education federations are with different needs and requirements on the data. However, we also believe that there exist some public needs asking for the participation and contribution from a large number of education institutions. It is desirable to build a public learning model that any education institution can participate in. However, this is also a non-trivial task. Firstly, it requires the experts in pedagogy to propose a potential public need that is beneficial to many education institutions. Secondly, it requires the IT experts to discuss the feasibility and possible solutions towards such need.
3) Data alliance forming: The willingness of a participant to join a federation is affected by many factors. Even education is a public affair, we still have to admit that there also exists competition between different education institutions. Although it seems that forming an education federation is beneficial on building a powerful model for better education activity assistance, the education institutions may have their own considerations. Therefore, it is significant, but also challenging, to form a data alliance where the members can “share” their local data for some common goal that can be achieved by federated learning.

Acknowledgements

This work is supported by Shenzhen Basic Research Funding Scheme (JCYJ20170818103849343), China University of Geosciences (Wuhan) Higher Education Reform Fund Project (YJG2019103).

References

[1] M. Ciolacu, A. F. Tehrani, R. Beer, and H. Popp. Education 4.0 fostering student’s performance with machine learning methods. In 2017 IEEE 23rd International Symposium for Design and Technology in Electronic Packaging (SIITME), pages 438–443, 2017.

[2] “students’ academic performance dataset”. [Online]. Avaliable: https: //www.kaggle.com/aljarah/xAPI-Edu-Data.

[3] “student grade prediction”. [Online]. Avaliable: https://www.kaggle. com/dipam7//student-grade-prediction.

[4] M. F. Masood, A. Khan, F. Hussain, A. Shaukat, B. Zeb, and R. M. Kaleem Ullah. Towards the selection of best machine learning model for student performance analysis and prediction. In 2019 6th International Conference on Soft Computing Machine Intelligence (ISCMI), pages 12– 17, 2019.

[5] S. C. Harris and V. Kumar. Identifying student difficulty in a digital learning environment. In 2018 IEEE 18th International Conference on Advanced Learning Technologies (ICALT), pages 199–201, 2018.

[6] M. Hussain, W. Zhu, W. Zhang, J. Ni, Z. U. Khan, and S. Hussain. Identifying beneficial sessions in an e-learning system using machine learning techniques. In 2018 IEEE Conference on Big Data and Analytics (ICBDA), pages 123–128, 2018.

[7] E. Tanuar, Y. Heryadi, Lukas, B. S. Abbas, and F. L. Gaol. Using machine learning techniques to earlier predict student's performance. In 2018 Indonesian Association for Pattern Recognition International Conference (INAPR), pages 85–89, 2018.

[8] J. Xu, K. H. Moon, and M. van der Schaar. A machine learning approach for tracking and predicting student performance in degree programs. IEEE Journal of Selected Topics in Signal Processing, 11 (5): 742–753, 2017.

[9] J. Arunrerk W. Punlumjeak, N. Rachburee. Big data analytics: Student performance prediction using feature selection and machine learning on microsoft azure platform. In Journal of Telecommunication, Electronic and Computer Engineering, volume 9, pages 113–117, 2017.

[10] Katy Jordan. "mooc completion rates: The data.". [Online]. Avaliable at: http://www.katyjordan.com/MOOCproject.html.

[11] Marius Kloft, Felix Stiehler, Zhilin Zheng, and Niels Pinkwart. Predicting mooc dropout over weeks using machine learning methods. pages 60–65, 2014.

[12] N. Kondo, M. Okubo, and T. Hatanaka. Early detection of at-risk students using machine learning based on lms log data. In 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAIAAI), pages 198–201, 2017.

[13] K. J. de O. Santos, A. G. Menezes, A. B. de Carvalho, and C. A. E. Montesco. Supervised learning in the context of educational data mining to avoid university students dropout. In 2019 IEEE 19th International Conference on Advanced Learning Technologies (ICALT), volume 2161- 377X, pages 207–208, 2019.

[14] Song Guo, Deze Zeng. Pedagogical Data Federation toward Education 4.0. In the 6th International Conference on Frontiers of Educational Technologies (ICFET 2020), June 5–8, 2020, Tokyo, Japan. ACM, New York, NY, USA, 5 pages.

[15] https://github.com/FederatedAI/FATE.

[16] https://github.com/tensorflow/federated.

[17] https://github.com/PaddlePaddle.

[18] https://blogs.nvidia.com.tw/2019/12/clara-federated-learning/.

[19] Chuan Ma, Jun Li, Ming Ding, Howard H Yang, Feng Shu, Tony QS Quek, and H Vincent Poor. On safeguarding privacy and security in the framework of federated learning. IEEE Network, 2020.

[20] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. ACM Trans. Intell. Syst. Technol., 10 (2), January 2019.

[21] Yu, Hsiang-Fu, Hung-Yi Lo, Hsun-Ping Hsieh, Jing-Kai Lou, Todd G. McKenzie, Jung-Wei Chou, Po-Han Chung et al. "Feature engineering and classifier ensemble for KDD cup 2010." *KDD cup* 11, 2010.

[22] SIGKDD. "sigkdd, kdd cup 2015-predicting dropouts in mooc.". [EB/OL]. http://www.katyjordan.com/MOOCproject.html.

[23] H Brendan Mcmahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera Y Arcas. Communication-efficient learning of deep networks from decentralized data. pages 1273–1282, 2017.