Journal of Network and Systems Management (2021) 29:42
https://doi.org/10.1007/s10922-021-09609-5

A Federated Learning Approach to Frequent Itemset Mining in Cyber-Physical Systems

Usman Ahmed[1] · **Gautam Srivastava**[2,3] · **Jerry Chun-Wei Lin**[1]

Received: 26 November 2020 / Revised: 23 April 2021 / Accepted: 22 May 2021 /
Published online: 1 June 2021

Abstract

Effective vector representation has been proven useful for transaction classification and clustering tasks in Cyber-Physical Systems. Traditional methods use heuristic-based approaches and different pruning strategies to discover the required patterns efficiently. With the extensive and high dimensional availability of transactional data in cyber-physical systems, traditional methods that used frequent itemsets (FIs) as features suffer from dimensionality, sparsity, and privacy issues. In this paper, we first propose a federated learning-based embedding model for the transaction classification task. The model takes transaction data as a set of frequent item-sets. Afterward, the model can learn low dimensional continuous vectors by preserving the frequent item-sets contextual relationship. We perform an in-depth experimental analysis on the number of high dimensional transactional data to verify the developed models with attention-based mechanism and federated learning. From the results, it can be seen that the designed model can help and improve the decision boundary by reducing the global loss function while maintaining both security and privacy.

Keywords Federated learning · Frequent itemset mining · Cyber-physical systems · privacy · Data mining · Embedding

✉ Jerry Chun-Wei Lin
jerrylin@ieee.org

Usman Ahmed
usman.ahmed@hvl.no

Gautam Srivastava
srivastavag@brandonu.ca

[1] Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Bergen, Norway

[2] Department of Mathematics and Computer Science, Brandon University, Brandon R7A 6A9, Canada

[3] Research Center for Interneural Computing, China Medical University, Taichung 40402, Taiwan, ROC

1 Introduction

Data is deemed as a major asset to explore different facts about the entities associated with it. For instance, medical centres, engineering, marketing, sports, warehouses, and cyber-physical systems contain a rich amount of data about different objects or individuals representing various kinds of information [1]. Data mining specifically focuses on extracting interesting patterns or "knowledge" from data through rigorous analysis. Various real-life applications demand to mine interesting patterns from data [1]. Currently, the scientific community has been interested in data mining techniques that pertain to pattern mining like frequent pattern mining (FPM) [2], association rule mining (ARM) [3], frequent episode mining (FEM) [4], and sequential pattern mining (SPM) [5]. These techniques' prime concern is to mine patterns from real-world applications by harnessing co-occurrence, frequency, and interestingness measures.

Cyber-physical systems (CPS) are known to be as strong at their ability to integrate computation with physical processes. Certain unique data mining problems emerge in the design and operation of CPS-based data sources. Unlike the Internet of Things (IoT), data from CPS often involves interactions between "cyber" and "physical", which tend to be more closely integrated into CPS, as seen with the databases used in them. A plethora of these techniques produces promising results in most scenarios. However, their applicability cannot be generally considered to be able to handle all kinds of data. The types of databases are often slightly different such as binary databases, quantitative databases, probabilistic databases, stream databases, and fuzzy databases, among others. Mining information from these types of databases is a non-trivial process that requires contemplation of various essential constraints. Most of the existing state-of-the-art algorithms in pattern mining focus on handling binary databases because they have comparatively less complicated structures than other types of databases. Furthermore, existing techniques return a bulk of redundant information in response to the model employed for the pattern mining task. The redundancy makes the knowledge or models larger and more challenging to use or understand by humans.

1.1 Motivation

CPS combines the dynamics of physical processes with those of software and communication. These interwoven concepts provide abstractions and modeling, design, and analysis techniques for the integrated parts [6]. Fifth-generation mobile networks (5G) are currently being deployed and likely to replace 4G in advanced countries in CPS environments [7]. Furthermore, we see Sixth-generation mobile networks (6G) on the horizon, which has led to an excess of security and privacy issues concerning data. The higher bit rates available in 5G/6G will enable the use of software-defined networking (SDN) techniques while providing faster data transfer rates. Although advantageous for faster transfer speeds, these faster rates can lead to bigger issues from a cybersecurity point of view [8]. CPS models are only reliable

when 5G/6G networks perform well, especially all the services that are deployed on cloud servers.

With CPS in 5G/6G, higher capacity and very low latency will allow applications designed for IoT networks to connect with data centres. This has and will lead to a fully mobile and connected society which has been the goal all along. From this, the data in databases can be dynamically inserted, deleted, and/or modified. Conventional methods used to handle generic data operate by updating the discovered information by re-scanning the updated database and re-mining the required information. However, this is a theoretical approach as it could increase the time complexity in updating databases for real-time decision making. There is a need for an ideal approach that can intelligently handle the three aforementioned dynamic situations (insertion, deletion, and modification). Thus, a reliable and stable CPS would be better to handle this situation and perform all dynamic operations locally, even if the 5G/6G network is not available. There is a need for an ideal approach that can intelligently handle these three dynamic situations.

Secondly, mobile nodes with personal information can be tracked down and become vulnerable through known attacks like eavesdropping, denial of service, replay, and repudiation [9, 10]. The higher speed requires a high quality of service (QoS) in terms of execution time, while a higher volume of data is transferred. Data-intensive devices' deployment to 5G heterogeneous networks requires addressing data privacy and security more seriously. Sensitive data is associated with several applications. For example, considering customer purchases and their location as personal information, this type of data is private and confidential and should be secured and preserved. In this scenario, an edge computing network is required to carry out some sensitive tasks without transmitting the data to central servers, i.e., a cloud server. In academia, the community has not looked into the ability to synthesize information from multiple heterogeneous data sources. Considering the various aspects of the learning environment's implementation can significantly help in evading the above-stated issues.

Machine learning-based classification models are also used to solve classification tasks in different domains. Random forest is an ensemble classifier that contains the number of decision tree-based models [11]. A random number of features is divided among each tree based classifier. The voting mechanism is adopted while predicting the unknown class. *K*-Nearest Neighbour (KNN) [12] is known to be the most straightforward classification technique, and it is an instant based learning, also known as a lazy learning classification [13]. Iterative Dichotomiser 3 (ID3) uses the information to select the attribute and classify the current subset of instances. For each level node, information gain is calculated recursively [14]. Another ID3 based approach is C4.5 [15], which uses the information gain and gain ratio to select attributes. The main advantage of C4.5 over ID3 is that it handles both continuous and missing attribute values.

In transactional data, a set of distinct and discrete items exist. The items represent different patterns in many different domains and applications. For example, in supermarket basket analysis, items represent the purchases of the products. In health data, items represent symptoms diagnosed during the patient's admission. Different applications have different orientations and concept of the items. Extracting useful

information from the analysis of the item is a challenging task. A common solution in data mining is the usage of the frequency of the items as features. The threshold values are being used as criteria for pattern extraction. The bitmap representation is used to represent items. If the item exists, then the binary value is assigned to it; otherwise, it will set it zero. The frequent item(sets) are represented with huge vector space since the number of items(sets) is often large, especially when data is collected from CPS environments. It is resulting in the curse of high dimensionality issues and data sparsity problems.

Researchers solve the curse of dimensionality and data sparsity problems by using measures [16]. However, the above approaches may have three limitations. Firstly, they consider the representation of transactional data for particular tasks, i.e., transaction classification. As a result, they cannot transfer the patterns or learning mechanisms to other different tasks. Secondly, the methods require the number of instances structured in the database. However, real-world applications are often dynamic, and some domain applications, such as stream scenarios, do not allow the re-scanning of databases. Third, the models do not tackle the problem of extracting patterns without using actual data. The databases often collaborate in centralized structures. This structuring often results in overhead as it requires time-consuming approvals because of data privacy and ethical concerns associated with data sharing of personal records. Even when these challenges are being addressed, data sets are valuable for organizations, so they prefer not to share full data sets. Furthermore, the datasets of mobile networks or IoT networks are often very large and often expensive to acquire central hosting storage. Consequently, a federated learning approach [17] can tackle the above issues, where only model weights are shared across the network without the raw data information. In this paper, we assume at least one cycle of federated learning is being achieved and models share their weights to the server. If the network is down, then the local model provides an embedding and uses it for the prediction until the network is online again.

1.2 Contribution

To overcome the static FIM-based model's weakness, we first propose an attention-based federated learning approach in this paper. To the best of our knowledge, this is the first attempt at this sort of federated learning pattern mining implementation application in IoT, mobile, and other modern computer networks. For transactional data, the proposed model can perform learning capacity for low dimensional representation (embedding) in a fully unsupervised manner. We use frequent itemsets as an input vector and then reduce it to low dimensional space using a deep learning attention-based model. Next, we apply the learned embedding to the different transaction classification tasks. The locally trained model and model sharing of federated learning helps to increase performance globally as well locally. The embedding helps to capture the association among individual items to produce the low dimensional continuous vectors. Therefore, based on the developed attention-based mechanism, it helps to capture the relationships among the transactions. In short, we summarize our contributions as follows:

1. We propose an attention-based transaction embedding model for transactional data to learn low dimensional continuous representation.
2. We prove that federated learning helps to improve the performance of learned embedding without raw data sharing.
3. We demonstrate that the developed embedding model helps in transaction classification tasks on several benchmark datasets.

2 Related Work

Over the past few decades, pattern mining algorithms [1] have shown effectiveness to discover valuable information for decision-making. Mining information from them is a non-trivial process that requires contemplation of various important constraints [5]. This section gives a thorough review of works related to mining information from databases and current state-of-the-art techniques.

2.1 Association-Rule Mining

Association rule mining (ARM) is a fundamental part of Knowledge Discovery in Database (KDD). It has been used extensively to ascertain association rules among different item sets in a database for decision-making. The first algorithm is called Apriori [18], which Agrawal and Srikant proposed to find the relationships of the items in the databases. Apriori uses two phases to first find the set of frequent itemsets in databases based on the minimum support threshold. After that, the combinations of the frequent itemsets are formed to generate the set of association rules based on a minimum confidence threshold. Many extensions [19] have been presented to handle ARM regarding the occurrence frequency for the efficiency problem. However, ARM determines the relationship among items of binary databases, thus ignoring important factors like weight, importance, interestingness, or quantity. For instance, ARM does not consider the price information of different items in a database. Furthermore, Apriori gives equal weight to all the items in a database which could cause poor decision making.

2.2 Classification

Data mining and classification are employed for market basket analysis wherein past transactions are explored to identify patterns that are likely to be purchased together. The techniques utilized to mine the data are association, sequential pattern analysis, clustering, and classification [20]. In association, associativity between two or more items is discovered based on their past relationship [5]. For instance, if a customer purchases a monitor screen, then there is a high probability that he/she will purchase a keyboard and association determine the extent of the relationship between a monitor and a keyboard. In market basket analysis, the association is used to determine items with a high probability of being purchased together by a customer. Another primary data mining technique is the discovery of sequential pattern wherein items

to be bought in a specific sequence or pattern by a customer is identified based on the historical sequence of items in the transaction [1]. Cheng et al. proposed an frequent itemset (FI) mining approach where discriminated patterns are extracted by using the information gain [16]. The same approach was adopted and selected by other researchers by using the support ratio, and difference [21]. FI helps to build rule-based classification, often called associated classification. It is based on a strong association between FI and labels. Embedding based learning was introduced in 2013 in the famous word2vec model [22]. The embedding vectors are learned based on the co-occurrence of the words in textual data. The same methods are adopted for networks [23] as well as transaction data [24]. However, federated learning for transaction data has not been studied to date, which will be considered as the main novelty in this paper.

2.3 Deep Neural Networks

A neural network can be used to learn hidden patterns and extract high dimension features from databases [25]. A neural network model uses the learned features for a conditional distribution of the input vector and output class. Different neural networks are then used based on specific problems and domains. The most common method is multi-layer perceptron architecture. In this network, each hidden layer takes input from the previous layer and then computes weights by taking averages. The final output is a non-linear activation function to the total input values. The learning task is a non-linear optimization problem that updates the weights based on the loss and gradient values. In supervised learning, the objective is to reduce the loss function by using input weights and bias values. The algorithms mostly fall under the category of gradient descent. In this method, the algorithm starts from random points for each input feature. It is then executed for several iterations called epochs that are used for a batch of instances. The trainers compute the loss values and gradient by using a non-linear objective function. Then, weights are updated in a way that reduces the loss function. This process leads to a convergence point or optimum local error. The predictive ability of neural networks comes from hidden layers of the structure. The correct selection of architecture and hypermeter results in higher performance. The network training helps to scale features and combine them in higher-order representation [26]. The higher representation helps to increase the predictive power and achieve generalizations.

Different architectures have been proposed over the past two decades [27]. The difference in neural architecture is in terms of hidden layers, layers type, layer shapes, and connection between layers [28]. To learn higher dimension features from tabular data, Wainberg et al. used fully connected networks [29]. The convolutional neural networks (CNN) learn feature embeddings from image pixels. The translation invariant pixel data brings benefits to the architecture and increases learning compatibility. Many studies were respectively presented in this area of research for solving distinct problems related to wildlife [30], X-ray scans [31] and autonomous driving [32], to name just a few. Several deep learning architectures were proposed

and used in the domain of natural language processing (NLP) [33] including Neural Machine Translation [34] and time series analysis [35].

Typically, RNN is an encoder and decoder framework where the encoder takes the input sequence and decodes into the vector's fixed length. RNN uses different gates to process the input features based on the loss function. However, while compressed, the model loses relevant information [26]. Another issue with the RNN encoder and decoder model is the alignment of input and output vectors. Neighbours' feature values influence the resulting sequence. The decoder lacks a selective focus on input features. Another variant for RNN is an attention-based sequence model [26]. The idea behind the attention in this model is to allow complete input sequence usage and then apply attention-based weights on selective input sequences to prioritize the importance and position of relevant information. Afterward, the decoder uses the position, context vector, and corresponding weights for the higher feature representation, then feed the learned weights to the RNN model for further usage [26]. The attention model takes multiple inputs and joins the learned attention weights [36]. There is also the soft attention model [37]. The model used the weighted average of the hidden states and then built a context vector. The method helps the neural network to efficiently learn the hidden pattern and reduce the loss function. The hard attention model was proposed by Xu et al. [38] that computes the context vector from sampling hidden states in the input vectors. The hard computation reduces the computation cost. However, the architecture's convergence is difficult to achieve. Luong et al. [39] proposed a local and global attention method. The global attention architecture is like soft attention, and global attention is an intermediate between soft and hard attention. The model picks an attention point or position of input features for each round. This creates a local attention model. The predictive function learns the attention position. The global attention takes advantage of soft and hard attention by remaining computational efficient and differentiability within the attention windows.

2.4 Federated Learning

Federated learning (FL) [40, 41] was proposed to prevent data leakage and build machine learning models based on distributed datasets. The federated learning term was proposed by McMahan et al. [42] in 2016: "We term our approach Federated Learning since the learning task is solved by a loose federation of participating devices (which we refer to as clients) which are coordinated by a central server". The partitioning of non-IID (identically and independently distributed) data among different unreliable devices with limited communication bandwidth is represented as a new research task. The research across different areas, including cryptography, databases, and machine learning analyzed data distribution learning without sharing data. The approaches related to cryptographic methods were introduced in the early 1980s [43, 44]. Agrawal and Srikant [45] used the methods to learn from local data using a centralized approach. Several challenges that overlapped across different fields and not limited to machine learning, i.e., distributed optimization,

cryptography, security, differential privacy, fairness, systems, information theory, and statistics are then addressed.

The FL framework only shares model weights to the network nodes, and the model trains locally without sharing actual datasets [46]. The above works focused on handling issues related to data distribution, unbalanced data, and device ability for optimization. Federated learning can be classified as two learning models, i.e., horizontal and vertical [47]. In horizontal federated learning, feature space is the same, but data distribution is different [47]. The method has overlapping characteristic with privacy preservation machine learning as it considers the privacy of data in decentralized collaborative learning. In vertical federated learning, the dataset feature space is different, and data distribution is overlapping.

The federated learning mechanism is proposed by Shokri and Shmatikov [48] to train multiple deep learning models on joint inputs. Hayes and Ohrimenko proposed the trusted model sharing mechanism [49]. Fredrikson et al. proposed a federated model by utilizing the output of the machine learning algorithm [50]. Mohassel and Rindal [51] proposed an aggregation function that uses the approximation of fixed-point multiplication protocols. The federated learning method in cybersecurity has some drawbacks [52, 53]. These include single-point failure and calling issues for an increase in network size. The 5G network integration with FL and decentralized connectivity is still an open research question [54]. Cloud supported systems face many problems that include Quality of Service (QoS) issues for implementing time-sensitive applications [55, 56]. Introducing 5G network services, efficiency, and better communication of connected IoT devices are areas that should be resolved [55].

3 Developed Attention-Based Federated Learning Model

Given a dataset D shown in Table 1, and the set of FIs discovered from $D, \mathcal{F}(\mathcal{D}) = \{X_1, X_2, \ldots, X_F\}$, as mentioned in Table 2, the feature vector of a transaction T is defined as $f(T) = [x_1, x_2, \ldots, x_F]$, where $x_i = 1$ if $X_i \subseteq T$ otherwise $x_i = 0$, as describe in Table 3. The label of transactions is manually given.

Problem statement: Consider Table 1 as an example that will be used by the proposed algorithm shown in Algorithm 1. There are five transactions in the table $(T_1, T_2, T_3, T_4, T_5)$. Given items $\mathcal{I} = \{i_1, i_2, \ldots, i_M\}$ and a transaction dataset $\mathcal{D} = \{T_1, T_2, \ldots, T_N\}$. Our goal is to map the transaction data by learning a function $f : \mathcal{D} \rightarrow \mathbb{R}^d$ such that every transaction $T_i \in \mathcal{D}$ is mapped to a d -dimensional

Table 1 Transaction database

TID	Items	Class Label
T_1	a, b, c	1
T_2	b, c, d	1
T_3	a, d	2
T_4	b, c, d, e	3
T_5	a, c, d	4

Table 2 FIs from the transaction data under minimum support of 0.6

FI	Items	Sup
X_1	a	0.6
X_2	b	0.6
X_3	c	0.8
X_4	d	0.8
X_5	b, c	0.6
X_6	c, d	0.6

Table 3 Transaction represented as set of FIs and used as input features

TID	FIs	Class label
T_1	$\{X_1, X_2, X_3, X_5\}$,	1
T_2	$\{X_2, X_3, X_4, X_5, X_6\}$,	1
T_3	$\{X_1, X_4\}$,	2
T_4	$\{X_2, X_3, X_4, X_5, X_6\}$,	3
T_5	$\{X_1, X_3, X_4, X_6\}$,	4

continuous vector. The learning requires to have the similarity among the transactions in a way that correlated item sets should have similar embedding. The embedding $\mathbf{X} = \left[f(T_1), f(T_2), \ldots, f(T_N)\right]$ can be used by any machine learning classifier for the data mining as well as classification task by using the class label as mentioned in Tables 1 and 3.

3.1 Client-Server-Based Federated Learning

Federated learning in transactional data for multiple stores (clients) has not been well studied. The centralized model with an aggregate server for multiple clients without exchanging sensitive information is a preferred situation for data security. An attention network can adapt using the global weights and gradient from the client's locally trained models. This results in improved overall performance. The proposed framework uses multi-client data as shown in Fig. 1. For experimentation purposes, the data is split among six clients with equal distribution. Additionally, each client's data has non-overlapping data, local model, and database. The data distribution is varied among clients randomly. Given these differences, the framework is exercised on real-world datasets, where a store's multiple outlets in a given city are connected to a central server for supply and demand analysis. Our experimental settings also assume that that environment itself is non-independent and identically distributed. The federated learning algorithm uses a server-client setup. The client data is locally stored for each model. The local data is used to train the initialized model. For each iteration, we require the client to transfer the locally trained weights or their gradients to the server (Algorithm 1—lines 3 to 9). The server receives and aggregates the weights (Algorithm 1—lines 5 to 6). The purpose of this is to share the client model's training without sharing the actual data and using weights. The global model then uses the aggregate weights to update the

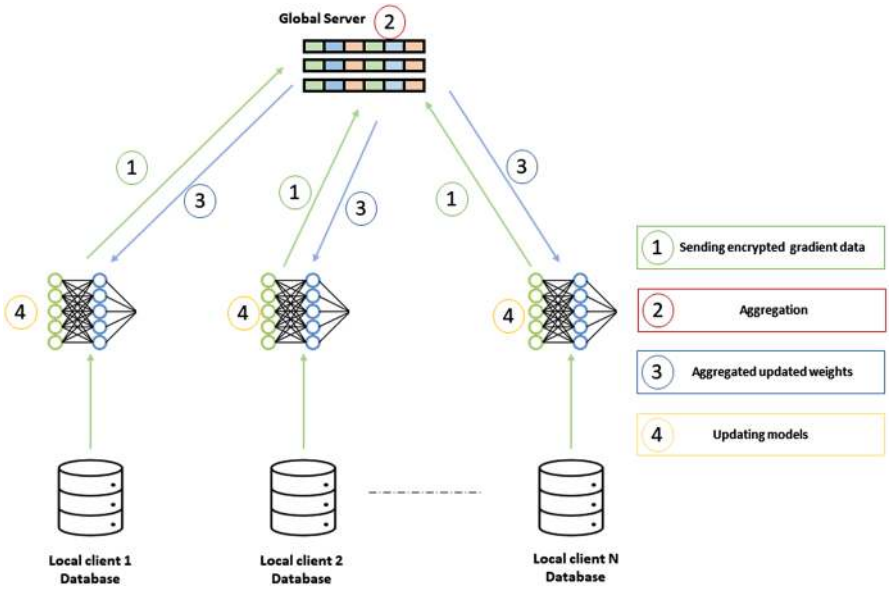

Fig. 1 The framework of the designed attention based federated learning for transaction embedding

global model (Algorithm 1—lines 8 to 9). After aggregation, the next round starts with each local client model possessing the global weights. After a certain number of federated learning iterations, a convergence point is reached by each client.

For experimental purposes, we used an early termination method to check the convergence point. During the empirical analysis, we set an early stopping patient value to 10. After that, the client may select the best model for each iteration based on hold out data. Each client monitors the validation loss on the local test set. From this, each client can select from the global aggregated model or the best local iteration model. We implemented the federated averaging method to converge early and reduce overfitting of the model [41]. In our empirical analysis, we found that embedding size must be set higher for optimal performance. The reason for this is that the decoder model can have a bigger vector space to map the positional attention. The federated averaging is used to reduce the global loss function L which is resulted from the weighted combination of K losses $\left\{\mathcal{L}_k\right\}_{k=1}^{K}$ (client losses) of the distributed aggregated function. The model could learn the embedding by using the parameter ϕ that minimizes the L on local data X_k where X combination of local data sets and representation of the embedding. Equations 1 and 2 represents the loss function.

$$\min_{\phi} \mathcal{L}(X;\phi) \tag{1}$$

$$\mathcal{L}(X;\phi) = \sum_{k=1}^{K} w_k \mathcal{L}_k\left(X_k;\phi\right) \tag{2}$$

The coefficient $w_K > 0$ denotes the weights of client K model. The model K is trained on local data, and only weights are distributed among the server. The weight is aggregated by summing the number of clients in the network.

Algorithm 1 Attention-based federated averaging method.

INPUT: T transaction data, R are the number of rounds, n_K are the local training epochs to minimize loss $\mathcal{L}_k\left(X_k; \phi^{(t-1)}\right)$ for client K

OUTPUT: Optimize weights

$Weights \leftarrow \phi^{random}$

▷ Initialize weights randomly

for *all* $r \in R$ **do**

  for *all client* $\in K$ **do**

    Send ϕ^{r-1}

    Receive $\left(\Delta\phi_k^{(r)}, n_k\right)$

  end for

  $\phi_k^{(r)} \leftarrow \phi^{(r-1)} + \Delta\phi_k^{(r)}$

  $\phi^{(r)} \leftarrow \frac{1}{\sum_k n_k} \sum_k \left(n_k \cdot \phi_k^{(r)}\right)$ ▷ Aggregate( )

end for

Return $\phi^{(r)}$

3.2 Deep Neural Network (DNN) Architecture

In our approach, we use transactional data represented as frequent itemsets as contextual information as described in Table 1 and Fig. 2. The frequency to purchase the likely items is always highly relevant. Therefore, the item frequency in transactional data is considered the developed model's contextual data points.

3.2.1 Attention Network

We use the attention mechanism to exploit the contextual relevance of similar items to create embedding and then used it to predict output class describe in Table 4. For comparison, we used Table 5 for configuration. We first transfer the items' features through dense 100 Relu units. We then pass the developed model through the Luong's attention method that uses the decoder hidden state [57].

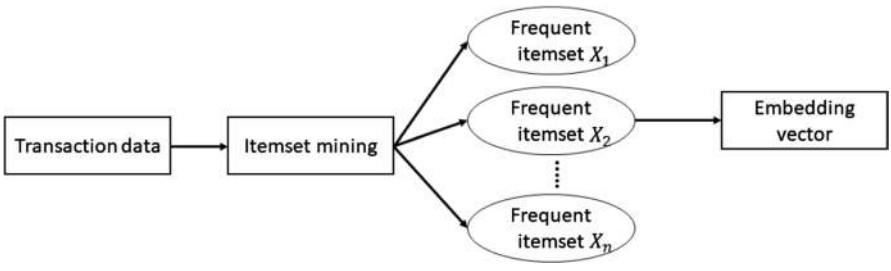

Fig. 2 Flow of the method

Table 4 Transaction data set input and output parameters

Dataset	Input	Items	Output	No. of classes	Type
Snippet	FIs	Set keywords e.g.,	Business, Computers, Culture-Arts-Ent,	8	Multi-class
		"supplier", "export"	Education-Science,Engineering, Health, Politics-Society, Sports		
Cancer		Diagnosedsymptoms	Re-admission status	3	Multi-class
		(e.g., "cough", "headache")	of apatient		
Retail		Set of products	United kingdom	2	Binary
		purchased by customers	orothers		
Food		List of foods (e.g., "milk")	Using coupon	2	Binary
		purchased	or not		

Table 5 Characteristic of the data set

Dataset	# trans	# train	# test	# items	Avg. length
Snippets	12,340	10,060	2280	23,686	13.00
Cancer	15,000	12,000	3000	3234	6.00
Retail	3000	2400	600	3376	26.93
Food	4000	3200	800	1559	25.87

The attention score is calculated and concatenated with the hidden state of the decoder. We then pass the obtained output sequence through 100 and 350 Relu units, respectively. The last layer is `softmax`, and we used an `Adam` optimizer for the learning rate. We mask the input data for padding and train the network for 10 epochs per iteration. The number of iterations is set to 100. The input features are padded, and the fixed-length vector is passed to the models.

4 Experimental Evaluation

For experimental analysis, we used four benchmark datasets from the SPMF library [58]. We mention the characteristic summary of the dataset in Tables 4 and 5. Snippets dataset is from the web search transaction, and each item represents the keywords. Cancer is the dataset of patient admission based on diagnoses symptoms. The Retail is a transaction data set from the United kingdom online retailer. A food dataset is a collection of food baskets where each item represents a product purchased by the customer.

4.1 Baseline

In the Bag-of-Words (BOW) and Term Frequency-Inverse Document Frequency (TF-IDF) methods, each transaction is treated as a document and items as a word [13]. By using this mechanism, we can apply NLP based models to transactional data. For comparative purposes, we used the BOW model, TF-IDF, and the Trans-2vec model [24] as the baseline models to compare with our proposed model. Moreover, we used the elbow method for the selection of the support threshold of the frequent itemsets as was shown by Nguyen et al. [24] and shown in Fig. 3. The vector representation of attention embedding was used for the classification task, as shown in Table 4. This is done using the `softmax` layer. The loss function is used to evaluate the classification task. The model is measured in terms of Accuracy and F-measure.

4.2 Results and Discussion

In Table 6, the proposed attention-based embedding model is shown to be able to perform better than the embedding-based model. The proposed model can achieve 2−13% improvement over BOW, and TF-IDF, respectively. Against word embedding, the developed model can perform 0.2% to 1% improvement for Snippets, Cancer, and Food datasets respectively that we can observe in Table 6. However, they have reduced performance for the retail dataset. The reason for this is that the developed model requires tuning for different dataset characteristics. Another potential reason is that it contains an imbalanced class problem that affects the classification accuracy. The model can improve by undersampling the majority class. However, we do not need to hyper-tune the classification layer with embeddings. The purpose is to see a federated learning environment for the transaction data. However, the model can still be improved by tuning the hyperparameters. The federated environment helps to learn the embedding without sharing actual data and the results demonstrate

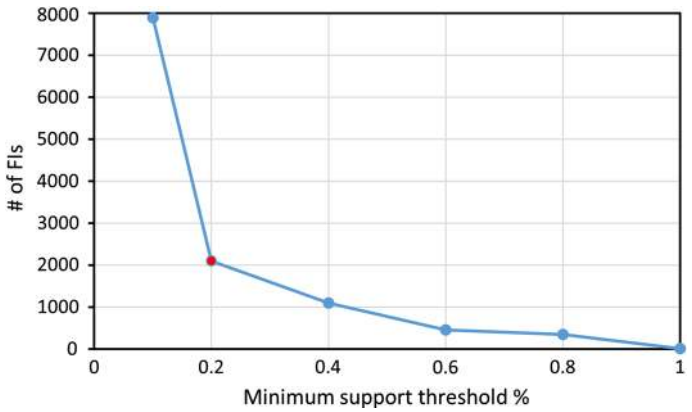

Fig. 3 The threshold selection based on the elbow curved method. The ϕ value selected for *Snippet data* is indicated by the red dot (Color figure online)

Table 6 Results in terms of the pre-defined support thresholds manually

Methods	Snippets		Cancer		Retail		Food	
	AC	F1	AC	F1	AC	F1	AC	F1
BOW	66.32	65.83	48.57	48.47	77.33	77.31	63.12	63.12
TF-IDF	70.26	69.52	49.43	49.43	81.67	81.56	64.12	64.49
Trans2Vec	79.05	78.3	50.34	50.28	**83.43**	**83.36**	72.51	72.47
Proposed	**81.02**	**80.01**	**57.03**	**57.01**	83.01	82.08	**72.55**	**72.48**
Support thresold	0.2		0.2		0.7		0.2	

The bold values show the best performance among all compared algorithms

that the learning of transaction information for frequent itemsets and then applying them on the classification can be enhanced with some hyper tuning. The attention network can extract meaningful relationships among the transaction data.

To obtain the robustness results, the proposed model needs to be run for more extended epochs. This adjustment helps to increase the weights per instance. However, we can also see that the correlated features set decreases the performance while training time increased. All methods are performed well for the classification task by using the transaction as a set of frequent itemsets. We also see that we should increase the feature set for our method. Other features, i.e., support, weight, or occupancy of itemset should be combined to get a suitable learning method, which can be explored in the future.

When invoking local and global updates, loss monitoring is vital. If we perform more local updates, then it will increase the divergence between local models as was seen with the averaged function. This will increase the convergence in terms of training loss in comparison to federated cycles. The increase of local updates helps to reduce communication costs and time per iteration. The optimization of the global and local loss depends upon the federated cycle, local updates, and error versus convergence time. This research should be evaluated further, and the optimization method should be introduced.

5 Conclusion and Future Work

With 5th Generation (5G) technology, data-intensive applications will be created using distributed sensors on various hardware levels. However, security considerations create privacy challenges. In this paper, we propose data privacy of a 5G device connected to a heterogeneous network. Our method aims to analyze sensitive information in datasets collected with smart objects in IoT environments. From the data, we wish to identify sensitive information via a federated learning approach. This paper uses an attention-based mechanism for transaction embedding in an attempt to classify transaction data. The developed model can achieve high accuracy. The federated averaging learning-based method significantly reduces the global loss of the shared weights. From our experimental results, it can be seen that the

federated learning approach can obtain practical benefits over traditional supervised learning methods. Our model can perform better locally without sharing and distributing private raw data across a network. This technique can achieve high generalization. In the future, we plan to optimize the network-tuning using an active learning mechanism. A weighted-based method for each class sub-sample selection can also be considered as a further extension. Nevertheless, a multi-label classification that considers uncertainty, utility, frequency, and co-occurrence as input features for the classification task can be investigated to improve the overall accuracy performance.

Funding Open access funding provided by Western Norway University Of Applied Sciences.

References

1. Fournier-Viger, P., Lin, J.C.W., Vo, B., Chi, T.T., Zhang, J., Le, H.B.: A survey of itemset mining. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. **7**(4), 1207 (2017)
2. Han, J., Pei, J., Yin, Y., Mao, R.: Mining frequent patterns without candidate generation: a frequent-pattern tree approach. Data Min. Knowl. Discov. **8**(1), 53–87 (2004)
3. Agrawal, R., Srikant, R., et al.: Fast algorithms for mining association rules. The International Conference on Very Large Data Bases, vol. 1215, pp. 487–499 (1994)
4. Mannila, H., Toivonen, H., Verkamo, A.I.: Discovery of frequent episodes in event sequences. Data Min. Knowl. Discov. **3**(1), 259–289 (1997)
5. Fournier-Viger, P., Lin, J.C.W., Kiran, R.U., Koh, Y.S., Thomas, R.: A survey of sequential pattern mining. Data Sci. Pattern Recogn. **1**, 54–77 (2017)
6. Shi, J., Wan, J., Yan, H., Suo, H.: A survey of cyber-physical systems. In: International Conference on Wireless Communications and Signal Processing, pp. 1–6 (2011)
7. Al Ridhawi, I., Aloqaily, M., Boukerche, A., Jararweh, Y.: Enabling intelligent IOCV services at the edge for 5G networks and beyond. IEEE Trans. Intell. Transp. Syst. **1–11**, (2021)
8. Lin, J.C.W., Srivastava, G., Zhang, Y., Djenouri, Y., Aloqaily, M.: Privacy preserving multi-objective sanitization model in 6G IoT environments. IEEE Internet of Things J. **8**(7), 5340–5349 (2021)
9. Ehsanfar, A., Grogan, P.T.: Auction-based algorithms for routing and task scheduling in federated networks. J. Netw. Syst. Manag. **28**(2), 271–297 (2020)
10. Ehsanfar, A., Grogan, P.T.: Mechanism design for exchanging resources in federated networks. J. Netw. Syst. Manag. **28**(1), 108–132 (2020)
11. Liaw, A., Wiener, M., et al.: Classification and regression by randomforest. R news **2**(3), 18–22 (2002)
12. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. Wiley, New York (1973)
13. Trstenjak, B., Mikac, S., Donko, D.: KNN with TF-IDF based framework for text categorization. Procedia Eng. **69**, 1356–1364 (2014)
14. Quinlan, J.R.: Induction of decision trees. Mach. Learn. **1**, 81–106 (1986)
15. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo (1994)
16. Cheng, H., Yan, X., Han, J., Hsu, C.W.: Discriminative frequent pattern analysis for effective classification. In: The International Conference on Data Engineering, pp. 716–725 (2007)

17. McMahan, B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: The International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 54, pp. 1273–1282 (2017)
18. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. In: ACM SIGMOD International Conference on Management of Data, pp. 207–216 (1993)
19. Savasere, A., Omiecinski, E.R., Navathe, S.B.: An efficient algorithm for mining association rules in large databases. Technical report, Georgia Institute of Technology (1995)
20. De Smedt, J., Deeva, G., De Weerdt, J.: Mining behavioral sequence constraints for classification. IEEE Trans. Knowl. Data Eng. **32**(6), 1130–1142 (2020)
21. He, Z., Gu, F., Zhao, C., Liu, X., Wu, J., Wang, W.: Conditional discriminative pattern mining: concepts and algorithms. Inf. Sci. **375**, 1–15 (2017)
22. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems (2013)
23. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 855–864 (2016)
24. Nguyen, D., Nguyen, T.D., Luo, W., Venkatesh, S.: Trans2Vec: learning transaction embedding via items and frequent itemsets. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining, vol. 10939, 361–372 (2018)
25. Nguyen, G., Dlugolinsky, S., Bobák, M., Tran, V.D., López, Á.L., Heredia, I., Malík, P., Hluchý, L.: Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey. Artif. Intell. Rev. **52**(1), 77–124 (2019)
26. Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Empirical Methods in Natural Language Processing, pp. 1724–1734 (2014)
27. Vinayakumar, R., Soman, K.P., Poornachandran, P.: Applying convolutional neural network for network intrusion detection. In: International Conference on Advances in Computing, Communications and Informatics, pp. 1222–1228 (2017)
28. Sze, V., Chen, Y.H., Yang, T.J., Emer, J.S.: Efficient processing of deep neural networks: a tutorial and survey. Proc. IEEE **105**(12), 2295–2329 (2017)
29. Wainberg, M., Merico, D., Delong, A., Frey, B.J.: Deep learning in biomedicine. Nat. Biotechnol. **36**(9), 829–838 (2018)
30. Van Horn, G., Aodha, O.M., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.J.: The inaturalist species classification and detection dataset. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8769–8778 (2018)
31. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D.Y., Bagul, A., Langlotz, C., Shpanskaya, K.S., Lungren, M.P., Ng, A.Y.: Chexnet: radiologist-level pneumonia detection on chest X-rays with deep learning. CoRR (2017). arXiv:1711.05225
32. Siam, M., Elkerdawy, S., Jägersand, M., YogamaniM S.K.: Deep semantic segmentation for automated driving: taxonomy, roadmap and challenges. In: IEEE International Conference on Intelligent Transportation Systems, pp. 1–8 (2017)
33. Lin, J.C.W., Shao, Y., Djenouri, Y., Yun, U.: ASRNN: a recurrent neural network with an attention model for sequence labeling. Knowl. Based Syst. **212**, 106548 (2021)
34. Sennrich, R., Haddow, B., Birch, A.: Neural machine translation of rare words with subword units (2015). arXiv:1508.07909
35. Fawaz, H.I.: Deep learning for time series classification. CoRR (2020). arXiv:2010.00567
36. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: Advances in Neural Information Processing Systems, pp. 289–297 (2016)
37. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: The International Conference on Learning Representations (2015)
38. Kelvin, X., Ba, J., Kiros, R., Cho, K., Courville, A.C., Salakhutdinov, R., Zemel, R.S., Bengio, Y.: Show, attend and tell: neural image caption generation with cisual attention. In: The International Conference on Machine Learning, vol. 37, pp. 2048–2057 (2015)
39. Luong, T., Pham, H., Manning, C.D..: Effective approaches to attention-based neural machine translation. In: The Conference on Empirical Methods in Natural Language Processing, pp. 1412–1421 (2015)
40. Posner, J., Tseng, L., Aloqaily, M., Jararweh, Y.: Federated learning in vehicular networks: opportunities and solutions. IEEE Netw. **35**(2), 152–159 (2021)

41. Konecný, J., McMahan, H.B., Ramage, D., Richtárik, P.L.: Federated optimization: distributed machine learning for on-device intelligence. CoRR (2016). arXiv:1610.02527
42. McMahan, B., Moore, E., Ramage, D., Hampson, S., Aguera y Arcas, B.: Communication-efficient learning of deep networks from decentralized data. In: The International Conference on Artificial Intelligence and Statistics, vol. 54, pp. 1273–1282 (2017)
43. Rivest, R.L., Adleman, L., Dertouzos, M.L., et al.: On data banks and privacy homomorphisms. Found. Secure Comput. **4**(11), 169–180 (1978)
44. Yao, A.C.: Protocols for secure computations. In: The Annual Symposium on Foundations of Computer Science, pp. 160–164 (1982)
45. Agrawal, R., Srikant, R.: Privacy-preserving data mining. In: ACM SIGMOD International Conference on Management of Data, pp. 439–450 (2000)
46. Mothukuri, V., Parizi, R.M., Pouriyeh, S., Huang, Y., Dehghantanha, A., Srivastava, G.: A survey on security and privacy of federated learning. Future Gen. Comput. Syst. **115**, 619–640 (2020)
47. Yang, Q., Liu, Y., Chen, T., Tong, Y.: Federated machine learning: concept and applications. ACM Trans. Intell. Syst. Technol. **10**(2):12:1–12:19 (2019)
48. Shokri, R., Shmatikov, V.: Privacy-preserving deep learning. In: The Conference on Computer and Communications Security, pp. 909–910 (2015)
49. Hayes, J., Ohrimenko, O.: Contamination attacks and mitigation in multi-party machine learning. CoRR (2019). arXiv:1901.02402
50. Fredrikson, M., Jha, S., Ristenpart, T.: Model inversion attacks that exploit confidence information and basic countermeasures. In: ACM SIGSAC Conference on Computer and Communications Security, pp. 1322–1333 (2015)
51. Mohassel, P., Rindal, P.L.: Aby^3: a mixed protocol framework for machine learning. In: ACM SIGSAC Conference on Computer and Communications Security, pp. 35–52 (2018)
52. Sedjelmaci, H., Guenab, F., Senouci, S., Moustafa, H., Liu, J., Han, S.: Cyber security based on artificial intelligence for cyber-physical systems. IEEE Netw. **34**(3), 6–7 (2020)
53. Liu, Y., Kang Peng, J., A.M. Iliyasu, Niyato, D., El-Latif, A.A.A.: A secure federated learning framework for 5G networks. IEEE Wirel. Commun. **27**(4), 24–31 (2020)
54. Savazzi, S., Nicoli, M., Rampa, V.: Federated learning with cooperating devices: a consensus approach for massive iot networks. IEEE Internet of Things J. **7**(5), 4641–4654 (2020)
55. Singh, K.D., Sood, S.K.: QoS-aware optical fog-assisted cyber-physical system in the 5g ready heterogeneous network. Wirel. Pers. Commun. **116**(4), 3331–3350 (2020)
56. Singh, K.D., Sood, S.K.: 5G ready optical fog-assisted cyber-physical system for IoT applications. IET Cyber-Phys. Syst. Theory Appl. **5**(2), 137–144 (2020)
57. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. arXiv preprint (2015). arXiv:1508.04025
58. Fournier-Viger, P., Lin, J. C.-W., Gomariz, A., Gueniche, T., Soltani, A., Deng, Z., Lam, H.T.: The spmf open-source data mining library version 2. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 36–40 (2016)

Usman Ahmed is a PhD candidate at the Western Norway University of Applied Sciences (HVL). He has rich experience building and scaling high-performance systems based on data mining, natural language processing, and machine learning. His research interests are sequential data mining, heterogeneous computing, recommendation systems, and machine learning.

Gautam Srivastava is an associate professor at Brandon University, Canada. He has published over 200 peer-reviewed papers. His research interested include privacy, security, blockchain technology, ML/AI, and the Internet of Things. He is a senior member of IEEE.

Jerry Chun-Wei Lin is a full Professor in Western Norway University of Applied Sciences. He has published more than 400 peer-review papers. His research interests include data analytics, privacy preserving and security, ML/AI, and optimization. He is a fellow of IET and the senior member of IEEE and ACM.