# Federated K-means Clustering for Adaptive OFDM-IM

Xueyu Wu, Andy M. Tyrrell, and Youngwook Ko

***Abstract*—In this letter, a new federated learning strategy for k-means clustering assisted adaptive orthogonal frequency devision multiplexing with index modulation (OFDM-IM) system, which develops a global model using local learning outcomes aggregated from distributed devices, is proposed. The proposed strategy aims to efficiently leverage the computing power of all the devices in a distributed system. Simulation results show that the proposed strategy reduces the training steps required at each device and simultaneously improve the throughput, with a federation cost for model aggregation and broadcast.**

***Index Terms*—Federated learning, k-means clustering, adaptive modultion, OFDM-IM**

## I. Introduction

Index modulation has been considered as an energy efficient method for the mMTC network in 6G. In conventional modulation schemes, all data bits are conveyed by the amplitude-phase modulation (APM) symbols, whereas in index modulation, part of data bits are conveyed by the indices of active subcarriers or antenna. Exploiting the index modulation concept in the multi-carrier systems, various OFDM-IM schemes that have been proposed. [1] proposed a spread-OFDM-IM scheme achieving high transmit diversity, which applies a precoding matrix to the transmit signal. [2] proposed a scheme called coordinate interleaved OFDM-IM (CI-OFDM-IM) which seperately transmit the real and imaginary parts of the data symbol using different active subcarriers. [3] proposed a super-mode OFDM-IM (SuM-OFDM-IM) which forms index symbol via both mode activation patterns (MAPs) and subcarrier activation patterns (SAPs) at the same time to maximize the number of data bits transmitted by index symbol. Mainly focusing on coding gain and diversity gain, such existing OFDM-IM schemes have neglected to discuss new insights into learning-driven adaptive modulation signals, which could improve the spectral efficiency and reliability in multi-user heterogeneous environments.

Conventional adaptive modulation schemes require channel state information (CSI) at the transmitter while learning driven schemes can learn the pattern from the environment. In time division duplexing (TDD) applications, the CSI of downlink and uplink are the same so that downlink signal can be adopted as the indicator of CSI for uplink. K-means clustering can efficiently extract the implicit pattern of multi-dimensional vectors by clustering them according to the Euclidean distance between them. [4] proposed an OFDM-IM adaptation with the use of single user k-means clustering. However, the process of constituting a sufficient size of training dataset is expensive for resource limited MTC devices. Centralized learning is a possible solution leveraging the resources of central servers but it leads to high communication cost for sharing the data.

Federated learning (FL) is a potential enabler for connected intelligence in 6G [5]. Compared to centralized learning strategy, FL only shares the model updates instead of all the data, which reduces the consumption of spectral and energy resources. For example, [6] proposed a federated deep learning strategy for automatic modulation classification (AMC), which avoids data leakage while the performance loss is within 2% compared to the centralized algorithm.

The model aggregation strategy is a crucial challenge in federated learning. In FedSGD, a part of clients upload samples randomly selected from their local dataset and the model parameters will be updated in the FL server, which causes expensive cost in terms of communications. To optimize this weakness, an algorithm named FedAvg was developed [7]. In FedAvg, clients update their local parameters locally, and only upload the parameters to the FL server. The FL server will average the local parameters with appropriate weights to get global parameters. In [6], FedSGD and FedAvg based algorithms are proved to have similar performance when solving AMC problem. [8] proposed a federated stochastic variance reduced gradient (FSVRG) optimization algorithm, which improved the performance for non-independent, and identical data distribution. Such strategies focus on using deep learning, which may not suit to constrained MTC devices. Only a few paper developed federation strategy for k-means clustering. [9] proposed a federated k-means clustering algorithm based on FedAvg for image recognition. [10] proposed a federated k-means scheme for proactive caching, where the training data are shared for model aggregation at the high cost. Potential of effectively federating k-means clustering has been overlooked in OFDM-IM variants at MTC devices.

In this paper, a federated k-means clustering called Fed-k-means is developed to obtain the precise adaptive OFDM-IM model enhancing the system throughput with less training data at devices. The main contributions of this work are: (i) to develop the multi-user adaptive OFDM-IM system with the use of federated k-means clustering; (ii) to develop a novel weighting strategy for the federated k-means clustering to provide the accurate adaptation strategy of OFDM-IM signals; (iii) to evaluate the effective throughput of the system by simulations and the simulated results clearly present that the proposed system can outperform the benchmarks, in terms of the effective throughput.

Xueyu Wu, Andy M. Tyrrell and Youngwook Ko are with the School of Physics, Engineering and Technology, University of York, United Kingdom (Emails: xueyu.wu@york.ac.uk, andy.tyrrell@york.ac.uk and youngwook.ko@york.ac.uk).

## II. System Model

Consider a distributed multi-carrier system where $B$ users send uplink data packets to the base station (BS) by employing OFDM-IM in TDD mode. For the OFDM-IM transmission, assume that the sub-block size for each user is $N$. Since the communication and learning process run independently across users, for simplify, the following discussion will focus on one user. The communication process are introduced here with the learning process in Sec. III.

Every transmission each user intends to adjust its modulation mode. The $KT$ modulation modes are combinations of $T$ modulation orders and $K$ different numbers of active subcarriers. Denote the $t$-th modulation order and the $k$-th number of active subcarriers by $M_t \in \{M_1, \cdots, M_T\}$ and $k \in \{1, \cdots, K\}$, respectively. $\mathbf{m}_q \in \{\mathbf{m}_0, \cdots, \mathbf{m}_{KT}\}$ is the $q$-th modulation mode, where $\mathbf{m_0}$ means no transmission and

$$\mathbf{m}_q \triangleq \begin{cases} (0,0) & \text{if } q = 0 \\ (k, M_t) & \text{if } q \in \{1, \cdots, KT\} \end{cases} \tag{1}$$

Given $\mathbf{m}_q$, there are $p$ bits to be transmitted. Denote $C(N,k)$ the number of possible combinations of $k$ active subcarriers over $N$ subcarriers, $p$ is given by

$$\begin{aligned} p &= p_1 + p_2 \\ &= \lfloor \log_2 C(N,k) \rfloor + k \log_2 M_t \end{aligned} \tag{2}$$

$\mathbf{s} = [s_1, ..., s_k]^T$ is the modulated symbol vector, $s_k \in \mathbf{M}_t$, where $\mathbf{M}_t$ is the constellation of $M_t$-ary modulation. Notice that for a given $k$, the index subset is $\mathbf{i} = \{i_1, \cdots, i_k\}$ where $i_a \in \{1, \cdots, N\}$ for $a = 1, \cdots, k$. The OFDM-IM signal vector of each user in the frequency domain is $\mathbf{x} = [x_1, ..., x_N]^T$, where

$$x_n = \begin{cases} s_m & \text{if } n = i_m \in \mathbf{i} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Denote by $\mathbf{H} = \text{diag}\{h_1, ..., h_N\}$, the frequency domain channel matrix whose elements $h_i$ are complex Gaussian random variables with $h_i \sim \mathcal{CN}(0,1)$. Applying $\mathbf{G}$ the $N \times N$ Zadoff-Chu precoding matrix by [1], in the frequency domain, the uplink received signal is

$$\mathbf{y} = \sqrt{P\bar{\alpha}}\mathbf{H}\mathbf{G}\mathbf{x} + \mathbf{n} \tag{4}$$

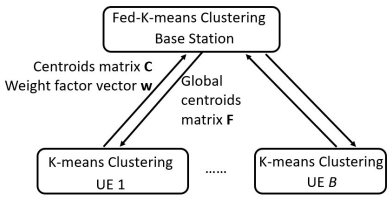
where $\mathbf{n}$ is the Additive White Gaussian Noise (AWGN) vector whose elements follow $\mathcal{CN}(0,1)$, and $\bar{\alpha}$ is the average SNR. At the BS side, maximum likelihood (ML) detector is adopted to recover the signal of each user.

$\omega(\mathbf{x}, \hat{\mathbf{x}})$ is the number of error bits when $\mathbf{x}$ is decoded as $\hat{\mathbf{x}}$. The upper bound of conditional bit error probability with $\mathbf{m}_q$ and channel matrix $\mathbf{H}$ is given by

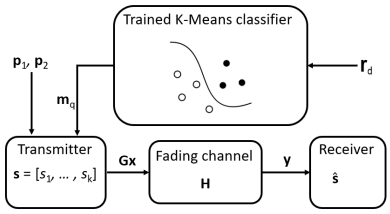
$$P_b(\mathbf{m}_q, \mathbf{H}) \leq \frac{1}{pCM_t^k} \sum_{\mathbf{x}} \sum_{\hat{\mathbf{x}}} \omega(\mathbf{x}, \hat{\mathbf{x}}) Q\left(\sqrt{\frac{\sqrt{P\bar{\alpha}}\|\mathbf{H}\mathbf{G}(\mathbf{x} - \hat{\mathbf{x}})\|^2}{2N_0}}\right) \tag{5}$$

$Z(\mathbf{m}_q)$, the indicator on whether modulation mode $\mathbf{m}_q$ is chosen in $i$-th transmission, is given by

$$Z(\mathbf{m}_q) \triangleq \begin{cases} 1 & \text{if } \mathbf{m}_q \text{ is chosen} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$




(a) Federated learning structure




(b) Learning-driven adaptive OFDM-IM framework

Fig. 1: The structure of federated clustering adaptive OFDM-IM

The lower bound of instantaneous effective throughput when using modulation mode $\mathbf{m}_q$ is given by

$$T_E(\mathbf{m}_q, \mathbf{H}) \geq p(1 - P_b(\mathbf{m}_q, \mathbf{H})) \tag{7}$$

The effective throughput is measured by

$$\begin{aligned} T_{Eavg} &= \mathbb{E}[\sum_{q=0}^{KT} T_E(\mathbf{m}_q, \mathbf{H}) Z(\mathbf{m}_q)] \\ &= \sum_{q=0}^{KT} (\mathbb{E}[T_E(\mathbf{H}|\mathbf{m}_q)] \Pr(\mathbf{m} = \mathbf{m}_q|\mathbf{F})) \end{aligned} \tag{8}$$

where $\Pr(\mathbf{m} = \mathbf{m}_q|\mathbf{F})$ is the probability of chosing modulation mode $\mathbf{m}_q$ with a given learned model $\mathbf{F}$. In the simulations presented in this paper, the effective throughput is measured by the average number of correctly decoded bits per transmission.

It can be seen that the $T_{Eavg}$ is effected by the learned-model $\mathbf{F}$. The main focus of this paper is to develop a federated learning assisted adaptive OFDM-IM algorithm to find the best modulation mode, enhancing the throughput without CSI.

## III. Federated Clustering Adaptive OFDM-IM

This section introduces the learning process of the proposed system. Figure 1a shows the structure of federated learning. In this phase, the BS uses the local models collected from $B$ users to develop a global model and send it to $B$ users. Figure 1b illustrates the structure of the learning-driven adaptive OFDM-IM, where $\mathbf{r}_d$ is the vector of downlink signal energy which is acquired by observing the downlink signal of each subcarrier at the user. In this stage, each user uses $\mathbf{r}_d$ as the input of the learned model to predict the modulation mode maximizing the effective throughput.

Consider that each user has $V$ different data vectors to train itself over the learning process. The $v$-th training data vector contains $(|h_{1v}|^2, ..., |h_{Nv}|^2, q)$, $v \in \{1, \cdots, V\}$, where the first $N$ entries, $|h_{iv}|^2$, represent uplink subcarrier gains. The $(N+1)$-th element is the best modulation mode, which

achieves maximum effective throughput for the uplink channel $\mathbf{H}$ over the learning process. The best modulation mode estimation $q$ for a given channel matrix $\mathbf{H}$ is obtained by

$$q = \arg\max_{j} T_E(\mathbf{m}_j, \mathbf{H}) \quad s.t. \;\; P_e(\mathbf{m}_j, \mathbf{H}) \leq \mu \tag{9}$$

where $\mu$ is the BEP threshold.

To this end, denote, first, $b$-th user exploits its local dataset obtained from experienment in advance to get local centroids matrix $\mathbf{C}^b$ by using Algorithm 1. The superscripts represent user index here and after.

**Algorithm 1** Fed-k-means local model training at the $b$-th user

Denote $\mathbf{Z}^b$ the matrix whose columns containing the training data, $\mathcal{S}_k^b$ the index set of the training data points assigned to the $k$-th cluster, and $|\mathcal{S}_k^b|$ the number of elements in $\mathcal{S}_k^b$
**Input(s):** $(N+1) \times V$ training data matrix $\mathbf{Z}^b$, number of local clusters $K_C$
**Output(s):** $(N+1) \times K_C$ local centroids matrix $\mathbf{C}^b$, $1 \times K_C$ weights vector $\mathbf{w}^b$
**Initialization:** Randomly generate initial local centroids matrix $\mathbf{C}^b$ and then assign each data point, $\mathbf{Z}_{1:N,v}^b$, $v \in [1, V]$, to its closest local cluster
**Repeat until $\mathbf{C}^b$ does not change:**
(1) Update $\mathbf{C}^b$
**for** $k = 1, ...K_C$ **do**
$\mathbf{C}_{1:N,k}^b \leftarrow \frac{1}{|\mathcal{S}_k^b|} \sum_{v \in \mathcal{S}_k^b} \mathbf{Z}_{1:N,v}^b$
Set $\mathbf{C}_{N+1,k}^b$ to the modulation mode which appears most frequserntly in the $k$-th cluster
**end for**
(2) Assign each data point, $\mathbf{Z}_{1:N,v}^b$, $v \in [1, V]$, to its closest cluster and then calculate the weight for each centroid
**for** $k = 1, ...K_C$ **do**
$\mathbf{w}_k^b \leftarrow |\mathcal{S}_k^b|$
**end for**

The BS collects local centroids from the users after the local training process, and performs another clustering algorithm with the local centroids to develop a global model. In particular, computing the federated centroids in the global model, the sum of weighted local centroids are iteratively considered as shown in Algorithm 2. The weight coefficients of local centroids represent the number of data points in the local clusters, which can be used to indicate the relative importance of local centroids with large data points against those with small data points. Based on these, the global model is computed. The global model accuracy increases when the number of users increases by implicitly leveraging more training data.

Such federated clustering is designed to decrease the loss function, which is given by

$$\mathcal{L}(\mathbf{F}) = \sum_{k} \sum_{l \in \mathcal{A}_k} \mathbf{w}_l \|\mathbf{F}_{1:N,k} - \mathbf{C}_{1:N,l}\|_2^2 \tag{10}$$

Once the global model is updated, each user is assumed to access the global model and predict their best modulation mode to be used at each adaptive transmission. During the process of predicting the modulation mode maximizing the effective throughput, the downlink signal energy vector $\mathbf{r}_d$ is adopted as the input of the online prediction model. The prediction scheme is shown in Algorithm 3.

**Algorithm 2** Fed-k-means global model updating

Denote $\mathcal{A}_k$ the index set of the local centroids assigned to the $k$-th global cluster
**Input(s):** $\mathbf{C} = [\mathbf{C}^1, \cdots, \mathbf{C}^B]$, $\mathbf{w} = [\mathbf{w}^1, \cdots, \mathbf{w}^B]$, number of global clusters $K_G$
**Output(s):** $(N+1) \times K_G$ global centroids matrix $\mathbf{F}$
**Initialization:** Randomly generate initial global centroids matrix $\mathbf{F}$ and then assign each data point, $\mathbf{C}_{1:N,l}$, $l \in [1, L]$, $L = BK_C$, to its closest global cluster
**Repeat until $\mathbf{F}$ does not change:**
(1) Update $\mathbf{F}$
**for** $k = 1, ...K_G$ **do**
$\mathbf{F}_{1:N,k} \leftarrow \frac{1}{\sum_{l \in \mathcal{A}_k} \mathbf{w}_l} \sum_{l \in \mathcal{A}_k} \mathbf{w}_l \mathbf{C}_{1:N,l}$
Set $\mathbf{F}_{N+1,k}$ to the modulation mode which appears most frequserntly in the $k$-th global cluster
**end for**
(2) Assign each local centroid, $\mathbf{C}_{1:N,l}$, $l \in [1, L]$, to its closest global cluster

**Algorithm 3** Prediction algorithm

**Input(s):** Global centroids matrix $\mathbf{F}$, downlink signal energy vector $\mathbf{r}_d$
**Output(s):** Index of modulation mode $q$
**for** every transmission **do**
Predict $q$ (assign $\mathbf{r}_d$ to its closest global centroid)
Find $l = \arg\min_k \|\mathbf{F}_{1:N,k} - \mathbf{r}_d\|_2^2$
$q \leftarrow \mathbf{F}_{N+1,l}$
**end for**

## IV. Simulation Results

Simulation results of the proposed algorithms in the distributed adaptive OFDM-IM systems are presented in this section. To measure their efficacy, the focus is on two simulation scenarios: (i) the effective throughtput and BER performance in different average SNRs of the Fed-k-means and the single user k-means strategy; and (ii) the sensitivity of the federated OFDM-IM adaptation with different number of users in terms of effective throughput. For all the simulations, Rayleigh fading channel is applied to each subcarrier, the number of subcarriers $N = 4$, the number of active subcarriers $k \in \{1, 2\}$, the cardinality of possible modulation constellations $M_t \in \{0, 2, 4\}$. The BEP threshold $\mu = 0.01$. By using the well known elbow method, the number of clusters of the local model, $K_C$, are found to be 10, 20, 40, 100 for the training dataset sizes of 50, 100, 200, 500, respectively, and the number of clusters of the global model, $K_G$, is chosen to 100.

In Figure 2 and Figure 3, the effective throughput and BER of the four schemes, (i) Classical k-means with 200 training data; (ii) Fed-k-means with 200 training data at each user;

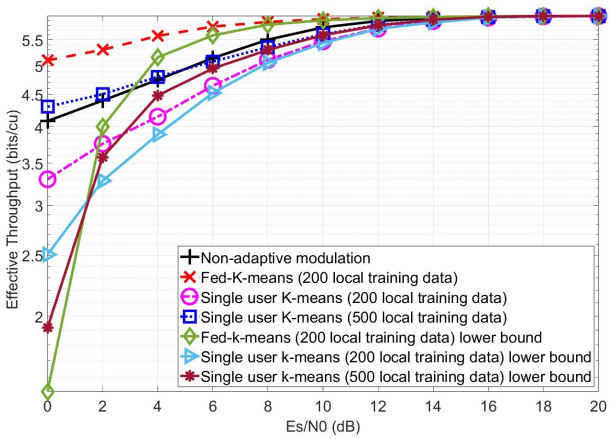



Fig. 2: Effective throughput versus average SNR of Fed-k-means adaptive OFDM-IM with 70 users.

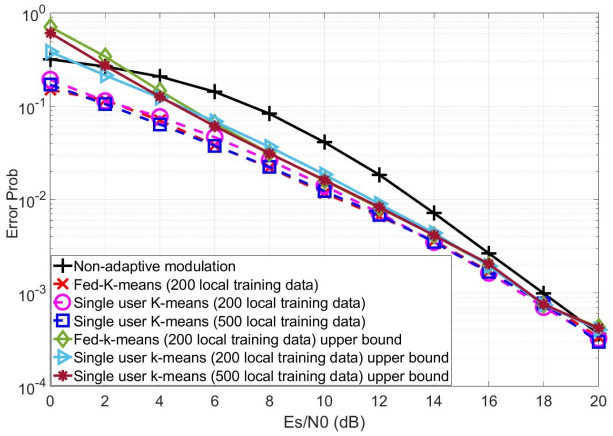



Fig. 3: Average BER versus average SNR of Fed-k-means adaptive OFDM-IM with 70 users.

(iii) Classical k-means with 500 training data; and (iv) Non-adaptive modulation, are presented in a 70 users scenario. The theoretical lower bounds of effective throughput and the theoretical upper bounds of BER, with a given learned model, are also depicted for validation. At mid and low SNRs, the Fed-k-means has higher effective throughput than the single user k-means with either 200 training data or 500 training data. The BER of the Fed-k-means and single user k-means are similar, which are lower than the non-adaptive modulation. When the SNR is greater than 14dB, the effective throughput of all the four schemes are similar because the mode with highest data rate becomes the majority choice. These results show that the proposed Fed-k-means algorithm can improve the effective throughput with an even smaller training dataset than the single user k-means algorithm.

In Figure 4, the effects of number of users on the effective throughput of the Fed-k-means schemes with different size of training dataset are depicted. Note that all the results in this part are average values of 50 simulations, and the SNRs for all the three settings are 4dB. The effective throughput increases when the number of users increases. The plots with 200 and 100 training data points reaches their highest effective throughput, at 5.5 bits/cu, at 50 users and 100 users,

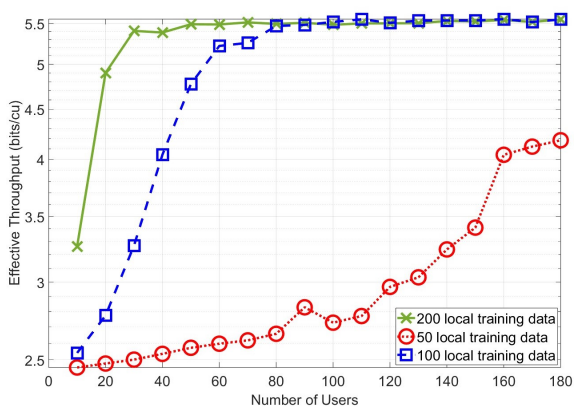



Fig. 4: Effective throughput versus number of users of Fed-k-means adaptive OFDM-IM.

respectively. This result indicates that the required number of users for achieving the best performance decreases when the size of training dataset in each user increases.

## V. Conclusion

This paper proposed the federated k-means clustering strategy for adaptive OFDM-IM. By aggregating the learning outcome of distributed users, the adaptation strategy developed at the BS improved the accuracy of the global learning model, requiring less training data from individual devices. With the global adaptation model, distributed users were able to reliably adjust OFDM-IM signals to their local conditions. The simulation results showed that the Fed-k-means OFDM-IM improved the throughput through the multi-user federation. Heterogeneous training features across users such as asymmetric sets of modulation modes will be investigated in the future.

## References

[1] T. Van Luong and Y. Ko, "Spread OFDM-IM with precoding matrix and low-complexity detection designs," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 12, pp. 11 619–11 626, 2018.

[2] E. Başar, "OFDM with index modulation using coordinate interleaving," *IEEE Wireless Communications Letters*, vol. 4, no. 4, pp. 381–384, 2015.

[3] A. T. Dogukan and E. Basar, "Super-mode OFDM with index modulation," *IEEE Transactions on Wireless Communications*, vol. 19, no. 11, pp. 7353–7362, 2020.

[4] Y. Ko and J. Choi, "Unsupervised machine intelligence for automation of multi-dimensional modulation," *IEEE Communications Letters*, vol. 23, no. 10, pp. 1783–1786, 2019.

[5] Y. Liu, X. Yuan, Z. Xiong, J. Kang, X. Wang, and D. Niyato, "Federated learning for 6G communications: Challenges, methods, and future directions," *China Communications*, vol. 17, no. 9, pp. 105–118, 2020.

[6] Y. Wang, G. Gui, H. Gacanin, B. Adebisi, H. Sari, and F. Adachi, "Federated learning for automatic modulation classification under class imbalance and varying noise condition," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 1, pp. 86–96, 2022.

[7] B. McMahan, E. Moore, and et al., "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[8] J. Konečnỳ, H. B. McMahan, D. Ramage, and P. Richtárik, "Federated optimization: Distributed machine learning for on-device intelligence," *arXiv preprint arXiv:1610.02527*, 2016.

[9] H. H. Kumar, V. Karthik, and M. K. Nair, "Federated k-means clustering: A novel edge AI based approach for privacy preservation," in *2020 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*. IEEE, 2020, pp. 52–56.

[10] Y. Liu, Z. Ma, Z. Yan, Z. Wang, X. Liu, and J. Ma, "Privacy-preserving federated k-means for proactive caching in next generation cellular networks," *Information Sciences*, vol. 521, pp. 14–31, 2020.