Federated Clustering Algorithm based on a Fuzzy Clustering Feature Vector

Asier Urío-Larrea
asier.urio@unavarra.es
Universidad Pública de Navarra
Pamplona, Navarra, Spain

ABSTRACT

The availability of a large amount of data produced in multiple devices and organizations provides great opportunities and challenges. A great amount of this data is privacy-sensitive and thus, sending it to a central server is not convenient. Furthermore, this data can have different characteristics in different producers making a unique globalized model not the best choice for its processing. Federated Learning is a concept developed to overcome these problems, through a local parameter learning step and a posterior parameter sharing step which does not involve sharing the data. In this work, a Federated Clustering algorithm is proposed. This algorithm is based on Fuzzy Cluster Feature vectors which can manage the uncertainty of points belonging to a certain cluster. The preliminary experimental results show that the proposal is equivalent to other State of The Art methods while requiring fewer communication rounds.

CCS CONCEPTS

• **Computing methodologies** → **Cluster analysis**; • **Information systems** → **Clustering**; **Clustering and classification**.

1 INTRODUCTION

Nowadays data is produced by a large amount of particular devices such as smartphones, wearables and home appliances (IoT). Furthermore, great volumes of sensitive data are generated in privacy-sensitive places such as health centres, schools and banks. Much knowledge can be extracted from these sources to improve people's lives, such as predictive typing in mobile phones [7] and drug discovery [3] in healthcare.

However, given the characteristics of that data, a number of problems arise. Of these, the most prominent are data privacy and the existence of agents that do not follow the same distribution of data. The first of these stems from a greater awareness of the management of personal data, which is already being regulated by various governments such as the European Union with its regulation on the protection of personal data (GDPR). The second can occur when data is generated at different points and in these cases, despite there being a global relationship between the data, each original agent follows a slightly different distribution (non-IID data) that decreases performance if only a global model is considered.

In order to overcome these challenges, the concept of Federated Learning (FL) arose. In 2016 McMahan [10] presented the FedAvg algorithm, which trained a neural network on each client and shared the set of weights with the server. The server averaged the weights learned by the clients and distributed the results back to them. McMahan reported the Federated model's good performance in image classification using convolutional neural networks and in language prediction tasks using recurrent neural networks. The challenge of data variability is studied in [14] where it is shown a performance reduction in situations where the data is non-IID and the algorithm does not take this into account. However, a correctly designed algorithm is able to overcome this limitation and generate a federated model with good performance in this type of heterogeneous environment [8].

Many studies have been carried out on FL in neural networks, but the concept is also applied to a variety of machine learning systems, such as Bayesian systems [4] and clustering [6].

In this work, a federated clustering algorithm based on a fuzzy version of a Clustering Feature (CF) vector is introduced. The fuzziness in this structure means that each point does not belong to a single cluster, but has a degree of belonging to each of them. This condition can be very valuable in a federated environment, where a given CF may slightly represent a certain structure in the client, but when aggregating it with the rest of the CFs from the other clients it could finally merge with another.

The structure for this work is as follows. Section 2 justifies the proposed structure for data summarization. In Section 3 the federated algorithm is presented, and Section 4 shows the preliminary experimentation and results for the proposal. Finally, Section 5 highlights some conclusions and suggests the directions for further research.

2 FUZZY FEATURE VECTOR

The BIRCH algorithm [13] was developed to deal with very large datasets, it employs a CF to incrementally store summary data information of the cluster. This structure comprises the number of data points (N), the linear sum of the data points ($\overline{LS}$) and the square sum of the data points (SS). As the data was analyzed incrementally, at certain moments some clusters had to be merged, this was possible due to the Additivity Theorem which states that the CF of the merge of two clusters is the addition of each of their components.

Zadeh introduced the theory of Fuzzy Sets [12] in which an object does not belong to one set and not to the others, instead, each object has a certain degree of membership, ranging from 0 to 1, to each of the sets.

This foundational work of fuzzy set theory led to the development of many related fields, among which is fuzzy clustering, in which each point has a degree of membership in each of the clusters,

KDD '24, August 25–29, 2024, Barcelona, ES

as the most representative example is the Fuzzy C-Means [2] which is a fuzzy version of the traditional K-means [9] algorithm.

Many researchers have developed fuzzy versions of hard-clustering algorithms, the work that inspired this proposal is the FuzzStream algorithm [5] which is a fuzzy version of the CluStream [1] algorithm. These algorithms deal with the clustering of data streams, however, the way in which the structure of the latter is extended to obtain a fuzzy version can be used analogously with BIRCH CFs.

The choice of the CF structure is due to its additive property, which will enable the server to merge the compatible structures from different clients. The proposed Fuzzy Cluster Feature Vector (FCF) structure differs from the original CF in that it includes the sum of memberships component (M) and the weighting strategy for the summation components. The description of the FCF is as follows:

- $\overline{LS}$: Linear sum of examples weighted by their membership to the FCF.
- SS: Quadratic sum of distances between the examples and the FCF prototype weighted by their membership to the FCF.
- N: Number of examples assigned to the FCF.
- M: Sum of memberships of the examples assigned to the FCF.

3 FFCF ALGORITHM

With the previously defined FCF structure, the Federated clustering algorithm based on a Fuzzy Clustering Feature vector (FFCF) consists of the following steps::

(1) Each client creates the necessary FCF structures from its data. These structures are created point by point, merging them when their similarity is above a specified threshold.
(2) Each client sends the created FCF structures to the server.
(3) The server combines the received structures merging them when necessary.
(4) The servers perform the Weighted FCM on the combined FCFs, considering the sum of memberships as weights.
(5) The server sends the combined FCF structures and the results of the WFCM to the clients.

This algorithm requires the following parameters: the minimum and maximum number of FCF, the merging threshold which determines whether two structures are similar enough to be merged. For the Weighted FCM, the number of clusters and the fuzziness degree m have to be specified.

4 EXPERIMENTAL RESULTS

To evaluate the performance of the algorithm, we reproduced the third case presented in [11] in which the dataset is composed of 4 clusters of 1000 points each. From this dataset, 4 partitions were made with a different number of points in each experiment. The clusters assigned to each client can be seen in Figure 1 and the number of points for each client is shown in the first column of Table 1, where each row represents a different experiment.

Two metrics are employed to test the performance. The Within Sum of Squared Error (WSSE), which is the sum of the squared distance from each point to the centre of its assigned cluster, divided by the product of the number of samples multiplied by the number of dimensions. The Outside SSE (OSSE) is analogous, but instead of the distance to its assigned cluster, it is the distance to the non-assigned clusters. It is desirable that the first be smaller and the second bigger. In order to assign each point to a cluster, we select the cluster with the greatest membership.

$$WSSE = \frac{\sum_{k=1}^{K} \sum_{x_j \in C_k} ||x_j - c_k||^2}{N \times d} \qquad OSSE = \frac{\sum_{k=1}^{K} \sum_{x_j \notin C_k} ||x_j - c_k||^2}{N \times d}$$

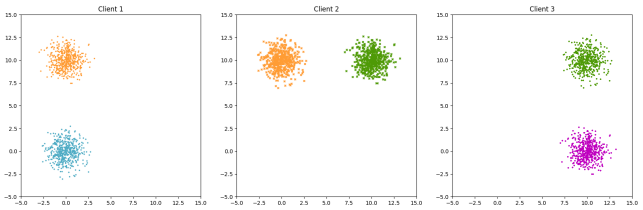

Figure 1: Distribution of the 4 clusters into the 3 clients

Table 1: Results

Points per Client	FFCM (avg2) WSSE	FFCM (avg2) OSSE	FFCF WSSE	FFCF OSSE
100 -1000-100	0.63	17.18	0.62	17.05
100 -1000-1000	0.63	17.21	0.61	17.11
1000-100 -100	0.63	17.18	0.62	17.09
1000-1000-1000	0.62	17.14	0.61	17.10

The FFCF algorithm's parameters were a merge threshold of 1 and a minimum number of FCF structures of 5 and a maximum of 100. The m parameter of the FCM was 2.

The obtained results can be found in Table 1. This table shows the best result for the reference FFCM and the ones obtained for the FFCF. On the one hand, we observe that FFCF performs slightly better regarding WSSE and slightly worse for OSSE. Despite these differences, we can say that this preliminary version of our algorithm is equivalent to the FFCM.

5 CONCLUSIONS AND FUTURE WORK

This work shows that the proposed Fuzzy Feature Vector-based Federated Clustering algorithm has a good performance in the studied case. One specific advantage of the proposal is that it requires fewer communication rounds than the FFCM. This demonstrates that the FCF structures are valid for a federated clustering application.

After this preliminary study, further experimentation is needed. On the one hand, more datasets have to be tested, both in IID and non-IID data distributions, synthetic and real. And, on the other hand, the performance of the algorithm has to be tested against a greater number of algorithms. The FFCF algorithm would benefit from an optimisation to address these new situations.

ACKNOWLEDGMENTS

Supported by MCIN/AEI/10.13039/501100011033/FEDER, UE through the project PID2022-136627NB-I00 and by Contratos predoctorales Santander-UPNA 2021

REFERENCES

[1] Charu C Aggarwal, S Yu Philip, Jiawei Han, and Jianyong Wang. 2003. A framework for clustering evolving data streams. In *Proceedings 2003 VLDB conference*. Elsevier, 81–92.

[2] James C Bezdek, Robert Ehrlich, and William Full. 1984. FCM: The fuzzy c-means clustering algorithm. *Computers & geosciences* 10, 2-3 (1984), 191–203.

[3] Shaoqi Chen, Dongyu Xue, Guohui Chuai, Qiang Yang, and Qi Liu. 2020. FL-QSAR: a federated learning-based QSAR prototype for collaborative drug discovery. *Bioinformatics* 36, 22-23 (2020), 5492–5498.

[4] Zhongxiang Dai, Bryan Kian Hsiang Low, and Patrick Jaillet. 2020. Federated Bayesian optimization via Thompson sampling. *Advances in Neural Information Processing Systems* 33 (2020), 9687–9699.

[5] Priscilla de Abreu Lopes and Heloisa de Arruda Camargo. 2017. Fuzzstream: Fuzzy data stream clustering based on the online-offline framework. In *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 1–6.

[6] Alessandro Epasto, Andrés Muñoz Medina, Steven Avery, Yijian Bai, Robert Busa-Fekete, CJ Carey, Ya Gao, David Guthrie, Subham Ghosh, James Ioannidis, Junyi Jiao, Jakub Lacki, Jason Lee, Arne Mauser, Brian Milch, Vahab Mirrokni, Deepak Ravichandran, Wei Shi, Max Spero, Yunting Sun, Umar Syed, Sergei Vassilvitskii, and Shuo Wang. 2021. Clustering for Private Interest-based Advertising. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (Virtual Event, Singapore) *(KDD '21)*. Association for Computing Machinery, New York, NY, USA, 2802–2810. https://doi.org/10.1145/3447548.3467180

[7] David Leroy, Alice Coucke, Thibaut Lavril, Thibault Gisselbrecht, and Joseph Dureau. 2019. Federated learning for keyword spotting. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 6341–6345.

[8] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. 2020. Federated Optimization in Heterogeneous Networks. In *Proceedings of Machine Learning and Systems*, I. Dhillon, D. Papailiopoulos, and V. Sze (Eds.), Vol. 2. 429–450. https://proceedings.mlsys.org/paper_files/paper/2020/file/1f5fe83998a09396ebe6477d9475ba0c-Paper.pdf

[9] James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1. Oakland, CA, USA, 281–297.

[10] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. 2023. Communication-Efficient Learning of Deep Networks from Decentralized Data. arXiv:1602.05629 [cs.LG]

[11] Morris Stallmann and Anna Wilbik. 2022. Towards Federated Clustering: A Federated Fuzzy c-Means Algorithm (FFCM). arXiv:2201.07316 [cs.LG]

[12] Lotfi Asker Zadeh. 1965. Fuzzy sets. *Information and control* 8, 3 (1965), 338–353.

[13] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. 1996. BIRCH: an efficient data clustering method for very large databases. *SIGMOD Rec.* 25, 2 (jun 1996), 103–114. https://doi.org/10.1145/235968.233324

[14] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. 2018. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582* (2018).