Community Cluster Method Federated Machine Learning for Secure & Private Survival Rate Prediction of Patients Infected from Coronavirus (COVID-19)

Divya[1,2], Vikram Singh[1], Naveen Dahiya[2]

[1]Chaudhary Devi Lal University, Sirsa, India
[2]Maharaja Surajmal Institute of Technology, New Delhi, India.

Abstract: In December 2019, a pneumonia outbreak was reported in Wuhan, China which was later contributed to a novel strain of coronavirus(COVID-19). By the month of February 2020, thousands of people around almost 27 countries of the world have been confirmed to having suffering from the disease with a large number of casualties. Electronic medical records and data can help the research community to predict the survival rate of the patient suffering from the deadly virus. But the data is existing in the form of isolated islands across the various healthcare institutions. Moreover, the privacy concerns and the security issues have further complicated the process of information sharing across the stakeholders and this has made training of Machine Learning models to predict the survival rate of patients a difficult task. In this context, Federated Machine Learning can provide a solution by keeping the training data localised, thereby preserving security and privacy of data. In this paper we propose a novel Community Cluster Method (CCM) Federated Machine Learning Model to predict the survival rate of a patient infected from Coronavirus (COVID-19).

Keywords: Distributed datasets, Data Privacy Preservation, Federated Learning, Community Cluster Method, identically independently distributed data.

1. Introduction

Since the time we have first used the term Artificial Intelligence in 1955, the concept has literally come of age nowadays. In 2016, when AlphaGo [1] defeated the top human players of the game, we have sort of seen an eye opener. All this has renewed the interest in the field with a new fervour. With the breakthrough technologies like the advancements in the field of mobile technology, connecting architectures, rapid rise in the number of mobile devices, and last but not the least, the internet, we are witnessing a huge potential in the field of artificial intelligence. Machine learning as an offshoot of artificial intelligence, has proved its effectiveness in a vast area of application including Pattern Recognition, Recommendation systems, Natural Language Translation, Virtual Assistants, Speech and Text Identification, predictive Analytics, Medical Diagnosis and Prognosis, etc. to name a few.

In order to develop effective, accurate and efficient models, the data plays a significant role. However, in real life scenario, most of the that could be used to train the Machine Learning models is found to be existing in the form of isolated cut-outs. Most of the industries either have a very limited access to data or don't have good quality data. Even if the industries are willing to co-operate for data sharing, there are technological, administrative, protocol level barriers, with industry competitions and cost also adding to the challenges. All this is a serious bottle neck in our vision of connected, more intelligent Machine Learning Models.

In the view of recent Facebook data breach [2], the industries at the macro level and the people at the micro level have increasingly become aware about the privacy and security of their data, and how, when and where their data is being used. This has even led to the enactment of GDPR[3] by European Union in 2018. The GDPR is designed to give users more control over their personal data [3][4][5]. In this eventuality, various countries all over the world have also framed and enacted similar legislations.

This has become a matter of concern for the research community, as the traditional manner in which the machine learning models were trained essentially involved data collection at one site, data cleaning at another, and model training at maybe some third site. The final model may be used elsewhere. Thus, there was a heavy reliance upon the data being shared among parties, but this needs to be reviewed keeping in mind the new regulations.

In the year 2017, Google introduced a new concept called Federated Learning [6][7][8] in which, the machine learning machine-learning models are built based on data that is not centralised like the traditional approach, but, distributed across multiple devices so as to prevent data leakage and address privacy and security concerns. In the traditional approach, the data owners send their data to a centralised location or cloud, where this data is aggregated to build the models. However, in the Federated approach, each of the contributing parties have their data at located at their respective sites and all the computations, data cleaning and model building are done

locally. All the participating parties send their individual model, not the data to the cloud where the individual models are aggregated to build a bigger and better centralised machine learning model. This framework applies to a data- partition framework where each partition corresponds to a subset of data samples collected from one or more users [9]. Thus, Federated Learning can be used effectively to train a machine learning model to be used in the healthcare domain.

2. Literature survey

The electronic record of health data or patient information has a key role in providing good healthcare. These records can be effectively used to build a machine learning model to perform various tasks like medical diagnosis and prognosis, effective drug delivery, personalised healthcare services to name a few. The electronic records have shown improved quality of healthcare in case of some serious diseases[10][11][12], reduction in undesired medical check-ups [13], cost economy for healthcare providers [14], improved medical education [15] to name a few.

In order to have better models for the healthcare sector, machine learning has used the electronic patient records in a disease prediction by using methods such as regression, k-nearest neighbour, decision trees and support vector machines for predicting disease like Type 2 Diabetes, one year in advance to the actual occurrence of the disease[16], for predicting the risk of a person committing suicide using EMR driven Boltzmann machines[17], age related muscular degeneration using deep neural networks[18], etc. These applications have made promising debut resulting into an improved sense of healthcare scenario among the stakeholders[19], but they all are based on the easy and readily availability of the data. Traditionally, healthcare data distributed across sites centralized in a database for access for analysis [20][21][22]. However, in the case of healthcare scenario, data transfers are complex due to strict regulations and sensitive nature of the data. These hurdles not only make data utilization expensive but also slow down information flow in healthcare where timely updates are often important. In reality, this is easier said than done, because being generated by multiple patients at diverse locations, integration and subsequent use of the data is a matter of concern.

Owing to the sensitive nature of the electronic healthcare records, the domain is facing a lot of challenges related to the concerns of data privacy, security and access. Nowadays, since the spread of diseases is vast both geographically and in terms of numbers, because of multiple factors, we have the data stored at multiple locations, which is not just limited to the hospitals or medical healthcare providers, but also to pharmacies and personal devices, to name a few[23][24]. This may impede adoption of machine learning methodology on the healthcare data in practice, and thus, has generated the interest of the researchers to find new methods or ways for secure, private, cost effective and easy medical data sharing among the parties [25][26]

In this regard Federated Learning has emerged as a good solution wherein both the data and computation are kept local and then the locally computed results are aggregated to train a global predictive model [27][28]. Indeed, FL foregoes the need of collection and sharing of data, and thus can act as a good framework for developing machine learning applications on privacy-sensitive electronic medical records. Federated Learning is found to be performing well with identically independently distributed data(IID) but with non-identically independently distributed data the results may not be that good[29][30][31][32]. Moreover, the healthcare records are generally of non-identically independently distributed nature [33].

Here we are proposing an algorithm inspired by deep embedding clustering in order to handle non-identically independently distributed healthcare data. Clustering has shown promising results in diagnosis of various diseases like diabetes, diagnosis of cancer symptoms and chronic pain treatment to name a few. We have used the data corresponding the the patients suffering from Coronavirus and proposed a Community Cluster Method based Federated Machine Learning model that can predict the survival rate of the patients.

3. Proposed algorithm/ Methodology

In this section we describe our Community Cluster Method based Federated Learning Model. We have considered three procedures viz., Encode, K-means clustering and community clustering. In the first procedure, each hospital is considered as a client that learns de-noising auto-encoder F_{enc} that is initialised with $W_{0,enc}$ for E_1

epochs and it returns only the trained weights of encoder $W_{1,enc}$ to the server for average. The total number of examples are denoted by M, the client index is denoted by h, the size of each client is denoted by N^h and the averaged encoder is denoted by F_{enc}.

In the second procedure i.e. k means clustering procedure, each client uses F_{enc} to transform its data into representations R^h and compute the average representation R'^h which are then sent to the server. The server, then uses k means clustering method F_{kmeans} with k communities on R'^h from all the clients.

Algorithm 1. Community Cluster Method Federated Machine Learning

1. **Procedure One: Encode (H, E_1)**
2. initialize weights $W_{0,enc}$ for de-noising auto-encoder F_{enc}
3. **for** each client h=1,2,….,H, parallely do
4. train F_{enc} for epoch E_1in order to obtain trained weights of encoder $W^h_{1,enc}$
5. **return** $W^h_{1,enc}$ to the server
6. perform update to obtain F_{enc}, $W_{1,enc} = \sum_{h=1}^{H} \frac{N^h}{M} W^h_{1,enc}$
7. **Procedure Two: K-means Clustering(K,H)**
8. **for** each client h=1,2,….,H, in parallel, do
9. use F_{enc} for finding encoded features R^h from each sample
10. **return** $R'^h = \sum_{i=1}^{N^h} \frac{R^i}{N^h}$
11. initialize k cluster centroids from $\{R'^1, R'^2, \ldots, R'^h\}$
12. train k-means clustering model F_{kmeans}
13. **for** each client h=1,2,….., H, **parallely do**
14. use F_{kmeans} on R^h to find the cluster of each example
15. count examples in each cluster $\{k_1^c, k_2^c, \ldots . k_n^c\}$
16. **return** example count value to the server
17. **Procedure Three: Community Cluster (N,H,E_2)**
18. initialize N community models $\{F_1, F_2, \ldots, F_N\}$ each having the same weight W_0
19. **while** the models are not converged, **do**
20. **for** n in 1. ..N, **parallely do**
21. **for** each client h=1,2,….,H, **parallely do**
22. train F_n^h for E_2 epochs to obtain W_n^h
23. **return** W_n^h to the server
24. $W_n = \frac{\sum_{h=1}^{H} K_n^h W_n^h}{\sum_{h=1}^{H} K_n^h}$
25. end while

The community cluster procedure starts with server being initialized by N neural network models $F_1, F_2, \ldots, F_N$, each having same weight W_0. The server provides all the N models to every client and each client learns every model on its full data for E_2. In the meantime F_{enc} and F_{kmeans} determine to which cluster does each example belongs to. The size of clusters is denoted by $k_1^c, k_2^c, \ldots, k_n^c$ and it is returned to the server along with the learnt weights. At the server each model is updated by taking the weighted average of j based on $k_1^c, k_2^c, \ldots, k_n^c$. The updated models are then sent to every client for the next round of training. This learning process is repeated till the convergence of the algorithm. The convergence condition for the model is that either the weights of the server-side global model converge to some given values, or that the number of maximum communication rounds are reached. So for a given test sample, the community cluster based federated learning, first converts the features into encodings by F_{enc} then define its community by F_{kmeans} and finally use the corresponding community model to make prediction.

4. Research Limitations

The coronavirus(Cov) has emerged as a major threat to the people not just in China or the neighbouring countries, but as a world level issue that needs to be addressed. At the time of doing this work, in February 2020, the number of confirmed cases of coronavirus is more than 75000, with the death toll in China being approx. 2500 persons. The spread of the disease on a worldwide level makes it necessary to integrate all the available data to build effective model to predict the survival rate of the patients, keeping their personal information secure and safe. The algorithm that we have proposed is based on the knowledge that is available to us so far, and this information is bound to be more clear in the upcoming time.

5. Conclusions & Future Scope

The Community Cluster Method Federated Machine Learning as proposed in this work provides for an efficient method to predict the survival rate of the Coronavirus patients, keeping the personal data safe and private. The model is built on the data that we have gathered so far, and since this has reached epidemic proportions, a lot of data is going to be available in the future. That will surely help in making the model better and improving its predicting accuracy. Further work can be done in the field to add more features to the predicting task based on the availability of the statistics.

References:

1. Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D.: Mastering the game of Go with deep neural networks and tree search. Nature. 529, 484–489 (2016). https://doi.org/10.1038/nature16961.
2. 2. Wikipedia. 2018. Facebook–Cambridge Analytica Data Scandal. https://en.wikipedia.org/wiki/Facebook-Cambridge_ Analytica%_data_scandal. - Google Search, last accessed 2020/02/25.
3. Hoofnagle, C.J., Sloot, B. van der, Borgesius, F.Z.: The European Union general data protection regulation: What it is and what it means. Inf. Commun. Technol. Law. 28, 65–98 (2019). https://doi.org/10.1080/13600834.2019.1573501.
4. Albrecht, J.P.: How the GDPR Will Change the World. Eur. Data Prot. Law Rev. 2, 287–289 (2016). https://doi.org/10.21552/EDPL/2016/3/4.
5. [Mayer-SchonbergerandPadova2015] Mayer-Schonberger, V., and Padova, Y. 2015. Regime change: Enabling big data through europe's new data protection regulation. Colum. Sci. & Tech. L. Rev. 17:315. - Google Search, last accessed 2019/12/24.
6. Konečný, J., McMahan, H.B., Ramage, D., Richtárik, P.: Federated Optimization: Distributed Machine Learning for On-Device Intelligence. (2016).
7. Konečný, J., McMahan, H.B., Yu, F.X., Richtárik, P., Suresh, A.T., Bacon, D.: Federated Learning: Strategies for Improving Communication Efficiency. (2016).
8. Mcmahan, H.B., Moore, E., Ramage, D., Com, B.: Federated Learning of Deep Networks using Model Averaging Blaise AgüeraAg˘Agüera y Arcas. (2012).
9. Cheng, K., Fan, T., Jin, Y., Liu, Y., Chen, T., Yang, Q.: SecureBoost: A Lossless Federated Learning Framework. (2019).
10. Cebul, R.D., Love, T.E., Jain, A.K., Hebert, C.J.: Electronic health records and quality of diabetes care. N. Engl. J. Med. 365, 825–833 (2011). https://doi.org/10.1056/NEJMsa1102519.
11. Neumann, A., Kalenderian, E., Ramoni, R., Yansane, A., Tokede, B., Etolue, J., Vaderhobli, R., Simmons, K., Even, J., Mullins, J., Kumar, S., Bangar, S., Kookal, K., White, J., Walji, M.: Evaluating quality of dental care among patients with diabetes: Adaptation and testing of a dental quality measure in electronic health records. J. Am. Dent. Assoc. 148, 634-643.e1 (2017). https://doi.org/10.1016/j.adaj.2017.04.017.
12. Dorr, D., Bonner, L.M., Cohen, A.N., Shoai, R.S., Perrin, R., Chaney, E., Young, A.S.: Informatics Systems to Promote Improved Care for Chronic Illness: A Literature Review. J. Am. Med. Informatics Assoc. 14, 156–163 (2007). https://doi.org/10.1197/jamia.M2255.
13. Podichetty, V., Penn, D.: The progressive roles of electronic medicine: Benefits, concerns, and costs, (2004). https://doi.org/10.1097/00000441-200408000-00005.
14. Simon, S.J., Simon, S.J.: An examination of the financial feasibility of Electronic Medical Records (EMRs): a case study of tangible and intangible benefits. Int. J. Electron. Healthc. 2, 185–200 (2006).

https://doi.org/10.1504/IJEH.2006.008832.

15. Tierney, M.J., Pageler, N.M., Kahana, M., Pantaleoni, J.L., Longhurst, C.A.: Medical education in the electronic medical record (EMR) era: Benefits, challenges, and future directions, http://www.ncbi.nlm.nih.gov/pubmed/23619078, (2013). https://doi.org/10.1097/ACM.0b013e3182905ceb.
16. Mani, S., Chen, Y., Elasy, T., Clayton, W., Denny, J.: Type 2 diabetes risk forecasting from EMR data using machine learning. AMIA Annu. Symp. Proc. 2012, 606–615 (2012).
17. Tran, T., Nguyen, T.D., Phung, D., Venkatesh, S.: Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM). J. Biomed. Inform. 54, 96–105 (2015). https://doi.org/10.1016/j.jbi.2015.01.012.
18. Lee, C.S., Baughman, D.M., Lee, A.Y.: Deep Learning Is Effective for Classifying Normal versus Age-Related Macular Degeneration OCT Images. Kidney Int. Reports. 1, 322–327 (2017). https://doi.org/10.1016/j.oret.2016.12.009.
19. Miotto, R., Wang, F., Wang, S., Jiang, X., Dudley, J.T.: Deep learning for healthcare: review, opportunities and challenges. https://doi.org/10.1093/bib/bbx044.
20. Sanchez-Pinto, L.N., Churpek, M.M.: Predictive Analytics and Machine Learning in Medicine. In: Actionable Intelligence in Healthcare. pp. 179–200. Auerbach Publications (2018). https://doi.org/10.1201/9781315208442-10.
21. Hashem, I.A.T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A., Ullah Khan, S.: The rise of "big data" on cloud computing: Review and open research issues, (2015). https://doi.org/10.1016/j.is.2014.07.006.
22. Holzinger, A.: Machine learning for health informatics. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 1–24. Springer Verlag (2016). https://doi.org/10.1007/978-3-319-50478-0_1.
23. Patel, A., Singh, N.M., Kazi, F.: Internet of Things and Big Data Technologies for Next Generation Healthcare. 23, (2017). https://doi.org/10.1007/978-3-319-49736-5.
24. Dubovitskaya, A., Xu, Z., Ryu, S., Schumacher, M., Wang, F.: Secure and Trustable Electronic Medical Records Sharing using Blockchain. AMIA ... Annu. Symp. proceedings. AMIA Symp. 2017, 650–659 (2017).
25. Yang, J.J., Li, J.Q., Niu, Y.: A hybrid solution for privacy preserving medical data sharing in the cloud environment. Futur. Gener. Comput. Syst. 43–44, 74–86 (2015). https://doi.org/10.1016/j.future.2014.06.004.
26. Brendan McMahan Eider Moore Daniel Ramage Seth Hampson Blaise AgüeraAg, H., Arcas, A.: Communication-Efficient Learning of Deep Networks from Decentralized Data. (2017).
27. Liu, D., Miller, T., Sayeed, R., Mandl, K.D.: FADL:Federated-Autonomous Deep Learning for Distributed Electronic Health Record. (2018).
28. Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., Chandra, V.: Federated Learning with Non-IID Data. (2018).
29. Jeong, E., Oh, S., Kim, H., Kim, S.-L., Park, J., Bennis, M.: Communication-Efficient On-Device Machine Learning: Federated Distillation and Augmentation under Non-IID Private Data.
30. Huang, L., Yin, Y., Fu, Z., Zhang, S., Deng, H., Liu, D.: LoAdaBoost:Loss-Based AdaBoost Federated Machine Learning on medical data.
31. Rosales, R.E., Rao, R.B.: Guest Editorial: Special Issue on impacting patient care by mining medical data, (2010). https://doi.org/10.1007/s10618-010-0167-9.
32. Xie, J., Girshick, R., Farhadi, A.: Unsupervised Deep Embedding for Clustering Analysis. (2016).
33. Greenfield, S., Kaplan, S.H., Kahn, R., Ninomiya, J., Griffith, J.L.: Profiling Care Provided by Different Groups of Physicians: Effects of Patient Case-Mix (Bias) and Physician-Level Clustering on Quality Assessment Results Background: Patient characteristics (case-mix bias) and physi. (2002).