Hierarchical fuzzy neural networks on heterogeneous big data with privacy preservation

Leijie Zhang, Ye Shi, CT Lin

***Abstract*—**

***Index Terms*—Hierarchical fuzzy neural networks, heterogeneous big data, distributed clustering, privacy preservation**

I. Introduction

Big data has been a ubiquitous term as enormous amount of data is being and continually generated at an unprecedented increasing scale in many areas including social networks, internet of things, commerce, astronomy, biology, medicine, etc [1], [2]. Due to the massive amount, high dimension, heterogeneity and unpredictability characteristics of the big data, various challenges are posed on the traditional data mining and machine learning methods. For instance, the massive and high dimensional big data will result in the requirement of extraordinary computational power; the heterogenenity ....

Heterogeneous big data ...

Why fuzzy? Mentioned some advantages of FNN to tackle data uncertainty.

Uncertainty: Loosely defined as "how well the data speaks to the question of interest with respect to the model", uncertainty over whether the available data will meet the demands of the task begins at the collection stage. Neither sensors or humans are immune to measurement errors; qualitative data, such as social media accounts, are often subjective in nature; and the data collected may not be relevant or might be incomplete. Currently, the best solution for dealing with uncertainty is to build the model through a fuzzy neural network which involves fuzzy rules and fuzzy inference systems to make optimal use of the given data [8-9]. DPCL will be capable of dealing with uncertainty by its fuzzification operation and if-then-rule architecture.

Every gain made by the last decades to improve our lives in social networks, commerce, astronomy, biology, medicine and beyond comes at threat of violations to the privacy and security of our personal data [1-2]. Even when data custodians take the greatest care not to improperly divulge our sensitive information, there are still countless ways an adversary may can illicitly access our information for nefarious purposes. One of the most difficult, and therefore common, attacks to counter is inference – observing seemingly non-sensitive information, such as one's current location, and using that to infer far more valuable data, such as medical histories or financial information. Satisfying differential privacy, DPCL will guarantee data protection from all currently-known adversarial attacks, including inference attacks, which most commonly occur during data transmission. Differential privacy guarantees are well proven to hold in a centralised settings where a single organisation owns all the data [4-5]. We intend to extend this guarantee to settings with multiple parties.

Multi-party computation: Privacy problems becomes even more severe when different organisations wish to collaboratively learn from their each other's data without fully disclosing that data to the other parties or breaching any existing privacy agreements. However, recently, distributed machine learning algorithms have emerged as a solution to preserving privacy in these multi-party data sharing scenarios [6-7]. The basic idea is that multiple all the participants jointly compute the model parameters and exchanging their results with their neighbours. DPCL's algorithm will be based on the ADMM approach, which is a well-known method of solving consensus optimisation problems in a distributed setting.

In summary, DPCL will allow multiple parties to benefit from sharing data, and the powerful insights analytics with shared data can allow, without compromising commitments given over data privacy, without leaving data vulnerable to security attacks, and without concern over whether the data provided by each party is relevant, complete, or compatible. The unique combination of an differential privacy guarantee in a multi-party setting with a distributed fuzzy neural network accomplishes all three of these goals.

Literature about heterogeneous data, ... Usually, heterogeneous databases can use principle component analysis (PCA) and correlation analysis to reduce the number of dimensions and consider the interactive relations between features in past research [13]. However, as PCA and correlation analysis do not consider nonlinear representative or nonlinear interactive relations between features, they might lose some important information for clustering. Hence, in this paper, instead of using PCA, we consider the double deep autoencoder model for the above issues, as follows:

Privacy concern: distributed clustering method ...

The uncertainty in the data may arise due to many factors including missing values, imprecise measurements, changes in process characteristics during the data generation period, lack of appropriate monitoring of data measurement process to name a few. Internet-of-Things (IoT) systems usually generate a large amount of unstructured and heterogeneous data demanding specialized techniques for data analytics. Thus, decision making in such an environment poses significant challenges and often demands new and innovative design techniques and algorithms for decision making.

Leijie Zhang, Ye Shi and CT Lin are with the School of Computer Science, University of Technology, Sydney, NSW 2007, Australia. Email: Leijie.Zhang@student.uts.edu.au,ye.shi-1@uts.edu.au, Chin-Teng.Lin@uts.edu.au.

High-dimensional types data leads to distributed machine learning. Additionally, the distributed algorithm has the following advantages: parallel and fast computation speed, privacy preserving and system robustness.

The distributed algorithm is built on the well-known alternating direction method of multipliers (ADMM)

Heterogeneous types big data: An intuitive way is to use a hierarchical structure to cluster the different types of the features.

The hierarchical FNN is built on the alternating optimization (AO).

No works on privacy-preserving hierarchical FNN.

The aim of this work is to bring together the advantages of both hierarchical and distributed framework into a single fuzzy system to tackle the massive and heterogeneous data, uncertainty issue and privacy concerns in big data environment.

Compared with gradient-type algorithms, alternating minimization has several advantages: (i) it is easy to implement as there is no need to tune optimization parameters like step sizes, (ii) it converges very fast in practice, and (iii) the subproblems are easy to solve as they usually have closed-form solutions. Thus, alternating minimization has been widely used in practice ...

Advantages Each iteration usually cheap (single variable optimization) No extra storage vectors needed No stepsize tuning No other pesky parameters that must be tuned Simple to implement Works well for large-scale problems Currently quite popular; parallel versions exist Disadvantages Tricky if single variable optimization is hard Convergence theory can be complicated Can slow down near optimum Non-differentiable case more tricky

Our contribution: 1. Use hierarchical structure of FNN to tackle heterogeneous big data 2. Use distributed clustering method for computational acceleration and privacy preservation. 3. Use AO to get a local optimal solution. The AO it is easy to implement as there is no need to tune optimization parameters like step sizes. Furthermore, it converges very fast in practice.

The effectiveness of our method is verified by large-scale regression and classification datasets.

II. Model Formulation for the HFNN

Mention why we use HFNN, heterogenous again ...

Put the structure of HFNN ...

III. Two-stage optimization algorithm to train the HFNN

A. Distributed clustering method for the low-level of HFNN

B. Alternating optimization for the high-level of HFNN

IV. Simulation results

V. Conclusion

Future work: semi-supervised learning and online learning of the HFNN. Another possible direction is to increase the layers of the consequent part of HFNN.

References

[1] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, "A survey of machine learning for big data processing," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, pp. 1–16, 2016.

[2] H.-Y. Tran and J. Hu, "Privacy-preserving big data analytics a comprehensive survey," *Journal of Parallel and Distributed Computing*, vol. 134, pp. 207–218, 2019.