**北 京 科 技 大 学**

**硕士学位研究生**

**选题报告及文献总结**

论文题目**Performance Optimization Of Privacy-Preserving Distributed Learning On Healthcare Data**



指导教师： 黄旗明副教授

单 位：计算机与通信工程学院通信工程系

学 号： S20191482

作 者： Ousman Manjang 奥斯曼

专业名称： 信息与通信工程

入学时间： 2019年9月

2021年 01 月 04

Table of Contents

[**Table of Contents 0**](#_Toc462651559)

[**1** **Introduction 1**](#_Toc462651560)

[1.1 PRIVACY OF HEALTHCARE DATA 1](#_Toc462651561)

[1.2 MACHINE LEARNING 1](#_Toc462651562)

[1.3 DISTRIBUTED LEARNING 2](#_Toc462651563)

[**2** **Related works 4**](#_Toc462651567)

[**3** **rESEARCH mETHOD** 6](#_Toc462651570)

[3.1 Data COLLECTION 6](#_Toc462651571)

[3.2 DATA PROCESSING 6](#_Toc462651572)

[**4** **Research CONTENT & OBJECTIVES**](#_Toc462651573)

[4.1 OBJECTIVES 6](#_Toc462651574)

[4.2 RESEARCH CONTENT 7](#_Toc462651575)

[4.3 SIGNIFICANCE 7](#_Toc462651576)

[**5** **References** 8](#_Toc462651577)

1. **INTRODUCTION**

**1.1 PRIVACY OF HEALTHCARE DATA**

In recent years, with a large amount of digitized data, and the demand for sharing and exchanging of data is high because it is essential to get the useful and previously unknown information. Many data mining exist and they require exchange of some or all of the useful information, but this research will be focus on medical or healthcare data. Studies have shown that a more convincing diagnosis can be obtained when the data is analyzed by different expert groups. This prompted the need for the sharing of sensitive information since data can be stored electronically and access remotely [2].

However, protecting the privacy and security of these data remain major barriers during this process. Since medical records are often related to related the confidential data obtained from patient and the physician may consider other data useful which are obtained via data sharing, and business entities [5], [17]. Medical data mining also has some unique features which should be put into consideration such as its heterogeneity, ethical issues, legal and social issues [1].

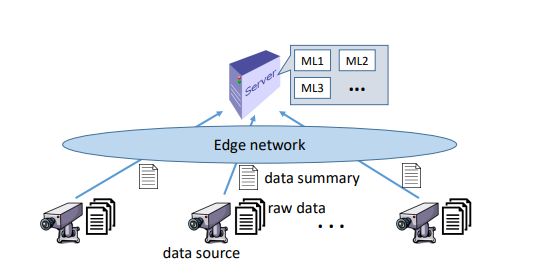
**1.2 MACHINE LEARNING**

Machine learning (ML) has been effective and successful in many domains, characterized by complex models training on large amount of data. In most cases, these data are being collected from different sources such as internet, phones, cars, Internet of thins, etc. Interestingly, large amount of data are produced on daily basis but the limitation lies at accessing these data due to some data protection laws [16] and also more people are now are of scandals surrounding data privacy such as the Cambridge Analytica [21]. Many different approaches have been studied and each with their limitations but one that is recognized by many is adding noise [23] and this will render the data almost useless. A more complex way of adding noise is elaborated in [6], basically there is always a tradeoff between protecting the privacy of the data and performance of the training, communication and privacy tradeoff.

**1.3 PRIVACY AND DISTRIBUTED LEARNING**

The privacy issues are handled by many algorithms and techniques in literature. But, always there exists a tradeoff between privacy loss and information loss. So, to preserve accuracy of results and to reduce loss of information, task based privacy preserving techniques are developed.

Distributed machine learning is of the most populous research fields dealing with large dataset [24], since the data is distributed among the computing nodes, it is often challenging to gather all the data required for effective training of the model. There are various ways of collecting these data but one that is traditionally accepted is to combine the outputs of the learned models [18]. Currently, the distributed learning techniques can be grouped into three main categories [25], the first one are those that globally aggregate the outputs of local models; the second one are those that construct global models from individual models derived from local data, and third are those that share representative local data with a global aggregator. My proposed work will be based on the third model which focuses on sharing the summaries of the source model into the creation of global model as illustrated in figure 1 below.



***Figure 1: Application scenario (ML i: Machine Learning Model i).***

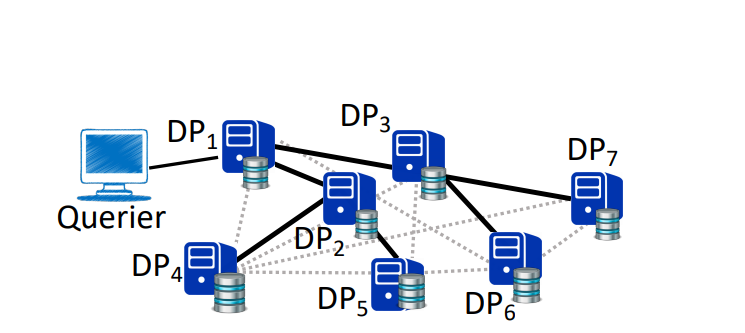
**ALGORITHM**

The stochastic gradient descents’ (SGD) performance can be optimized and represented in polynomial form. A typical example is illustrated by the Nesterov Accelerated SGD [4] (NASGD). The main ideal of this approach is to reduce the required number of training iteration thereby improving it performance and achieving the optimization. The algorithm proceeds by considering the previous gradients through the update rule.

.

where , with v being the velocity vector and µ controlling how much the accumulated previous gradients are taken into account. For a DPi , the local update rule (during local iteration l and global iteration j) becomes and

Where W(i,j,l) DPi’s local model at global iteration j and local iteration l W(i,j) DPi’s local model at global iteration J, DPi i th Data provider.



***Figure 2: Data Acquisition steps***

During this training iteration, one important factor to consider is over-fitting. An undesirable condition referred to as over-fitting and under-fitting occurs when we pre-define the required number of iteration. As such the performance of the training model will be ineffective and therefore will not be able to achieve the optimization. The data acquisition steps from the querier sending their request and the encrypted data being finally returned to the querier is classified into five major steps in [26] as illustrated in figure. The “Prepare” is the first stage of this procedure and it involves each data provider pre-processing their data. Next is the training of these local data by the individual data provides and the process is referred to as “map”. These locally trained data has to be aggregated in a “combine” stage in order to update the global model which is the “reduce” stage. The prediction is the final phase of this procedure.

The goal of this research is to eliminate this over-fitting and achieve optimization. A new stopping criterion is proposed that is based on the reduction of mean squared error (MSE) for a validation data set. An excellent result can obtained by stopping the network training when the reduction of MSE between training iterations became marginal.

1. **RELATED WORK**

The works related to this proposal can be view in two dimensions. One is in relation to the collective test protocol which is one method of eliminating if under-fitting and over fitting. Those conditions usually exist when we predefine the required number of training iterations. The other part has to do with protecting the privacy and confidentiality of the training data by various encryption schemes.

In recent years, the research of privacy-preserving has received a considerable attention. Most of the proposed protocols focused on the data which are partitioned either horizontally or vertically across distributed servers [7], [8], [9], [15]. A majority of these protocols employed secure multi-party computation (SMPC) framework in build a linear model over the joint dataset cooperatively. While many of the others rely on trusted third party, some of the approaches are based on secret sharing techniques [7], [8]. A typical example is the work done by author in [7], who framework presented two secure linear regression methods. These two methods are based on secure data integration and secure multi-party computation. The first method integrates datasets, which implies that no party can learn others data which is the main goal of the privacy-preserving. The other one uses the secure summation protocol that facilitates each party to share the local statistical information for computing the global regression coefficients.

In some approaches which are cryptography-based, Cryptographic tools are often incorporated so as to protect and give privacy to the dataset. In theory, secure multiparty computation [19], [20] can be very useful to resolve multi-party privacy-preserving deep learning, but the very high communication and computational complexity usually makes it impractical, even in the two-party case. In [27], Bansal et al. use a slightly different approach. It made use of the secure scalar and secret sharing in order protect against privacy leakage during the learning process. It is important to note that this approach has its own limitations as the scheme applies to the two-party setting, and it is not trivial to extend it to multi-party models. In [3], the author propose a privacy-preserving which goes by back-propagation algorithm and it is also suitable for multi-party deep learning over arbitrarily partitioned datasets based on homomorphic encryption [22]. However, unlike the other schemes mentioned above, this scheme requires that all the parties must be online and work interactively to decrypt the ciphertext of the intervening parameters in each of the iterations.

The de facto standard for privacy-preserving deep learning is [13], in this method, each party downloads the learning model directly from the centralized server and the local data is used to train the model. Upon completion of the training at each node, the resulting data is aggregated to the central server. One benefit of this scheme is that privacy can be easily achieved since the parties do not need to share the raw data. However, it requires a huge network bandwidth in order to share that trained data and the computation being done in every platform. Still on the methods of preserving the privacy of sensitive data, the data publishing techniques such as data perturbation can transform the data while maintaining its privacy [10] [11] [12]. Data perturbation can be very effective in this case is adapted in [11]. It is a technique of adding/multiplying noise to the original data. It is extremely difficult to reconstruct the original data set and it has a very high accuracy level with regards to its sensitivity to the reconstruction algorithms. Since only the perturbed table and the distribution function is used and not the original data, the chances for the leakage becomes very less.

While some focuses on how the data might be shared between different computing nodes, some techniques lay emphasis on how data is collected without breaking the individual’s privacy. This is referred to as local differential privacy (LDP). The LDP is a way of adding designed random noises to the original data and sends the noisy data to a data collector, guaranteeing that the data contributor’s original data is not leaked during the data collection process. It has recently attracted attention of many researchers as a promising way of guaranteeing individual privacy in the process of data collection [14], [15].

1. **METHODOLOGY & MATERIALS**

In this section, I will discuss the proposed methodology to accomplish the work as follows:

**3.1 DATA COLLECTION**

The data required for this work is a very sensitive data; that is healthcare data. It is often not always readily available from the internet for everyone. Different sets of data will be used. I proposed to find and download excel files to be used to train my model. Another set of data can be collected by granting permission to access to real data used by either medical schools or researchers to train the model.

**3.2 DATA PROCESSING**

The data collected based on the above procedures will e processed, put to test any analyzed. The data will be trained using the Nesterov Accelerated Gradient Algorithm. This will be implement with the aid of Go Programming Language. Python will also be used if necessary at the later stage to help in the data analysis.

### **RESEARCH CONTENT & OBJECTIVITY**

**4.1 RESEARCH OBJECTIVES**

* Danger, major concern and topic
* To address the need of privacy-preserving of healthcare data.
* Emerging issues to protect sensitive data.
* Find easy way to control unauthorized data access.
* Improve the training of the machine learning model to be fast and more accurate (optimization)
* Investigate healthcare data privacy-preserving schemes and investigate .

## 4.2 WORKING PROPOSAL:

Based on the above problems and the current research progress, the main research contents and the exact solution that this research is aimed to address is as follows.：

1. Nesterov Accelerated Stochastic Gradient Decent NASGD.
2. The methods of collective test protocols for early stop. The following methods will be used extensively to eliminate the overfitting problem

* Stop when the generalization loss exceeds a certain threshold,
* Stop when the quotient of generalization loss and progress exceeds a certain threshold and
* Stop when the generalization error increased in s successive strips.

### **RESEARCH SIGNIFICANCE**

The concern for data protection from malicious partner and data miners is increasing rapidly. Healthcare data is one of the most sensitive data that needs to be protected. While a lot of healthcare data are being generated globally, numerous issues are raised as this data are related to individuals privacy. There is a need for medical experts to consider data from different places to make a well-informed diagnosis as a result of getting a lot of information. This can be successfully achieved by distributed learning algorithms. This research aimed at find new methods of implementing the distributed learning scheme to learn from the medical data sets. It will make it easier for researchers and doctors to easily access a large range of data for their work. With proper privacy-preserving scheme employed, patients will be not have to worry about the breach of their personal data and thus will not be reluctant to provide sensitive data. This too comes as a benefit to the people as it gives too much assurance.

To alleviate the problem discussed in above section, numerous published data are available, but there is still a lot of work that have to be done. So, there is a need to develop an alternative approach that will resist privacy breach from malicious attackers and computing nodes simultaneously.

5. **REFERENCES**

[1] K. J. Cios & G. W. Moor, “Uniqueness of medical data mining”, Artificial Intelligencein Medicine, Vol. 26, Issue 1-2, pp.1-24.

[2] D. B. Taichman et al., ‘‘Sharing clinical trial data: A proposal from the international committee of medical journal editors free,’’ PLoS Med., vol. 13, no. 1, pp. 505–506, Apr. 2016.

[3]. Learning with privacy at scale. <https://machinelearning.apple.com/docs/learning-with-privacy-at-scale/appledifferentialprivacysystem.pdf>, 2018.

[4] Y. Nesterov. Smooth minimization of non-smooth functions. Mathematical programming, 2005.

[5] H. M. Krumholz and J. Waldstreicher, ‘‘The yale open data access (YODA) project—A mechanism for data sharing,’’ New England J. Med., vol. 375, no. 5, pp. 403–405, 2016.

[6] C. Huang, P. Kairouz, X. Chen, L. Sankar, and R. Rajagopal. Context-aware generative adversarial privacy. Entropy, 19(12), 2017.

[7]. M.D. Cock*et al.* Fast, privacy preserving linear regression over distributed datasets based on pre-distributed data Proceedings of the ACM Workshop on Artificial Intelligence and Security (AISec) (2015)

[8]. W. Du*et al.* Privacy-preserving cooperative scientific computations Proceedings of the IEEE Computer Security Foundations Workshop (CSFW-14) (2001)

[9]. W. Du*et al.* Privacy-preserving multivariate statistical analysis: Linear regression and classification Proceedings of the International Conference on Data Mining (SDM) (2004).

[10] Agrawal R., Srikant R.,”Privacy-Preserving Data Mining”, ACM SIGMOD Conference, 2000

[11] Agrawal D., Aggarwal C. C., “On the Design and Quantification of Privacy- Preserving Data Mining Algorithms”, ACM PODS Conference, 2002

[12] Wenliang Du and Zhijun Zhan, "Using Randomized Response Techniques for Privacy-Preserving Data Mining”, SIDKDD 2003.

[13] H.B. McMahan, E. Moore, D. Ramage, S. Hampson, and B.A.y. Arcas, “Communication-efficient learning of deep networks from decentralized data,” in Proc. AISTATS, 2017.

[14].G. Fanti, V. Pihur, and U. Erlingsson. Building a RAPPOR with the unknown: Privacy-preserving learning of associations and data dictionaries. In *Proceedings of the Privacy Enhancing Technologies Symposium*, 2016.

[15]J. Tang, A. Korolova, X. Bai, X. Wang and X. Wang. Privacy loss in Apple’s implementation of differential privacy on MacOS 10.12. , 2017.

[16] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union, L119:1–88, 2016.

[17] W. Raghupathi and V. Raghupathi, ‘‘Big data analytics in healthcare: Promise and potential,’’ Health Inf. Sci. Syst., vol. 2, no. 1, p. 3, 2014.

[18] P. K. Chan and S. J. Stolfo, “Toward parallel and distributed learning by meta-learning,” in AAAI Workshop in Knowledge Discovery in Databases, 1997.

[19] Jiexing Li, Yufei Tao, Xiaokui Xiao, “Preservation of Proximity Privacy in Publishing Numerical Sensitive Data”, ACM SIGMOD, 2008

[20] S. R. M. Oliveira and O. R. Zaïane, "Privacy Preservation When Sharing Data for Clustering", International Workshop on Secure Data Management in a Connected World, 2004

[21] M. Rosenberg, N. Confessore, and C. Cadwalladr. How Trump consultants exploited the Facebook data of millions. https://www.nytimes.com/2018/03/17/us/politics/ cambridge-analytica-trump-campaign.html, 2018. Retrieved May 23, 2019.

[22].B. Ding, J. Kulkarni and S. Yekhanin. Collecting telemetry data privately. In *Proceedings of Advances in Neural Information Processing Systems*, 2017.

[23] J. C. Duchi, M. I. Jordan, and M. J. Wainwright. Minimax optimal procedures for locally private estimation. Journal of the American Statistical Association, 113(521):182–201, 2018.

[24] D. Peteiro-Barral and B. Guijarro-Berdinas, “A survey of methods for distributed machine learning,” in Progress in Artificial Intelligence, November 2012.

[25]. Lu, H., Li, M.-J., He, T., Wang, S., Narayanan, V., & Chan, K. S. (2020). *Robust Coreset Construction for Distributed Machine Learning. IEEE Journal on Selected Areas in Communications, 1–1.*

[26] D. Froelicher, J. R. Troncoso-Pastoriza, A. Pyrgelis, S. Sav, J. S. Sousa, J.-P. Bossuat, and J.-P. Hubaux. Scalable privacy-preserving distributed learning. CoRR, abs/2005.09532, 2020.