SCHEMES TO DETECT DATA POISON UNDER DISTRIBUTED ENVIRONMENT

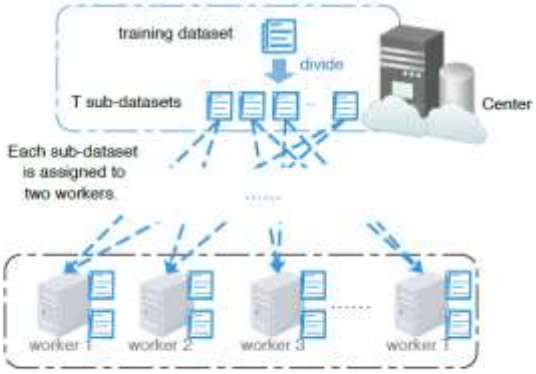
Dr. BANDI ASHA LATHA Professor, Department of CSE, SRK Institute of Technology, Vijayawada, A.P., India.

GAMPA RADHA KRISHNA, CHENNAMSETTY HEMANTH SAI, HUPATHI RAMYA RAJU, REPALLE BHARATH CHAITANYA Student, Department of CSE, SRK Institute of Technology, Vijayawada, A.P., India.

ABSTRACT:IN THIS PAPER we proposed cross-learning mechanism would generate training loops, based on which a mathematical model is established to find the optimal number of training loops. Then, for semi-DML, we present an improved data poison detection scheme to provide better learning protection with the aid of the central resource. To efficiently utilize the system resources, an optimal resource allocation approach is developed. Simulation results show that the proposed scheme can significantly improve the accuracy of the final model by up to 20% for support vector machine and 60% for logistic regression in the basic-DML scenario. Moreover, in the semi-DML scenario, the improved data poison detection scheme with optimal resource allocation can decrease the wasted resources for 20-100%.

INTRODUCTION

Distributed machine learning (DML) has been widely used in distributed systems , where no single node can get the intelligent decision from a massive dataset within an acceptable time. In a typical DML system , a central server has a tremendous amount of data at its disposal. It divides the dataset into different parts and disseminates them to distributed workers who perform the training tasks and return their results to the center. Finally, the center integrates these results and outputs the eventual model.Unfortunately, with the number of distributed workers increasing, it is hard to guarantee the security of each worker. This lack of security will increase the danger that attackers poison the dataset and manipulate the training result. Poisoning attack is a typical way to tamper the training data in machine learning. Especially in scenarios that newly generated datasets should be periodically sent to the distributed workers for updating the decision model, the attacker will have more chances to poison the datasets, leading to a more severe threat in DML.



LITERATURE SURVEY

Dalvi et al. [1] initially demonstrated that attackers could manipulate the data to defeat the data miner if they have complete information.

Lowd et al. [2] claimed that the perfect information assumption is unrealistic, and proved the attackers can construct attacks with part of the information. Afterwards, a series of works were conducted , focusing on non-distributed machine learning context. Recently, there are a couple of efforts devoted in preventing data from being manipulated in DML.

Zhang et al. [3] and Esposito et al. used game theory to design a secure algorithm for distributed support vector machine (DSVM) and collaborative deep learning, respectively. However, these schemes are designed for specific DML algorithm and cannot be used in general DML situations. Since the adversarial attack can mislead various machine learning algorithms, a widely applicable DML protection mechanism is urgent to be studied.

PROPOSEDSYSTEM

In this paper, we classify DML into basic distributed machine learning (basic-DML) and semi distributed machine learning (semi-DML), depending on whether the center shares resources in the dataset training tasks. Then, we present data poison detection schemes for basic-DML and semi-DML respectively. The experimental results validate the effect of our proposed schemes. We summary the main contributions of this paper as follows.

We put forward a data poison detection scheme for basic-DML, based on a so-called cross-learning data assignment mechanism. We prove that the cross-learning mechanism would consequently generate training loops, and provide a mathematical model to find the optimal number of training loops which has the highest security.

• We present a practical method to identify abnormal training results, which can be used to find out the poisoned datasets at a reasonable cost.

• For semi-DML, we propose an improved data poison detection scheme, which can provide better learning protection. To efficiently utilize the system resources, an optimal resource allocation scheme is developed.

IMPLEMENTATION

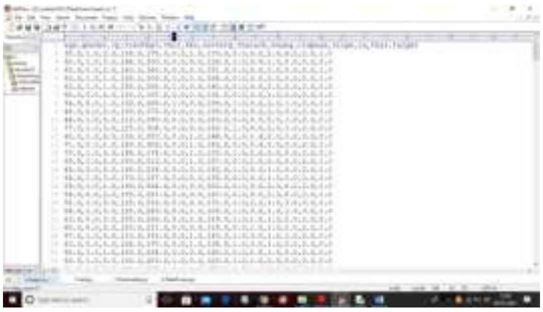
- Unfortunately, with the number of distributed workers increasing, it is hard to guarantee the security of each worker. This lack of security will increase the danger that attackers poison the dataset and manipulate the training result. Poisoning attack is a typical way to tamper the training data in machine learning. Especially in scenarios that newly generated datasets should be periodically sent to the distributed workers for updating the decision model, the attacker will have more chances to poison the datasets, leading to a more severe threat in DML.

- “Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In the SVM algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding thehyper-plane that differentiates the two classes very well
- DML into basic distributed machine learning (basic-DML) and semi distributed machine learning (semi-DML), depending on whether the center shares resources in the dataset training tasks. Then, we present data poison detection schemes for basic-DML and semi-DML respectively. The experimental results validate the effect of our proposed schemes.

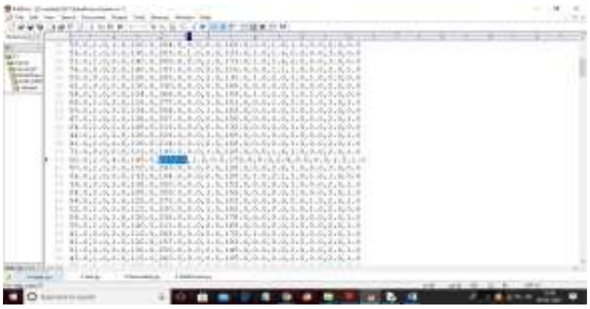
- We classify DML into basic-DML and semi-DML, which are shown in Fig.1, respectively. Both of the two scenarios have a center, which contains a database, a computing server, and a parameter server. However, the center provides different functions in these two scenarios. In the basic-DML scenario, the center has no spare computing resource for sub-dataset training, and will send all the sub-datasets to the distributed workers. Therefore, in the basic-DML, the center only integrates the training results from distributed workers by the parameter server.

SAMPLE SCREENS

To implement this project we have used heart disease dataset and in below dataset screen we can see dataset contains invalid data which called as Data Poision.



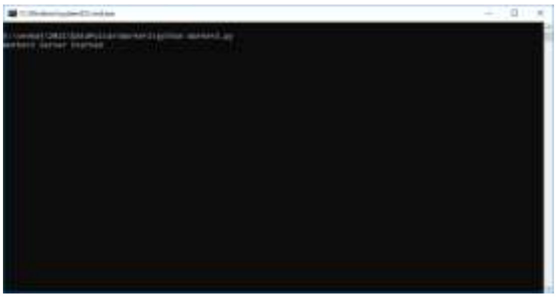
In above screen heart dataset first row contains column names and remaining rows are the column values and in below dataset screen we can see odd or invalid value



In above screen in selected blue value we can see recorded blood pressure value as 2233 which is wrong value and if ML train on such data then it may predict wrong result and it will reduce prediction accuracy and to avoid such problem we can apply Data Poison Detection technique. In python we can 'IsolationForest' class to detect and remove such poison data.

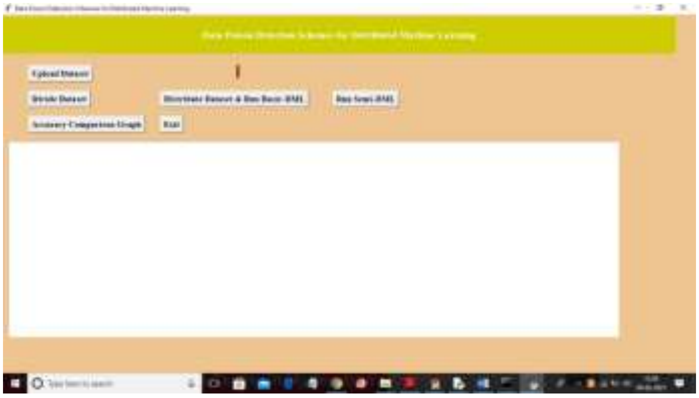
To run project first double click on 'run.bat' file from Worker1 folder to start worker 1 node and to get below screen

In above screen worker 1 server started and now double click on 'run.bat' file from worker2 folder to start worker 2



```
worker2 Server Started
```

In above screen worker2 server started and now double click on ‘run.bat’ file from ‘CenterServer’ folder to start distributed server and to get below screen



Data Poison Detection Schemes for Distributed Machine Learning
Upload Dataset
Divide Dataset
Distribute Dataset & Run Basic-DML
Run Semi-DML
Accuracy Comparison Graph
Exit

In above screen click on ‘Upload Dataset’ button to upload dataset and to get below screen

CONCLUSION

In this project, we classify DML into basic-DML and semi-DML. In basic-DML, the center server dispatches learning tasks to distributed machines and aggregates their learning results. While in semi-DML, the center server further devotes resources into dataset learning in addition to its duty in basic-DML. We firstly put forward a novel data poison detection scheme for basic-DML, which utilizes a cross-learning mechanism to find out the poisoned data. We prove that the proposed cross-learning mechanism would generate training loops, based on which a mathematical model is established to find the optimal number of training loops. Then, for semi-DML, we present an improved data poison detection scheme to provide better learning protection with the aid of the central resource. To efficiently utilize the system resources, an optimal resource allocation approach is developed. Simulation results show that the proposed scheme can significantly improve the accuracy of the final model by up to 20% for support vector machine and 60% for logistic regression in the basic-DML scenario. Moreover, in the semi-DML scenario, the improved data poison detection scheme with optimal resource allocation can decrease the wasted resources for 20-100%.

REFERENCES

[1] G. Qiao, S. Leng, K. Zhang, and Y.He, "Collaborative task offloading in vehicular edge multi-access networks," IEEE Communications Magazine, vol. 56, no. 8, pp. 48–54, 2018.

[2] K. Zhang, S. Leng, X. Peng, L. Pan, S. Maharjan, and Y. Zhang, "Artificial intelligence inspired transmission scheduling in cognitive vehicular communications and networks," IEEE Internet of Things Journal, vol. 6, no. 2, pp. 1987–1997, 2019.

[3] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "Tensorflow: A system for large-scale machine learning." in 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI), vol. 16. USENIX Association, 2016, pp. 265–283.

[4] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang, "Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems," CoRR, vol. abs/1512.01274, 2015.

[5] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges," Neurocomputing, vol. 237, pp. 350– 361, 2017.

[6] S. Yu, M. Liu, W. Dou, X. Liu, and S. Zhou, "Networking for big data: A survey," IEEE Communications Surveys & Tutorials, vol. 19, no. 1, pp. 531–549, 2017.

[7] M. Li, D. G. Andersen, J. W. Park, A. J. Smola, A. Ahmed, V. Josifovski, J. Long, E. J. Shekita, and B.-Y. Su, "Scaling distributed machine learning with the parameter server." in 11th USENIX Symposium on Operating Systems Design and Implementation (OSDI), vol. 14. USENIX Association, 2014, pp. 583–598.

[8] B. Fan, S. Leng, and K. Yang, "A dynamic bandwidth allocation algorithm in mobile networks with big data of users and networks," IEEE Network, vol. 30, no. 1, pp. 6–10, 2016.

[9] Y. Zhang, R. Yu, S. Xie, W. Yao, Y. Xiao, and M. Guizani, "Home m2m networks: Architectures, standards, and qos improvement," IEEE Communications Magazine, vol. 49, no. 4, pp. 44–52, 2021.