Ingénierie des Systèmes d'Information
Vol. 26, No. 6, December, 2021, pp. 515-522
Journal homepage: http://iieta.org/journals/isi

An Intelligent Approach for Protecting Privacy in Distributed Information Mining Using Secured Computation of Multiple Participating Sites

Vibhor Sharma[1], Shashi Bhushan[2*], Bhim Singh Boahar[3], Pramod Kumar[4], Anuj Kumar[4]

[1] HSST, Swami Rama Himalayan University, Dehradun 248001, India
[2] School of Computer Science, University of Petroleum and Energy Studies, Dehradun 248001, India
[3] Department of Computer Science, BlueCrest University College, GA-184, West Africa
[4] Department of Computer Science & Engineering, Krishna Engineering College, Ghaziabad 110093, India

Corresponding Author Email: sbhushan@ddn.upes.ac.in

https://doi.org/10.18280/isi.260601

Received: 23 September 2021
Accepted: 19 November 2021

Keywords:
privacy-preserving information mining, distributed data, multiparty computation, secret sharing

ABSTRACT

Information mining is a very well-known task using which managers can take a better decision regarding data operations. For that purpose, they need to get useful information from a large amount of raw data. This kind of big data mining is usually carried out on unstructured data that is huge in context of its size. Due to the vast size, the data mining process faces confidentiality and security breaches issues. There are many technologies through which we can search for useful information for getting fruitful results towards the fulfilment of organizational goals. However, it's a big challenge for researchers to get knowledge from the large amount of data that is owned by multiple parties which are located at different sites. For that, they need to perform Distributed Data Mining task with the concern of privacy leakage. While secure multiparty computation presents the solution to this problem but there are some issues which are untouched yet. In this paper, we presented two data privacy issues not even solved by multiparty computation. The presented algorithms are designed using the process of matrix computation with the help of encoding and decomposition methods. The implementation of the work is based on a secure Multi-Party Computation (MPC) protocol named SPDZ so that we can perform the sharing of data values. We analyzed the results produced by single machine and proposed design and implementation to show the similarity between both. Experimental results show the effectiveness of implemented algorithms and their implementation to preserve the privacy of distributed data in the process of distributed data mining.

1. INTRODUCTION

In the past years, we have seen the remarkable outburst in the field of data mining that was not done previously. In an organizational environment it is a matter of concern to use private data in a distributed manner. It is necessary to develop a secured infrastructure that could work efficiently for different beneficial purposes for organizations such as secured computation for multiple participating sites to obtain useful information without disclosing of confidential information of any participating site. We need some advance technologies for securing the confidentiality of data in the process on information mining. Most of the data frameworks are based on a distributed system that is spread all over the world. It becomes important for those companies which are having confidential data values in their big data. This kind of data holds electronic transactional data, online research data or private banking data related to credit or debit transactions. But in these kinds of distributed environment confidential values can be leaked due to duplicity attacks or attacks by intruders. These kinds of issues put challenges for researchers in the field of information mining [1]. Therefore, novel approaches are required for maintaining the privacy of certain data values so that secured sharing of data could take place among different organizations. There are many methods which are introduced to conserve the confidentiality of private own data of an organization [2]. But these approaches just hide the confidential information with the help of masking or erasing the original data. This can be done using different data mining functionalities such as Association rules, clustering, randomization, K-anonymity and many more. Furthermore, these privacy conserving methods may be categorized into five categories. These are randomization based methods, anonymization based methods, disruption based methods, condensation based methods and Cryptography based methods [3]. In the first two methods, large data loss takes place that is why their use is limited. The use of disruption and condensation based methods are limited due to qualification loss of the confidentiality that is at the time of reconstruction of the values how close can it be obtained to the actual value of a field. Cryptographic techniques are applied in Cryptography based methods to the confidentiality of private data values. These methods transform the original data into two other data to not disclose any kind of private information during the data mining task. Basically, the process of multi-party computation is based on cryptography methods to perform secure data mining. Using the multiparty computation process in a secured way, all participating sites used statistical function for getting results and any site does not get any confidential information of each other. If we talk about real-

world situations for data, base point and decimal point operations can be performed during MPC [4, 5]. Other operations like addition, subtraction, multiplication and division can be used with linear regression [6, 7]. Because of these benefits MPC has got attention why most of the researchers in the past years. However, there are some untouched issues which are not supported by MPC in real-world scenario.

In this paper, the main center of attention is two issues that are not supported during the multiparty computation process. These are evaluating stats of the data that is having different types and analyzing the correlation among different attributes using linear operations/regression. Furthermore, algorithms are designed which are based on matrix evaluation that depends on the optimization with encoding and decomposition methods [8]. For that purpose, the programming structure of MPC protocol named SPDZ is used. The main key points of this paper are: first we perform lower upper factorization (LU). After that the flow of data using an encoding method is implemented. At last, the suggested design and implementations are compared with previously developed state-of-art methods.

Remaining part of this paper holds different sections where section II represents different approaches towards information mining. Section III represents the related works. In section IV, problem description and methodology are presented. Section V shows the evaluated results followed by Section VI and VII related to conclusion and future work respectively.

2. DIFFERENT PRIVACY PRESERVING DATA MINING APPROACHES

Different approaches have been implemented for conserving the sensitivity of data. There are three levels where privacy conserving data mining approaches can be implemented as shown in Figure 1. First level represents the databases extracted from different sources for mutual gain. Second level shows the data mining algorithms implementation to get the useful information or patters. Third level shows the outcomes of data mining process after applying data mining algorithms that may be in the form of rules, patterns or useful information.

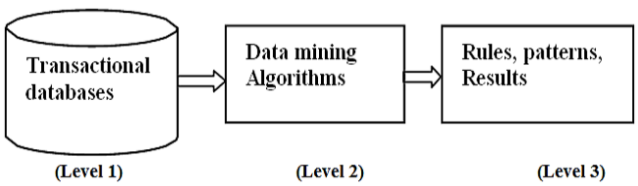

Figure 1. Data mining approaches' implementation levels for protecting the sensitivity

Following are the different dimensions on to which classification is based.

2.1 Data modification

Data Modification is the scheme of modifying the data i.e., to change the actual values of a dataset which are required to be released publicly and to ensure the confidentiality of sensitive information. Any technique for modifying the data should be adopted as per the security policy that an organization acquires. Following are different methods for modifying the data:

- Perturbation: Attribute value is changed by another value i.e., 0 to 1 or 1 to 0.
- Swapping: To interchange the values of a record as per requirement.
- Blocking: To replace the existing value of an attribute by aggregation.

Sampling: Delivery of data for a sample population.

2.2 Data distribution

Category of data mining technique used for protecting the sensitivity of the information which depends on the distribution of mined data i.e., centralized or distributed. In centralized environment, data is stored in a central location. On the other hand, in the distributed environment, data is placed in different data stores. This distribution takes place in two ways i.e., horizontal data distribution and vertical data distribution. Horizontal distribution means different records of the same attribute are stored in different places and vertical distribution means different attributes of the same record is stored in different places.

2.3 Hiding

Privacy-preserving algorithms can be further divided in to two categories: data hiding and rule hiding.

Data hiding means to hide the original data from dataset such as identity, name or address which is directly or indirectly linked to an individual's information.

Rule hiding means to remove the sensitive information (rule) inherited from the original data set when data mining technique is applied to that particular data set. Most of the algorithms which are used hide the sensitive information by modifying the data.

2.4 Data mining algorithms

Classification, clustering and association are the main tasks on to which algorithms for securing the mined data are used. Classification is a well-known process of discriminating the data into different classes to predict the class under which an object holds. Clustering is an unsupervised classification where class labels of objects are unknown. The idea is to decompose a data set in to groups called clusters. Objects having the same properties are considered into a cluster [9]. Association analysis is to derive the association rule for getting frequent patterns or item-sets.

2.5 Privacy preservation

The last approach is very much important that refers to different techniques for conserving the privacy of data coming from different sources. Such techniques are randomization method, anonymization method and encryption method. Randomization is very well-known approach that is being used for preserving the privacy of individual's personal data in the task of data mining. Using this method, noise is added to the existing data so that masking process could be implemented on individual's record. Then, reconstruction on the aggregated data distribution can be done through removing the previously added noise from the existing data. Anonymization method is utilized to conserve the privacy of sensitive information as information sanitization. This approach can be implemented by encrypting the existing personal data from the data set or

by removing it from the place where data mining will take place. Using this method, the private data will be anonymous after data mining task implementation. There are certain cases where data can be distributed over multiple places or sites for gaining mutual benefits. In these kinds of cases, there is a requirement of a kind of cryptographic protocol which can be implemented on different participating sites to conserve the secrecy of individual's personal/private data that is embedded with the whole data that is coming from a site to a central location for data mining function. Using which, a secure processing could be done on sensitive data which should not be disclosed.

3. RELATED WORK

To obtain useful information from the data that is gathered from multiple sites for decision making in the environment of the business becomes a key piece of work. To resolve the problem of privacy breach during the process of data gathering from different sites is a challenging task.

The problem of securing the competition between two sites is specialized and enlarged for consolidating it into computation among multiple sites called multi-party computation by Ankele et al. [10]. After that more efforts were made from different researchers to implement multi-party computation practically by Shukla et al. [11]. After this breakthrough, many more encryption methods were proposed for practically implementing multi-party computation by Sundari et al. [12-14]. Moreover, all these methods can be divided mainly into two types. First can be named as homomorphic encryption and second can be called as secure sharing of data. On the other hand, Nieminen [15] proposed the method of secured data sharing that is able to perform finite number of computations with more data interchange. Therefore, the comparatively first method is less effective than second method. By Hazay et al. [16], a study is made regarding the conversion of data inputted from multiple sites in common form into challenges of multi-party computation. By Giacomelli et al. [17], homomorphic encryption method was proposed to provide the solution of the environment where data is divided among multiple sites and distributed. However, this work produces complexity when the number of parameters is increased due to lack of procedure to get the product of two values which are encrypted for security purposes.

Zheng et al. [18] presented a secured protocol for securing the confidentiality of linear regression model which uses both offline and online method for initialization and processing states respectively which makes it possible to process a large amount of data for achieving privacy.

Our presented work is motivated by the works explained above to make multi-party computation more accurate and secured.

4. PROBLEM DESCRIPTION AND METHODOLOGY

First, we will explain the mentioned two untouched issues of data mining which are practical in real-world and not carried by MPC. There are some traditional ways to solve those issues by secret sharing which holds some risks. A traditional solution is defined as well as a solution for these issues. There are two issues which are solved in this paper are as discussed below:

First Issue: Evaluating stats of the data based on its types. Usually, organizations compute the indicator that represents the data stats for a defined period of time. Many companies possess market records of different commodities. Each commodity holds different codes which are based on its type. Ordered pair (code, cost) decides the cost of a commodity. All companies' market plan and future strategy for gaining benefits depend on the average of commodity's cost and the variation in cost value according to a given code of that commodity. For getting the average price of a commodity type, they need to be collected at a centralized location. These kinds of market records are highly confidential for a company. They cannot provide the actual data for security reasons. However, this problem can be solved by traditional methods of MPC named secret sharing, but there is a chance of privacy breach of centralized data.

Here, the problem is to compute the indicator value for the ordered pair (type, merit) of a commodity such as highest and lowest value, mean, mode, variation without any kind of privacy breach for both of attribute of a given ordered pair.

Second Issue: Analyzing the correlation among different attributes using a linear operation. Usually, organizations compute the indicator that represents the data stats for a defined period of time. Many companies possess market records of different commodities. Each commodity holds different codes which are based on its type. Ordered pair (code, cost) decides the cost of a commodity. All companies' market plan and future strategy for gaining benefits depend on the average of commodity's cost and the variation in cost value according to a given code of that commodity. For getting the average price of a commodity type, they need to be collected at a centralized location. These kinds of market records are highly confidential for a company. They cannot provide the actual data for security reasons. However, this problem can be solved by traditional methods of MPC named secret sharing, but there is a chance of privacy breach of centralized data.

Here, the problem is to compute the indicator value for the ordered pair (type, merit) of a commodity such as highest and lowest value, mean, mode, variation without any kind of privacy breach for both of attribute of a given ordered pair.

4.1 Traditional approach (secret sharing)

A traditional MPC based approach that can be used is secret sharing. In this approach, a secret can be shared among all the participating sites which are involved in the distributed data mining process. This secret is decomposed into different parts shared among different sites. That is, a single site cannot do with that part of the decomposed secret. But, if all the decomposed part of secretes composed off together then-secret can be obtained and privacy can be breached. Suppose, an element '*d*' is taken that belong to a set 'D'. We can divide d into multiple parts such as $d=d_1+d_2$ where $d_1$and d_2 can be given to parties P_1 and P_2 respectively. All the data values in D are presented by different elements. Mainly additive approach is used for this task of secret sharing. On the other hand, other schemes like multiplicative approach.

The additive sharing approach can be specialized in two algorithms Distribute and Rebuild among multiple participating sites discussed as Distribute and Rebuild Algorithms.

Distribute Algorithm

Input: d: Secret value
M: Total Number of participating sites.
Output: $\mathbf{N_{record}}=[n_k]_{1\le k\le M}$ (decomposed secret for each site)
Begin
Initialization
Let t=0 // t is a variable to hold temporary value
i=1
Repeat
n_k=Choose an arbitrary number (P) // P is a large prime number
t = t + n_k
Until(i≤M-1)
n_M= (d – t) Mod P
return N_{record}
End

Rebuild Algorithm

Input: $\mathbf{N_{record}}=[n_k]_{1\le k\le M}$ // Total Number of Secret records from all participating site
Output: d: Additive Value of all shared secret records
Begin
Initialization
add = 0
for each $n_k$inN_{record}**repeat**
add=add + n_k// add all the shared values
end for
d = add Mod P
return d
End

Sharing secret among participating sites for computing multiple sites contains theoretical background. A processing structure or framework is required for that purpose. SPDZ protocol is such a secured protocol to perform the mentioned task. Processing multiple sites data is executed in two steps. The first step is preprocessing into which numerical variables are defined and shared among multiple participating sites. All of this is done offline. It works like a shield against fraudulent sites or parties and enhances the conduct. The original processing is done is the second step that becomes online.

The first task i.e., to calculate the average cost of multiple types of a commodity can be done by constituting data in to a matrix and then processing it, that is based on SPDZ. But in that process, different types of data are revealed accidentally which may be confidential for companies.

Second task cannot be done with the help of actual SPDZ protocol i.e., analyzing the correlation among different attributes using linear operation. The least square method can be implemented for resolving the second issue that is not directly solved by SPDZ.

4.2 Proposed approach (a solution)

In this part, the solution of previously discussed issues is implemented:

A Solution of issue 1:

As it has been brought up previously that there is a chance of privacy leakage of data shared among multiple sites if we find out the cost of different types of a commodity individually. As a solution, we can combine the different types of a commodity to resolve the privacy breach of a particular type. Using this, we can compute the indicator for data stats by picking the individual type of data instead of sharing it among multiple participating sites so that different types of a commodity could be discriminated from a combined set of data. For that, we use an encoding method named one hot encoding that will combine all the data. This is represented as an algorithm named **WMD** as an example that is weighted mean and difference.

WMD Algorithm

Input: $Orecord = [(type, cost)]\ 1 \le i \le 2000$ // the data of a commodity taken from each site.
Output: Mean: Weighted mean of different types of a commodity
Diff: Difference
Begin
Initialization
Input type and cost with a matrix (2000 ×8) will all 0s in it initially.
for each row c_i in cost **do**
Set $ci[typei] = 1$ // 1 for each respective type
end for

for each row t_i in type **do**
Set $t_i[type_i] = cost_i$
end for
Mean = Weighted_Mean (type, cost)
repeat
$t_i\ [type_i] = (cost - Mean\ (type_i))^2$
until (each t_i in type)

Diff = Weighted_Mean (type, cost)
return Mean and Diff
End

The above process mentioned in an algorithm can be implemented by all the participating sites on their individual data. That will encode all the types of a commodity's data in to a matrix with 0s and 1s. All the data values will be considered together in a matrix and weighted mean can be calculated by column. In each row only the value that holds 1 will be considered and all the types of a commodity will be discriminated. The presented WMD algorithm calculates weighted mean and difference of different data types.

A Solution of issue 2:

As a solution to task2, we can use the least square method to solve the problem. If we talk about in the mathematical sense, the matrix can be decomposed/ factorized. There are different methods for that purpose; we selected LU factorization. In this, we will implement LU factorization. We will analyze the performance afterwards.

LU Factorization (Lower-Upper Product):

In this kind of factorization, the product of a lower triangular matrix and the upper triangular matrix is estimated. LU Factorization algorithm given below shows the implementation in the SPDZ protocol environment. This algorithm produces the values of lower triangular value and upper triangular value in recursive order in the processing structure of SPDZ. M should be full ranked to optimize the least square approach. So that it could be used in the process of data mining.

LU Factorization algorithm

Input: M = $[m_{pq}]_{1\le p\le i,\ 1\le q\le i}$ // Square matrix (i×i).
Output: L = $[l_{pq}]_{1\le p\le i,\ 1\le q\le i}$ // lower triangular matrix (i×i).

U = $[u_{pq}]_{1\le p\le i,\ 1\le q\le i}$// upper triangular matrix (i×i).

Begin

Initialization

First check M to satisfy the full_rank condition.

if M is not full_rank **then**

Terminate

else

Put all 0s to U (beneath the diagonal)

Put all 0s to L (over the diagonal) and 1s at the diagonal place

set j=1

repeat

$u_{jj} = m_{jj}$// Setting the pin point

for p = j+1 to i **do**

$l_{pj} = m_{pj} \div u_{jj}$//Calculating j^{th}column of L

$u_{jp} = m_{jp}$// Calculatinhjth rows in U

end for

for p = j+1 to i **do**

for q = j+1 to i **do**

$m_{pq} = m_{pq} - l_{pj} \times u_{jq}$//Updation on M

end for

end for

until (j ≤ i)

return L and U

End

Furthermore, LRCQ (Linear Regression of Cost and Quality) algorithm for a commodity, describes the LU factorization of linear regression of commodity's cost and quality in MPC environment.

LRCQ Algorithm

Input: O_{record}= [(MIC_i, $length_i$, $shade_i$, $cost_i$)] $_{1\le\ i\ \le 2000}$ // Commodity data accumulated from all participating sites

Output: LRO (Linear Regression Output)

Begin

Initialization

Put 1s in the first row of X (2000 × 10) and 0s in remaining rows

Put 0s in all rows of Y (2000 × 10)

define method **indicator** (i);

return the types indicator value of i in one-hot encoding

for each row x_i in X **do**

x_i [*indicator* (MIC_i)] = 1

x_i [*indicator* ($length_i$)] = 1

x_i [*indicator* ($shade_i$)] = 1

end for

Set i = 1

repeat

Y[i] = $cost_i$

until (i ≤ 2000)

L,U = LU_Factorization (X^T X)

LRO = $X^T Y L^{-1} U^{-1}$ // Linear Regression Output

return LRO

End

5. EXPERIMENTAL RESULTS

In this section, we represent the results of performed experiments on a dataset for conserving the confidentiality of distributed data among multiple participating sites.

5.1 Scenario of experiment

For the implementation of the proposed algorithm, we used Python language as per SPDZ processing structure. A virtual machine with 16 GB memory is used that is based on kernel having 6 CPU virtual cores. Cent Operating System is installed in the mentioned machine.

5.2 Dataset information and task execution

Cotton market data was taken to perform the experiment. Table 1 holds the name of attributes, their types and considered values as an example. Two jobs were executed related to presented issues for authentication and measuring the performance of the proposed work.

Table 1. Explanation of dataset

Attribute Name	Type	Considered Values
MIC	Categorical	B1 Level Value
Length	Categorical	29 Level Value
Shade	Categorical	22 Level Value
Type	Categorical	b-3128
Cost	Numerical	15260

Job 1: Calculating mean and difference of given commodity cost

The process of data mining for this job is implemented using WAD algorithm as discussed above. The focus is on calculating mean and difference to find out the mean value of the cost of different variants of a given commodity. Obtained data is represented in the form of an ordered pair (type, cost). Encoding process that is used to perform evaluation is one hot encoding processed on 8 different types of data. Cost is scaled between 10000 and 20000. Three participating sites are considered. Each of sites defines 300 rows.

Job 2: Reviewing the process of quality parameters effectiveness on commodity's cost

The process of data mining for this job is solved using LU factorization algorithm. The data is represented in the form of four attributes (*MIC, Length, Shade and Cost*). Again, three participating sites are involved and define 300 rows individually. We examine the association between each pair of quality parameters. If there is a high association between two quality parameters, we can remove one column in the assumed matrix.

5.3 Comparison of Cholesky factorization and LU factorization for computation of multiple parties

We created seven data sets to analyze the difference between the performance of LU factorization and Cholesky factorization. The processes of Cholesky and LU factorization are performed on the given dataset. Figure 2 shows the processing time results of both the processes. For all the data sizes, the processing time for Cholesky factorization increases when the size of data increased. On the other hand, the processing time of LU factorization remains almost the same. That means in comparison to Cholesky factorization, the LU factorization process is effective in the data size increasing context. So, we select LU factorization for our proposed work implementation i.e., linear operation.

5.4 Effectiveness testing

To test the effectiveness of our proposed work, we execute both the jobs on single machine as well as distributed

clustering environment for MPC.

Table 2 and 3 are showing the outcome of job 1 and job 2. The processing outcome of both single machine and distributed clustering environment for MPC are nearly equal. The differences can be avoided. The effectiveness of the proposed design and its execution can be seen from the demonstrated results in Tables.

5.5 Analyzing performance

Based on some important parameters, we examine the effectiveness of this work for given job1 and job2. These parameters are described in Table 4 and Table 5.

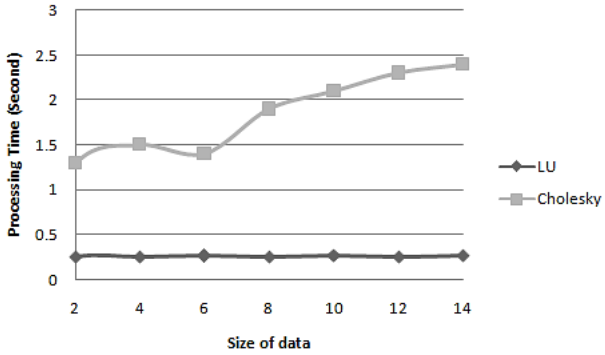

Figure 2. Processing time of LU and Cholesky factorization

Table 2. Effectiveness testing of job 1

Type	Mean	Mean of MPC	Diff.	Diff. of MPC
i.	13352.1	13352.1	84825.5	84825.5
ii.	12995.2	12995.2	73451.1	73451.1
iii.	13211.3	13211.3	77342.3	77342.3
iv.	14457.7	14457.7	39845.9	39845.9
v.	14245.2	14245.2	59871.6	59871.6
vi.	14627.1	14627.1	37658.4	37658.4
vii.	15371.1	15371.1	33810.2	33810.2
viii.	13107.2	13107.2	75463.2	75463.2

Table 3. Effectiveness testing of job 2

Type	Compound Data		Commodity Data	
	lro	lro of MPC	lro	lro of MPC
i.	0.183907	0.183907	12856.5	12856.5
ii.	11.029	11.029	539.443	539.443
iii.	18.5231	18.5231	204.23	204.23
iv.	28.8681	28.8681	-217.483	-217.483
v.	39.5991	39.5991	282.895	282.895
vi.	50.0857	50.0857	-139.975	-139.975
vii.	58.7991	58.7991	487.456	487.456
viii.	68.8058	68.8058	66.5314	66.5314

Table 4. Job 1 effectiveness parameters

Parameter	Performed Values
Processing Time	0.339s
Data Transmitted	0.498726 MB
Bytecode (Interpreted)	3.7 MB

Table 5. Job 2 effectiveness parameters

Parameter	Performed Values
Processing Time	0.873s
Data Transmitted	12.1865 MB
Byte code (Interpreted)	8.9 MB

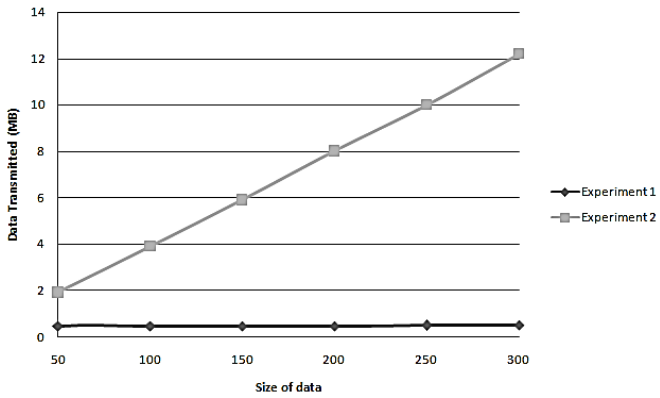

Figure 3. Transmitted Data in Job 1 and Job 2

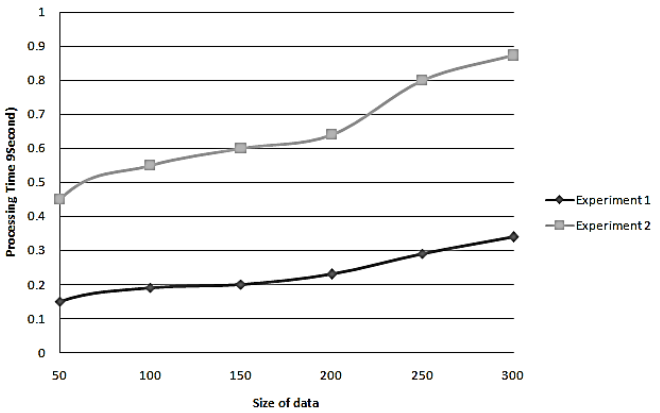

Figure 4. Processing time of job 1 and job 2

Cost of our proposed work is based on these parameters and as we can see in the tables that the given parameters indicate the admissibility of our defined cost. Because of processing time, we can state that our work is also applicable for performing this type of data mining task offline.

5.6 Effectiveness with increasing data size

To check the scalability of proposed work in case of increasing data size, five datasets are considered. For each participating site, records vary between 50 and 300. Then, we performed job 1 and job 2. Figures 3 and 4 shows the differences in transmitted data and processing time when data is being increased where processing time extension maintains admissibility that is slower than the increased amount of data in a linear way.

5.7 Performance comparability

We accomplished substantial experiments to signify the effectiveness of our proposed methodologies by comparing them with subsisting technologies. For that purpose, we selected data set of wine quality [19]. This data set is often used in many research papers which are related to multiparty computation for protecting the privacy in distributed information mining task. Two experiments were accomplished through our proposed methodology for comparison with the outcome of experiments in [19, 20]. Outcomes of experiments are shown in Table 6. The outcome of our proposed work is shown as A (with ten participating sites) and C (with three participating sites). The outcomes of [20, 21] are shown as B, C, E and F respectively.

Table 6. Performance comparison

	A (Ours)	B [17]	C [19]	D (Ours)	E [20]	F [21]
Evaluator's Total Processing Time (s)	4.06	4.09	4.08	1.96	3.05	3.01
Total Number of Participating Sites	10	10	10	3	2	2
Data Distribution	Horizontal	Horizontal	Horizontal	Horizontal	Vertical	Horizontal
Machine CPU	Six-Core	40 Core	Six-Core	Six-Core	Quad-Core	Six-Core
Machine Memory	8 GB	500 GB	8 GB	8 GB	7.5 GB	12 GB
Machine Type	Virtual Machine	Server	Virtual Machine	Virtual Machine	EC2 C4	Virtual Machine

When compared with literature [21] and literature [22] by considering ten parties, less processing time was achieved by us with the hardware of low configuration. When compared with literature [23] literature and [24, 25] in which Cholesky factorization and private regression tree generation was used respectively, our proposed work again received better performance as less processing time.

6. CONCLUSIONS

Conserving the privacy of confidential data during the data mining task in a distributed scenario is very much important and basic need of today's information mining environment. We discussed two important issues related to this task which are difficult to implement by the MPC processing structure. As a solution, we suggested an encoding method to solve the first issue and LU factorization to resolve the second issue. The solution is based on MPC-SPDZ processing framework. To test the effectiveness of the proposed work, we performed experiments on compound data and cotton market dataset. The experimental result shows the consistency between the outcomes of proposed work with results produced by a single machine framework. Performance analysis indicates that our implementation is empirical for tasks included in information mining.

7. FUTURE WORK

In future, the effectiveness can be boosted by making use of GPU i.e., Graphics Processing Unit. The dataset with large size can be mined for further improvement of the presented work. More factorization methods can be explored instead of LU factorization and Cholesky factorization.

REFERENCES

[1] Binjubeir, M., Ahmed, A.A., Ismail, M.A.B., Sadiq, A.S., Khan, M.K. (2019). Comprehensive survey on big data privacy protection. IEEE Access, 8: 20067-20079. https://doi.org/10.1109/ACCESS.2019.2962368

[2] Lin, J.C.W., Fournier-Viger, P., Wu, L., Gan, W., Djenouri, Y., Zhang, J. (2018). PPSF: An open-source privacy-preserving and security mining framework. 2018 IEEE International Conference on Data Mining Workshops (ICDMW), Singapore, pp. 1459-1463. https://doi.org/10.1109/ICDMW.2018.00208

[3] Kiran, A., Vasumathi, D. (2017). A Comprehensive survey on privacy preservation algorithms in data mining. 2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Coimbatore, India, pp. 1-7. https://doi.org/10.1109/ICCIC.2017.8524294

[4] Sharma, A.R., Kaushik, P. (2017). Literature survey of statistical, deep and reinforcement learning in natural language processing. 2017 International conference on computing, communication and automation (ICCCA), Greater Noida, India, pp. 350-354. https://doi.org/10.1109/CCAA.2017.8229841

[5] Blanton, M., Aguiar, E. (2016). Private and oblivious set and multiset operations. International Journal of Information Security, 15(5): 493-518. https://doi.org/10.1007/s10207-015-0301-1

[6] Bourse, F., Sanders, O., Traoré, J. (2020). Improved secure integer comparison via homomorphic encryption. In Cryptographers' Track at the RSA Conference, San Francisco, CA, USA, pp. 391-416. https://doi.org/10.1007/978-3-030-40186-3_17

[7] Catrina, O. (2018). Towards practical secure computation with floating-point numbers. 3rd Annual International Conference on Cryptography and Information Security, Iasi, Romania.

[8] Liu, J., Liang, Y., Ansari, N. (2016). Spark-based large-scale matrix inversion for big data processing. IEEE Access, 4: 2166-2176. https://doi.org/10.1109/ACCESS.2016.2546544

[9] Aggarwal, C.C., Philip, S.Y. (2008). A general survey of privacy-preserving data mining models and algorithms. Privacy-Preserving Data Mining, pp. 11-52. https://doi.org/10.1007/978-0-387-70992-5_2

[10] Ankele, R., Simpson, A. (2017). On the performance of a trustworthy remote entity in comparison to secure multi-party computation. 2017 IEEE Trustcom/BigDataSE/ICESS, Sydney, NSW, Australia, pp. 1115-1122. https://doi.org/10.1109/Trustcom/BigDataSE/ICESS.2017.361

[11] Shukla, S., Sadashivappa, G. (2014). Secure multi-party computation protocol using asymmetric encryption. 2014 International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, pp. 780-785. https://doi.org/10.1109/IndiaCom.2014.6828069

[12] Sundari, S., Ananthi, M. (2015). Secure multi-party computation in differential private data with Data Integrity Protection. 2015 International Conference on Computing and Communications Technologies (ICCCT), Chennai, India, pp. 180-184. https://doi.org/10.1109/ICCCT2.2015.7292742

[13] Saiyad, M. (2019). Secure multiparty computation and privacy preserving scheme using homomorphic elliptic curve cryptography. 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Madurai, India, pp. 776-780.

https://doi.org/10.1109/ICCS45141.2019.9065645
[14] Laird, P., Delany, S.J., Dondio, P. (2017). Multiparty computations in varying contexts. 2017 IEEE Conference on Dependable and Secure Computing, Taipei, Taiwan, pp. 400-407. https://doi.org/10.1109/DESEC.2017.8073825
[15] Nieminen, R., Jrvinen, K. (2020). Practical privacy-preserving indoor localization based on secure two-party computation. IEEE Transactions on Mobile Computing, 20(9): 2877-2890. https://doi.org/10.1109/TMC.2020.2990871
[16] Hazay, C., Scholl, P., & Soria-Vazquez, E. (2020). Low cost constant round MPC combining BMR and oblivious transfer. Journal of Cryptology, 33(4): 1732-1786. https://doi.org/10.1007/s00145-020-09355-y
[17] Giacomelli, I., Jha, S., Joye, M., Page, C.D., Yoon, K. (2018). Privacy-preserving ridge regression with only linearly-homomorphic encryption. International Conference on Applied Cryptography and Network Security, Leuven, Belgium, pp. 243-261. https://doi.org/10.1007/978-3-319-93387-0_13
[18] Zheng, W., Popa, R.A., Gonzalez, J.E., Stoica, I. (2019). Helen: Maliciously secure coopetitive learning for linear models. 2019 IEEE Symposium on Security and Privacy (SP), San Francisco, CA, USA, pp. 724-738. https://doi.org/10.1109/SP.2019.00045
[19] Vu, D.H., Luong, T.D., Ho, T.B. (2020). An efficient approach for secure multi-party computation without authenticated channel. Information Sciences, 527: 356-368. https://doi.org/10.1016/j.ins.2019.07.031
[20] Gascón, A., Schoppmann, P., Balle, B., Raykova, M., Doerner, J., Zahur, S., Evans, D. (2017). Privacy-preserving distributed linear regression on high-dimensional data. Proceedings on Privacy Enhancing Technologies, 2017(4): 345-364. https://dx.doi.org/10.1515/popets-2017-0053
[21] Bhushan, S., Singh, A.K., Vij, S. (2019). Comparative study and analysis of wireless mesh networks on AODV and DSR. 2019 4th International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU), Ghaziabad, India, pp. 1-6. https://doi.org/10.1109/IoT-SIU.2019.8777466
[22] Chowdhury, S.I., Lee, W.I., Choi, Y.S., Kee, G.Y., Pyun, J.Y. (2011). Performance evaluation of reactive routing protocols in VANET. The 17th Asia Pacific Conference on Communications, Sabah, Malaysia, pp. 559-564. https://doi.org/10.1109/APCC.2011.6152871
[23] Zhao, L., Ni, L., Hu, S., Chen, Y., Zhou, P., Xiao, F., Wu, L. (2018). Inprivate digging: Enabling tree-based distributed data mining with differential privacy. IEEE INFOCOM 2018-IEEE Conference on Computer Communications, Honolulu, HI, USA, pp. 2087-2095. https://doi.org/10.1109/INFOCOM.2018.8486352
[24] Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J. (2009). Wine quality data set. Decision Support Systems.
[25] Bhushan, S., Kumar, M., Kumar, P., Stephan, T., Shankar, A., Liu, P. (2021). FAJIT: A fuzzy-based data aggregation technique for energy efficiency in wireless sensor network. Complex & Intelligent Systems, 7: 997-1007. https://doi.org/10.1007/s40747-020-00258-w