New Algorithm Development of Categorized Randomization Based Privacy Preservation Data Mining

Anjali Vishwakarma
Department of Computer Science and Engineering
SIRTS Bhopal, INDIA

Dr. Ritu Shrivastava
Department of Computer Science and Engineering
SIRTS Bhopal, INDIA

Abstract- The most trending concept is of Privacy Preservation Data Mining (PPDM) which is a top research area concerned with the privacy of personally identifiable information when considered for the concept of data-mining. And this privacy is obtained through the use of the randomization, k-anonymization, and distributed privacy-preserving data mining. In order to provide better privacy, multi-level frameworks are used. The concept of PPDM has now become the most emerging domain because it allows sharing of the private data in the process of analysis. In this paper we will discuss the computational and theoretical limits associated with privacy preservation having the data-sets of high-dimension. Several techniques of PPDM have also been discussed here. This paper also gives a detail analysis of randomization based concept of PPDM along with some experimental analysis on the concept.

Keywords – **Randomization, Privacy Preservation, Data Mining,**

I. INTRODUCTION

Privacy-preserving uses the Data Mining approaches in order to protect the user's information from the intruders. The various methods involving randomization, k-anonymity model and l-diversity, distributed privacy preservation etc downgrade Application efficiency. The concept of preserving the privacy may be used in many real time applications including the surveillance, identity check etc. PPDM [1,2] is a new research direction in Data Mining and statistical databases [3], in which the Data Mining algorithms have been analyzed for the side effects they acquire in data privacy. The main consideration of suggested concept of privacy-preserving Data Mining is having a twofold concept.

On the one hand, after getting published all the private details of the user or the individuals can be revealed, that can generate various confidential and privacy related problems. And because of these problems of privacy various users or customers have become cautious while sharing their information with the public which may result in the unavailability of data. Therefore, the privacy must be considered as the most significant aspect in the domain of data-mining. And the Privacy-Preserving Data Mining concept is now becoming the most trending domain of research that may address the several problems regarding privacy. In order to face the challenging risk, some researchers have proposed a remedy which target at accomplishing the balance of data utility and information privacy when publishing dataset. This ongoing research is called Randomization dependent PPDM which is balancing the privacy of the data as per the legitimate need of the user.

The Preservation of privacy of each person's details or data is the most important function for the owners of data in order to maintain the privacy. So in this way privacy performs a vital part within the data-mining process. The process of Data Mining may enable the organization to utilize huge volume of data in order to establish the correlations and associations in between the data for enhancing the efficiency of business. Hence the Privacy-preserving Data Mining

has now become the appreciable domain of research in these days. With the help of PPDM approach the researchers can study the data without compromising privacy of any individual. Techniques of the Privacy-preserving Data Mining are clearly dependent on the definition of privacy, which captures what information is sensitive in the original data and should therefore be protected from either direct or indirect disclosure.

Privacy can also be preserved by decision tree learning on unrealized data sets [4]. And the analytical features of the randomization process which is high dimensional can also be studied for examining strengths and weaknesses of randomization. Data privacy is also preserved by using randomization with encryption method. In this paper, we provide a classification and description of the various techniques and methodologies that have been developed in the area of PPDM.

The privacy-preserving concept may permit sharing of sensitive data for a detailed examination of elements purposes so it is very popular technique. So people are ready to share their data. It is used for protecting the privacy of the risky and sensitive data of data mining .In Randomization Response Technique random noise is added to the original data to preserve privacy. The main goal of privacy preserving DM is to develop algorithms for modifying the original data and securing the information, so that the private data and private knowledge remain as it is after mining process.

II. PPDM

Privacy preserving Data Mining is going to be achieved in different ways specifically by using randomization methods, cryptography algorithms and anonymization methods. A modern survey is being done on various methods using PPDM and major PPDM techniques are reviewed based on merits & demerits. The current scenario privacy preserving Data Mining [2] propose some future research directions for research people.
The aim of privacy preserving Data Mining is to develop data mining methods without increasing the loss of data. The topic of privacy preserving Data Mining has been of great importance in Data Mining community in recent years. A number of effective methods for privacy preserving Data Mining are proposed earlier. Most methods use some form of transformation on the original data in order to perform the privacy preservation [5].

Privacy preserving Data Mining can be achieved in various ways by use of randomization techniques, cryptographic algorithms, anonymization methods etc. A recent survey on some of the techniques used for privacy-preserving Data Mining may be found in [1] which reviews main PPDM techniques based on a PPDM framework and compare the advantages and disadvantages of different PPDM technique and discuss the open issues and future research trends in PPDM.

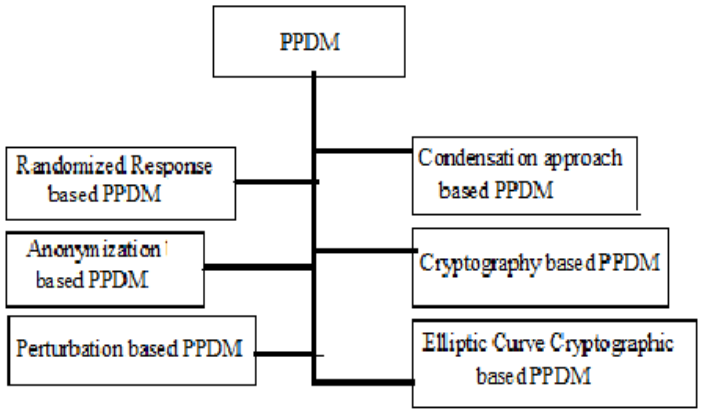

Fig.1 Techniques of PPDM

Many techniques have recently been proposed for privacy preserving Data Mining of multidimensional data set. Some techniques of the privacy-preserving Data Mining are discussed in this paper along with their merits and demerits by analyzing different approaches like the l-diversity is compared with the k- anonymity[2]

The concept of Privacy-preservation is taken as the one of the significant aspect for efficiently using excessive amount of the data. So this type of data also has been getting recorded within the electronic form, and this storing does not disturb any other user or data. It is also maintain the privacy of data while gathering and the mining process of data. The main reason of the PPDM is introduce some algorithms for modifying the original data in the similar manner, such that personal data or the private information will keep private even after the mining process. In this paper, we provide a classification and description of the various techniques and methodologies.

To solve the problem there are various techniques that are going to be presented in this paper and with each technique drawbacks and benefits are discussed. The primary aim of this paper is to grasp the existing privacy preserving data mining techniques to achieve efficiency. Data mining is the technique of analyzing the data set from different perspectives and get the useful information and specially to discover the knowledge is the ultimate aim of the data mining technique.

III. RANDOMIZATION

It is relatively simple and does not require knowledge of the distribution of other records in the data the randomization method can be implemented at data collection time itself. Since it treats all records equally irrespective of their local density, outlier records are more susceptible to adversarial attacks. The quantity used to measure privacy should indicate how closely the original value of an attribute can be estimated.
This technique of Randomization is also one of the cheapest and effective techniques used to secure the privacy of each user [7].

One key advantage of the randomization method is that it is relatively simple, and does not require knowledge of the distribution of other records in the data. Therefore, the randomization method can be implemented at data collection time, and does not require the use of a trusted server containing all the original records in order to perform the anonymization process.

In the available techniques of privacy-preserving DM, the concept of randomization is treated as the very significant approach. And this technique may offer the process of knowledge-discovery also provide the balance in between the utility and privacy [1]. Also balance in between them has been obtained through merging noise within the data. This technique of randomized by balancing data will be sent towards the recipient device.

And recipient device may obtain the data by the use of an approach of distribution-reconstruction-algorithm. In this method, noise is fused in the data at data collection time. It creates private representations of the records using different data distortion methods. The randomization method is easily implemented at data collection time, because the added noise is independent of the behavior of other data records.

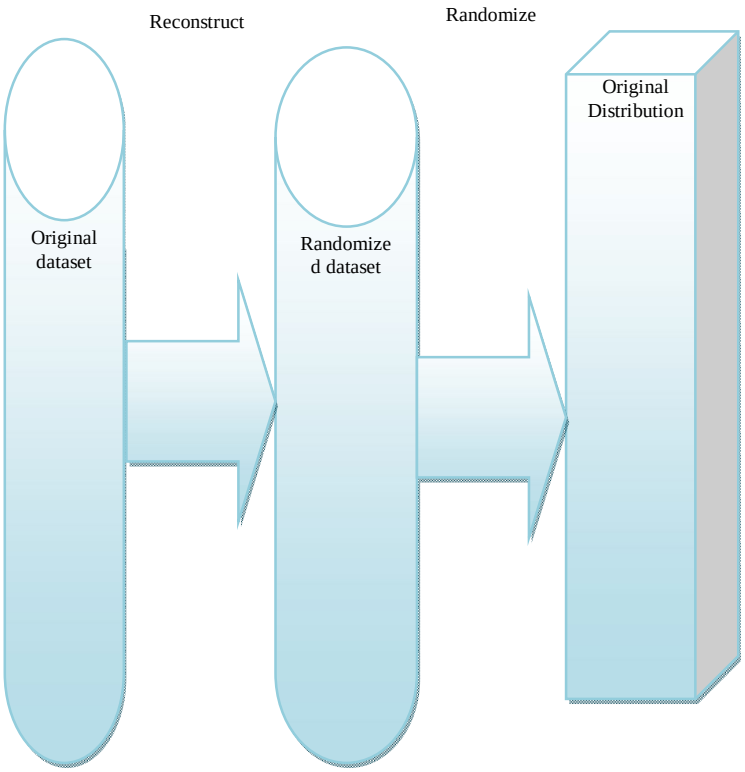

Fig.2 Model of Randomization

IV. RANDOMIZATION IN PPDM

The randomization approach is particularly well suited to privacy-preserving Data Mining of streams, since the noise added to a given record is independent of the rest of the data. However, streams provide a particularly vulnerable target for adversarial attacks with the use of PCA techniques because of the large volume of the data available for analysis. Many interesting techniques for randomization has been proposed which uses the auto correlation in different time series while deciding the noise to be added to any particular value [8].
The randomization method has been enhanced for the different types of Data Mining issues. A number of other techniques have also been proposed which seem to work well over a variety of different classifiers. Techniques have also been suggested for offering the approaches of privacy preserving in order to improve the effectiveness of classifiers. The problem of association rules is especially challenging because of the discrete nature of the attributes corresponding to presence or absence of items. The randomization method has been traditionally used in the context of distorting data by probability distribution for methods such as surveys which have an evasive answer bias because of privacy concerns [5].

Main idea for the research of privacy-preserving Data Mining is to develop these types of solutions that may provide the security of data along with data consistency and confidentiality with low computational complexity. So the technique of Randomization is the cheapest and also the effective method that is used for offering the privacy-preserving-Data Mining. In order to assure the performance [7] of the Data Mining and also in order to preserve the privacy of each user by this type of randomization approach this is required to be applied.

By this randomization mechanism we secure the data of user through offering them some random changes within their data just before transmitting those data and also consider few relevant details and also generating few noises in network. So a number of approaches within the randomization are the type of numerical-randomization and also the item-set randomization type of noise may be generated through adding or by multiplying the random values with the numerical data or through removing the actual items and also merging few "fake" values within the attributes set.

Randomization has been first suggested in numerical data for the data distribution of X. And the Value distortion is also performed over the data record xi through adding to it the random value of r from some predefined set of random distribution like R. In order to resolve the current trend issue there are several approaches like the algorithm which is going to be represented within this paper along with its separate technique demerits and merits. The basic purpose of this paper is to grasp the existing privacy preserving data mining techniques to achieve efficiency. Data mining is the technique of analyzing the data set from different perspectives and get the useful information and specially to discover the knowledge is the ultimate aim of the data mining technique.

V. LITERATURE REVIEW

Jiang, J. and Umano, M [9], emphasized the risk of disclosing the personal information during the data-mining process.. Hence, they suggested an approach for deriving the fuzzy-rules by the shared data along with similar type of attributes in the format of privacy-preserving. And this type of approach can deliver only the required values for the process of extraction by not gathering any type of data in single place and also it may achieve the global-fuzzy-rules in every place.

Kasugai H, Kawano A, Honda, K, Notsu A. [10] describes the essential techniques used within the privacy-preserving-Data Mining which is a type of priori data-anonymization, where every record is anonymous such that any of the records may not get related with any specific user. And for the efficient type of data anonymization, the approaches of clustering has also been implemented. Earlier, it has been represented that the fuzzy-clustering technique may obtain the data-anonymization process by not having any major loss in information as it efficiently combine the same types of records within the clusters in which every data is easily separated after performing the cluster merging process.This implementation is based on the fuzzy-k-member process of clustering into the privacy-preserving method. The main objective is to conduct the supervised method of recognition which maintains anonymization.

D.Karthikeswarant, V.M.Sudha, V.M.Suresh and A.J. Sultan [11] discussed various type of needs of the sharing of data, discovery of knowledge, privacy-preserving and also the PPDM all these have been the interesting topics of research in Data Mining and also in the domains of security of database. And the researchers may have also made various efforts for hiding the association-rules that have been generated. So this paper presented an ideal approach which can strategically updates some transactions within the database. Also it changes the support-value or the confidence-values in order to hide the sensitive-rules by not causing various side effects. Moreover, the unwanted type of side effects are like non-sensitive rules get hidden and also generate the fake-rules falsely, during the process of rule hiding.

Maiwand Khishki and Vijay Kumar [12] presented that without modifying the original algorithms for association rule mining it is possible to get accurate results. The privacy breach level calculated is high for Uniform randomization with single level. With multiple levels and in weighted randomization methods, the privacy breach level reaches an acceptable level of 50% or less. In this paper, the major contributions are; to show that multiple levels of privacy can be implemented without lowering the accuracy greatly. Multiple levels also decreased the privacy breach level slightly (Increased privacy). Weighted randomization is based on all the transactions that have been encountered. In this method, the privacy breach level falls further below the uniform randomization methods.

Li, Yaping, Chen, Minghua [13] enhances the assumption also it expands range of the perturbation-dependent privacy-preserving-Data Mining to the Multi-level-Trust approach. Within this simulation work, very trusted component is the data-miner and the very less perturbed copy of data which it may get used. Within this type of

setting, the malignant type of data-miner can have the access right over the various perturbed copies of similar data for different purposed and also merge these types of copies in order to infer the information regarding the actual data which the owner of data don't want to disclose. And so preventing this type of attacks is major issue of offering the services of MLT-PPDM. So this paper refers this issue through associating the perturbation with the copies at all the levels of trust. And also proves that this solution is more powerful for the attacks.

Pathak, F.A.N. ,Pandey [14], described about the innovative type of protocol. And this protocol can use the actual and the idyllic type of models. Through the use of these models we can provide more level of privacy and also security can be offered. For this the blocks of data get divided into the different segments and also again the segments get distributed within the parties. And the basic concept is the value got computed through the party which is contributing within the protocol and the computation is dependent over only given input and the generated. So in this condition it is not possible to get the personal data of other users.

Charu C. Aggarwal [15], describes the initial analysis on the randomization of high-dimensional data. And the objective is to determine the power and the shortcomings of the process of randomization and also represent the efficiencies and pitfalls of this process. The theoretical observation may result in various types of attractive and useful conclusions like the effects of privacy over the randomization may decrease with the rise of dimensionality. Features of given data-set may influence the level of anonymity of this approach and public information uses in the attack offers the selection of the perturbing distribution to be very complicated as compared to the earlier reasons. Particularly, the approach of Gaussian perturbations can be more efficient as compare to the equally distributed type of perturbations. And these observations are essential for the future work and also for the designing of randomization approach.

VI. PROPOSED WORK

This section deals with the proposed work. This section mainly classified the proposed work into two parts. In the first part, it describes about the proposed work architecture. This proposed architecture is shown in figure 1. This figure shows clearly that after reading the dataset, the first thing is that we divide the data into quasi and sensitive attribute. Then we perturb quasi attribute with the Root Mean Square value of previous column .Then we combine data of quasi and sensitive attribute then choosing or picking the any random column. Now Divide column in three categories say 1, 2, 3. After the division of the column, sort whole data according to column. Now select random row. Finally, Dived row in three categories say 1, 2, 3 and Sort whole data according to row. Now Randomization has been applied on dataset.

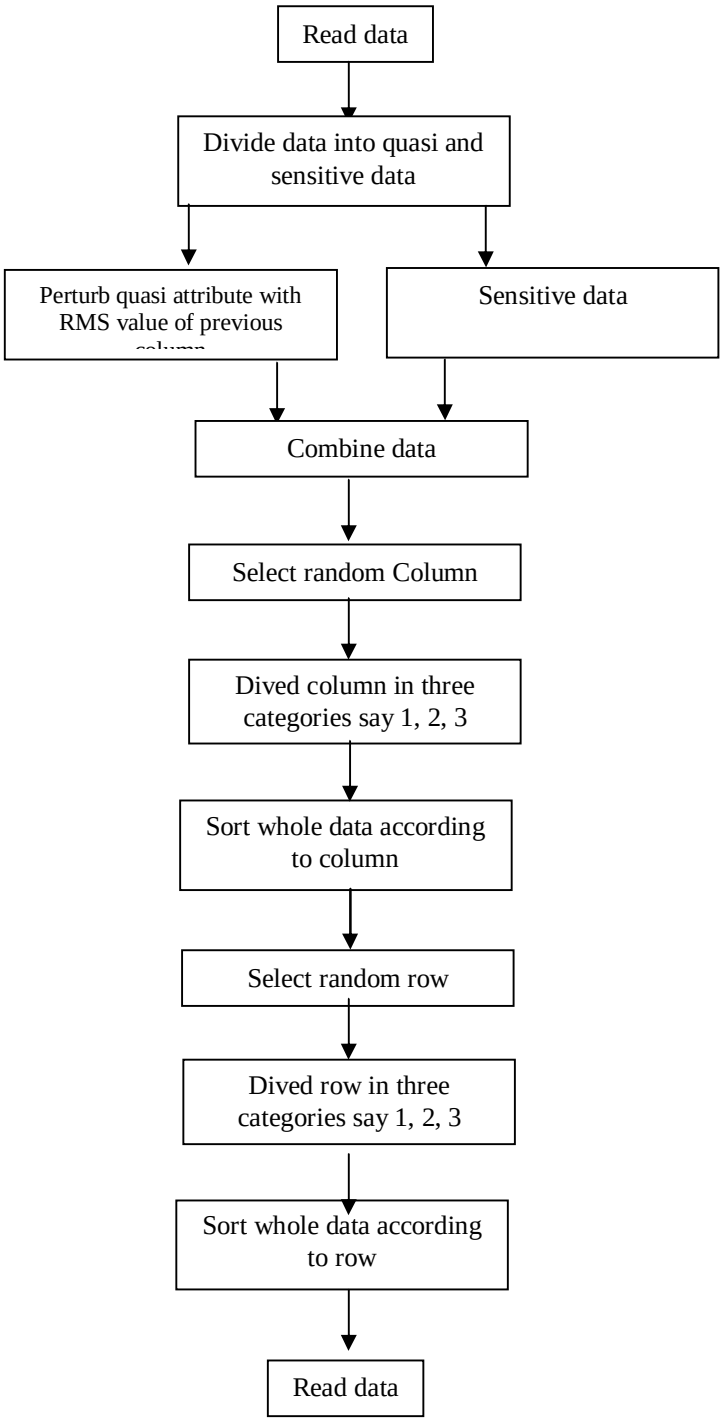

Figure 1: Architecture of proposed work Proposed Algorithm:

//read dataset in t and randomize data in t find quasi attribute in data let qi col be no of quasi attribute in data

1. RMS=0;
2. For j=1:qicol
3. For i=1:size(t,1)
4. RMS(j)=RMS+t(i,j)^2;

```
5.  End for
6.  End for
7.  For i=1:qicol
8.     RMS(i)=RMS(i)^0.5;
9.  End for
10. For j=1:(qicol-1)
11. for i=1:size(t,1)
12.    t(i,j)=RMS(j+1)+t(i,j);
13.  end
14. //Select random column j
15. Rangemin= min(t(j),:));
16. Rangemax= max(t(j),:));
17. // decied point of dicision
18. pointOne=Rangemin+(Rangemax+Rangemin)/3;
19. pointTwo=Rangemin+2*(Rangemax+Rangemin)/3;
20. // coding column order
21. for i=1:size(t,2)
22.    if (t(m(1),i)<pointOne)
23. col_oder(i)=1;
24. elseif(t(m(1),i)<pointTwo)
25. col_oder(i)=2;
26.    else
27. col_oder(i)=3;
28.    end
29. End
30. // select random row k
31. m = randperm(nCol);
32. Rangemin= min(t(i,k));
33. Rangemax= max(t(i,k));
34. pointOne=Rangemin+(Rangemax+Rangemin)/3;
35. pointTwo=Rangemin+2*(Rangemax+Rangemin)/3;
36. for i=1:size(t,1)
37.    if (t(i,m(1))<pointOne)
38. row_oder(i)=1;
39. elseif(t(i,m(1))<pointTwo)
40. row_oder(i)=2;
41.    else
42. row_oder(i)=3;
43.    end
44. End
45. Sorting data with sorted Colum oder and row oder
46. [~,sortedColOder]=sort(col_oder);
47. [~,sortedRowOder]=sort(row_oder);
48. t=t(:,sortedColOder);
49. t=t(sortedRowOder,:);
```

VII. RESULT ANALYSIS

This section deals with following points in order to discuss the effect of the proposed work. These are as follows:

1. Dataset

2. System and
3. Result Parameters

Dataset:

Dataset is taken from UCI website. This dataset is Hepatitis. There are total 155 no. of records. Total number of attributes are 20 (including the class attribute).

Attribute Information:

1. Class: DIE, LIVE
2. AGE: 10, 20, 30, 40, 50, 60, 70, 80
3. SEX: male, female
4. STEROID: no, yes
5. ANTIVIRALS: no, yes
6. FATIGUE: no, yes
7. MALAISE: no, yes
8. ANOREXIA: no, yes
9. LIVER BIG: no, yes
10. LIVER FIRM: no, yes
11. SPLEEN PALPABLE: no, yes
12. SPIDERS: no, yes
13. ASCITES: no, yes
14. VARICES: no, yes
15. BILIRUBIN: 0.39, 0.80, 1.20, 2.00, 3.00, 4.00
16. ALK PHOSPHATE: 33, 80, 120, 160, 200, 250
17. SGOT: 13, 100, 200, 300, 400, 500,
18. ALBUMIN: 2.1, 3.0, 3.8, 4.5, 5.0, 6.0
19. PROTIME: 10, 20, 30, 40, 50, 60, 70, 80, 90
20. HISTOLOGY: no, yes

System:

All the experiments are performed on Dual code system with 4 GB RAM. The operating system is Windows 7, 32 bits on which all of these experiments are performed.

Result Parameters:

This section considered two parameters for analysis of the performance of the proposed work over existing work.

a. Privacy
b. Execution Time

Privacy Level:

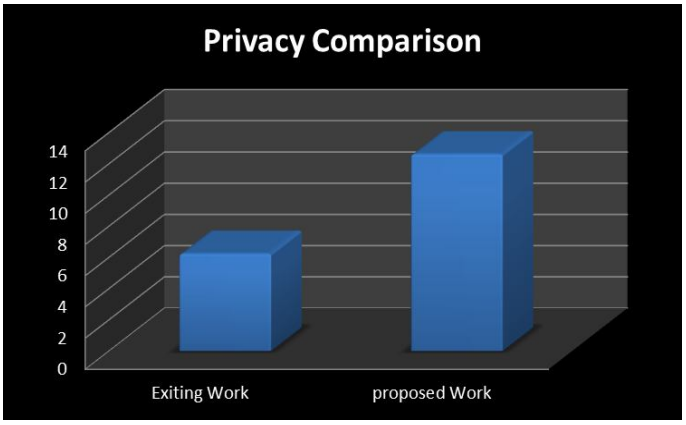

Figure 4: Privacy level of the existing work and proposed work.

This privacy level is showing in table I and its visual impact is shown in figure 4.

Table I: Privacy level of existing and proposed work.

Existing Work	Proposed Work
6.233	12.61

Execution time:

This execution time is showing in table II and its visual impact is shown in figure 5. Experiment is performing 10 times and the value shown here is average of that.

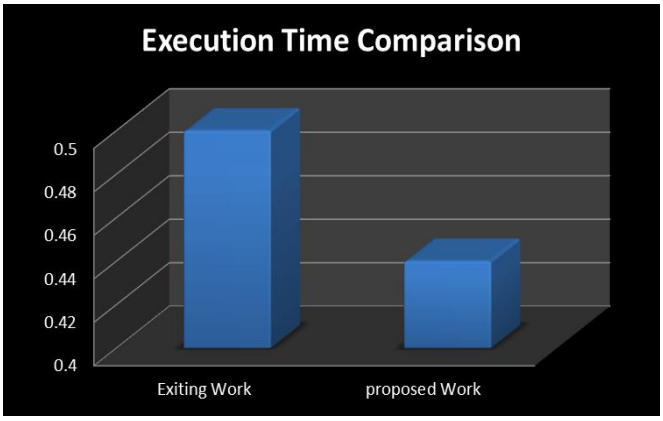

Figure 5: Execution time of the existing work and proposed work.

Table II: Execution Time of existing and proposed work.

Existing Work	Proposed Work
0.5	0.44

VIII. CONCLUSION

This paper suggested a categorized randomization PPDM method and a simplified taxonomy to help understand the problem. Other section of the paper talks about several approaches of the PPDM for managing the various issues regarding the privacy within the process of data-mining. Also this paper targeted over the concept of data mining, randomization along with the application of randomization within PPDM with some privacy concerns for the user which stores data and shares data with others. The authors have mainly detailed about randomization technique which protects privacy and briefly define major data mining task such as clustering, classification, association-rule mining, etc. the figure 4 and 5 along with the table i & ii clearly show that the performance of the proposed work as compare to existing work is more efficient.

REFERENCES

[1] Alpa K. Shah, Ravi Gulati, "Contemporary Trends in Privacy Preserving Collaborative Data Mining– A Survey", Proceedings in IEEE International Conference on Electrical, Electronics, Signals, Communication and Optimization (EESCO), 2015

[2] Malik M.B, Ghazi M.A, Ali R. 2012, Privacy Preserving DM Techniques: Current Scenario and Future Prospects, Third International Conference on Computer and Communication Technology (ICCCT), pp: 26 – 32.

[3] Kokkinos, Y., Margaritis, K., 2013, Distributed privacy-preserving P2P DM via probabilistic neural network committee machines, Fourth International Conference on Information, Intelligence, Systems and Applications (IISA), 2013, pp: 1-4

[4] Shrivastava A., Dutta U.: "An Emblematic Study of Different Techniques in PPDM". International Journal of Advanced Research in Computer science and Software Engineering (IJARCSSE), Vol.3, Issue.8, pp.443-447, 2013.

[5] Sachinjanbandhu, Dr. S.M. Chaware "Survey on DM with privacy preservation" International Journal on Computer Science & Information Technology Vol-5(4) 2014.

[6] Dhanalakshmi.M, Siva Sankari.E, "Privacy Preserving DM Techniques-Survey", IEEE, Information Communication and Embedded Systems (ICICES), 2014.

[7] Mohnish Patel, PrashantRichariya, AnuragShrivastava, (2013),,,,,A review paper on Privacy-Preserving Data Mining", Review article on Scholars Journal of Engineering and Technology (SJET) , pp.359-361

[8] A.S.Shanthi, , Dr. M. Karthikeyan" A Review on Privacy Preserving DM "IEEE International Conference on Computational Intelligence and Computing, 2012 .

[9] Jiang, J. and Umano, M. 2014, Privacy preserving extraction of fuzzy rules from distributed data with different attributes, Joint 7th International Conference on and Advanced Intelligent Systems (ISIS), 15th International Symposium on Soft Computing and Intelligent Systems (SCIS), 2014, pp : 1180-1185.

[10] Kasugai H, Kawano A, Honda, K, Notsu A. 2013, A study on applicability of fuzzy k-member clustering to privacy preserving pattern recognition, IEEE International Conference on Fuzzy Systems (FUZZ), 2013, pp:1-6 .

[11] D.Karthikeswarant, V.M.Sudha, V.M.Suresh and A.J. Sultan, "A Pattern based framework for privacy preservation through Association rule Mining" in proceedings of International Conference On Advances In Engineering, Science And Management (ICAESM -2012), IEEE 2012.

[12] MaiwandKhishki and Vijay Kumar, " Research Paper on Randomization-based Privacy-Preserving Association Rule Mining", IJARCSSE,Volume 5, Issue 6, June 2015 ISSN: 2277 128X.

[13] Li, Yaping, Chen, Minghua ; Li, Qiwei ; Zhang, Wei, 2012, Enabling Multilevel Trust in Privacy Preserving Data Mining, Knowledge and Data Engineering, IEEE Transactions on (Volume:24, Issue: 9), pp: 1598 – 1612

[14] Pathak, F.A.N. ,Pandey, S.B.S., 2013, An efficient method for privacy preserving DM in secure multiparty computation, Nirma University International Conference on Engineering (NUiCONE), 2013, pp: 1 – 3 Pathak, F.A.N. , Pandey, S.B.S., 2013

[15] Charu C. Aggarwal "On the Analytical Properties of High Dimensional Randomization." IEEE Transaction on KnowledgeAnd Data engineering Vol. 25, Issue-7 July 2013.