ISSN:0975-9646

Privacy Preservation using On Demand Computational of Gain using Vertical Partition

Ankita Jain
Computer Science Of Dept.
Radha Raman Institute Of Technology and Science, RGPV
Bhopal, India

Shubha Dubey
Computer Science Of Dept.
Radha Raman Institute of Technology and science, RGPV
Bhopal, India

Anurag Jain
(HOD)
Computer Science of Dept.
Radha Raman Institute of Technology and Science, RGPV
Bhopal, India

***Abstract*— Data mining means the extraction of some meaningful information so that analysis can be done quickly. But privacy is an important factor during the data mining process. Hence various techniques are implemented to provide privacy preservation in data mining. The existing technique implemented for privacy preservation in data mining provides efficient privacy at multi level trust [8]. But the technique implemented is not efficient in terms of computational time and error rate. Hence an efficient technique is implemented which is based on the concept of multiple parties for the computation of data mining. The proposed technique implemented here provides privacy preservation with less error rate and time.**

***Index Terms*—SSN, privacy preserving data mining, Multi level trust, clustering, classification.**

I. INTRODUCTION

Privacy is becoming an important issue in various data-mining applications such as health care, security, financial, and other types of sensitive data. It has become important in counter terrorism and homeland defence-related applications. These applications require creating of profiles, constructing of social network models, and detecting terrorist communications from privacy sensitive data. However, combining such diverse data sets may violate the privacy laws. Even though health organizations are allowed to give data as long as the identifiers (e.g., name, SSN, address, etc.,) are eliminated, it is considered unsafe since re-identification attacks can be constructed for linking different sets of public data to identify the initial subjects. This requires a well designed technique that provides careful attention for hiding privacy-sensitive information, while securing the inherent statistical dependencies which are important for the applications of data mining.

Today the data storage requirements are becoming large, as the large number of data is evolving day by day. As the Large database is required to store the large amount of data, so the privacy of the data is also an important factor. This has caused concerns that personal data may be used for a variety of intrusive or malicious purposes. Privacy preserving data mining [1], [2], [3] helps to achieve the aim of data mining by preserving the privacy of sensitive data here data mining goals without scarifying the privacy of the individuals and without learning underlying data values. Privacy-preserving data mining (PPDM) refers to the area of data mining that seeks to safeguard sensitive information from unsolicited or unsanctioned disclosure. The previous privacy preserving solutions were limited to only single level trust, which was not sufficient to preserve the privacy of information. So by expanding the scope from single level trust, here in the proposed system, multilevel trust solution for privacy preservation is applied in which data owner generates the different perturbed copies of same data for data miners of different trust levels. In privacy-preserving data mining (PPDM), data mining algorithms are analyzed for the side-effects they incur in data privacy, and the main objective in privacy preserving data mining is to develop algorithms for modifying the original data in some way, so that the private data and private knowledge remain private even after the mining process [4]. A number of techniques such as Trust Third Party, Data perturbation technique, Secure Multiparty Computation and game theoretic approach, have been suggested in recent years in order to perform privacy preserving data mining. However, most of these privacy preserving data mining algorithms such as the Secure Multiparty Computation technique, were based on the assumption of a semi-honest environment, where the participating parties always follow the protocol and never try to collude. As mentioned in previous works on privacy-preserving distributed mining [5], it is rational for distributed data mining that the participants are assumed to be semi-honest, but the collusion of parties' for gain additional benefits cannot be avoided.

However, in the real world different trust levels might be required. It is the motivation behind exploring multilevel trust. For instance a government organization might have internal and external miners. Moreover, it also wants to give data to public. In this case obviously it needs to generate multiple copies of data with different perturbation applied based on the trust level of the data miners and general public. There is a problem in this approach too. When adversaries are able to get internal and external miner copies, they can establish the identity information from the data. Therefore it is very challenging to implement multiple level trust based PPDM. To avoid this problem before giving data to data miner, it is perturbed in such a way that sensitive information is encoded or altered in order to ensure that the privacy of data is preserved. This feature is known as Privacy Preserving Data Mining (PPDM). There were many researches on the PPDM [6], [7],When compared to single level trust

scenario, many perturbed copies are required by the data owner to ensure non-disclosure of sensitive details. The number of perturbed copies depend on the trust level of the data miner. If the miner is trusted more, then it is likely that the number of perturbations of data to be published is less. However, from multiple diverse and perturbed copies the miner might produce original information accurately. This is the problem with the approach. Preventing such diversity attacks is a challenging task in multi-level trust based PPDM.

II. LITERATURE SURVEY

In this paper [8, 9], author has the possibility of preservative perturbation based PPDM to multilevel trust (MLT) is spread out, by comforting an implicit assumption of single-level trust in the exiting work. MLT-PPDM make available data owners to generate in a different way perturbed copies of its data for different levels of trust. This proposed work is based on perturbation based privacy preserving data mining. Here random perturbation come within reach of is functional to provide privacy on the data set. Until that time privacy is inadequate to single level trust in on condition that privacy to the data but now it is developed to multi level trust. The difficulty with existing multi level trust PPDM algorithms is that they fail to protect form non linear attacks. The most significant confront dishonesty in preventing the data miners from combining the copies at singular trust levels to in cooperation modernize the initial data more accurately than what is allowed by the trusted data owner. The problem is addressed by properly show a relationship noise diagonally the copies at different trust levels in the record. It is to prove that if one can design the noise covariance matrix to have corner-wave property, then data miners will have no assortment gain in their cooperative modernization of the unusual data. This maintains is demonstrated and make obvious the efficiency of the explanation through mathematical valuation. Last but not the least, this solution allows data owners to generate perturbed copies of its data at arbitrary trust levels on-demand. This offers the data owner maximum flexibility as a property. One should believe that multilevel trust privacy preserving data mining can find applications in many of the fields. This work takes the initial step to enable the MLT-PPDM services. There are many more concentration and significant directions attraction investigating. For example, it is not smallest amount comprehensible how to spread out the possibility of other move towards in the areas of incomplete information hiding, such as indiscriminate rotation-based data perturbation, preservation replacement and k-anonymity, to multilevel trust. It is also of great concentration to make bigger this come within reach of touch developing data streams this proposed work make uses developed batch generation to make available privacy in the multi level trust in which data will perturb multiple times so that it can keep away from non linear attacks.

Privacy Preserving Data Mining (PPDM) is used to mine appropriate knowledge from large quantity of data and at the same time save from harm the susceptible information from the data miners. The predicament in privacy-sensitive domain is get to the underneath of by the development of the Multi-Level Trust Privacy Preserving Data Mining (MLT-PPDM) where multiple in a different way perturbed copies of the same data is corrected to data miners at different trusted levels. In MLT-PPDM [10] data owners manufacture perturbed data by a variety of methods like Parallel generation, Sequential generation and On-demand generation. MLT-PPDM is strong in opposition to the diversity attacks. Here author present incomplete information hiding methodologies like random rotation perturbation, maintenance alternate and K-anonymity are have as a featured with MLT-PPDM to increase data security and to avoid leakage of the sensitive data. As a final point MLT-PPDM approach is improved to deal with alongside the non-linear attacks. So the Multi-Level Trust in Privacy-Preserving Data Mining when put together with incomplete information hiding methodologies help to find the accurate set of scales between maximum investigation consequences and maintain the assumptions that make known private information about organizations or entity at a smallest amount.

Data mining representations dataset to the examiners alternatively, it is possible to ascertain sensitive information from the given dataset. In this paper [11] author has suggested a new approach for privacy preserving data mining. The existing perturbation based PPDM models guess single level trust on data miners. In this work author mainly work on multilevel trust based PPDM which make available more elasticity to data owner in choosing the level of privacy to data. To defeat this crisis Privacy Preserving Data Mining (PPDM) came into subsistence. This kind of attack is prevented by using noise correlation matrix across the copies to deny the attackers not to have variety alternative. PPDM make certain that the datasets are produced in such a way that they cannot make known distinctiveness of the entities in attendance in the existing dataset. Before data is published, perturbation of data is made with the intention of preserve privacy in the data. Accessible PPDM systems take for granted single level trust on data miners. In recent times Li et al. relaxed this by introducing Multilevel Trust in PPDM. They consider that if the data miner is more trusted less perturbation is involved. On the other hand, malicious data miners can create distinctiveness information by combining multiple perturbed copies. It checks malicious attacks and it is a proper PPDM. In this paper [11] author has put into practice that multilevel trust based PPDM which make possible data owners to have lack of restrictions to decide the level of privacy required. Based on this trust level perturbations are made. Here author try to construct a sample application that shows the confirmation of perception. The experimental consequence make known that the proposed move toward is efficient and presents elasticity to data owners is robust and effective.

III. PROPOSED METHODOLOGY

The proposed methodology implemented here consists of following phases:

1. Take an input dataset from which some meaningful information can be extracted.
2. Now "On Demand" of the untrusted third party the dataset can be partitioned vertically into 'N' parties.
3. Each of the party contains a set of attributes with their respective classes.
4. Computation of Information Gain by each of the party and send to UTP.
5. UTP on the basis of information Gain will select the attributes having information gain and the remaining attributes with less information gain can be removed from the dataset.
6. Now clustering is done for each of the party on the basis of classes available.
7. Finally decision tree is generated from the available clustered dataset.

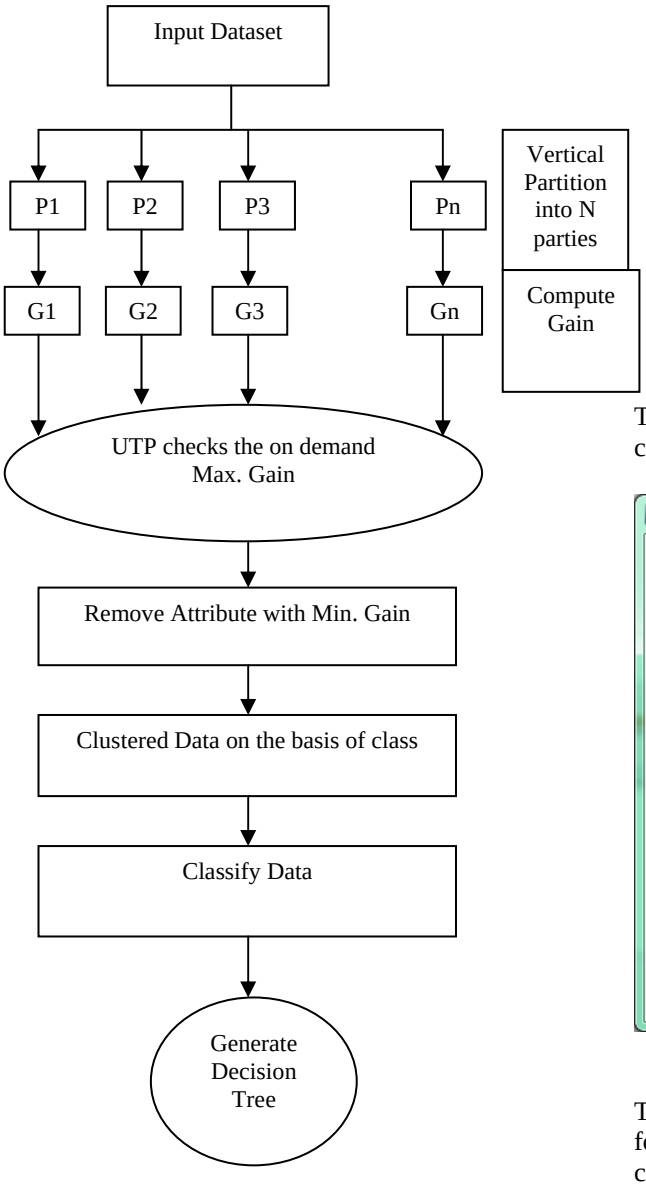

Figure 1. Flow Chart of the methodology

On demand Vertical Partition

Define P_1, P_2… P_n Parties. (Vertically partitioned).
Each Party contains R set of attributes A_1, A_2, …., A_R.
C the class attributes contains c class values C_1, C_2, …., C_c.
For party P_i where i = 1 to n do
If R is Empty Then
Return a leaf node with class value
Else If all transaction in $T(P_i)$ have the same class Then
Return a leaf node with the class value
Else
Calculate Expected Information classify the given sample for each party P_i individually.
Calculate Entropy for each attribute (A_1, A_2, …., A_R) of each party P_i.
Calculate Information Gain for each attribute (A_1, A_2,…., A_R) of each party P_i
Calculate Total Information Gain for each attribute of all parties (TotalInformationGain( )).
$A_{BestAttribute} \leftarrow$ MaxInformationGain( )
Let V_1, V_2, …., V_m be the value of attributes. $A_{BestAttribute}$ partitioned P_1, P_2,…., P_n parties into m parties
$P_1(V_1)$, $P_1(V_2)$, …., $P_1(V_m)$
$P_2(V_1)$, $P_2(V_2)$, …., $P_2(V_m)$
. .
. .
$P_n(V_1)$, $P_n(V_2)$, …., $P_n(V_m)$
Return the Tree whose Root is labelled $A_{BestAttribute}$ and has m edges labelled V_1, V_2, …., V_m. Such that for every i the edge Vi goes to the Tree
NPPID3(R – $A_{BestAttribute}$, C, ($P_1(V_i)$, $P_2(V_i)$, …., $P_n(V_i)$))
End.

IV. RESULT ANALYSIS

The experimental results are performed on two datasets consensus and Iris dataset.

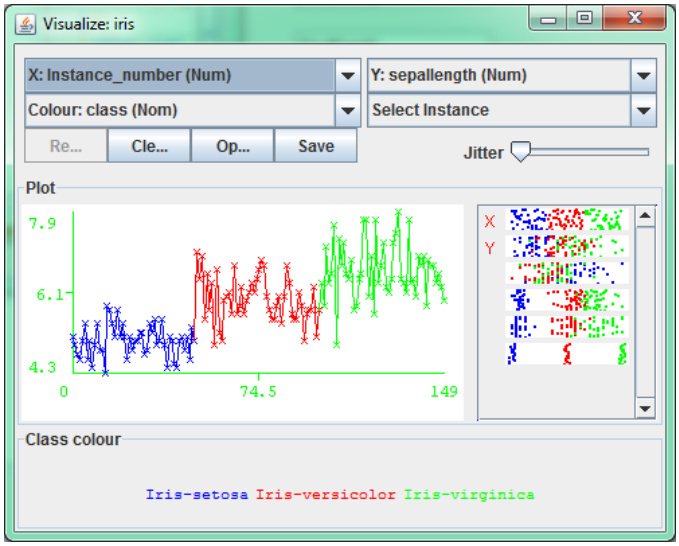

Figure 2. Threshold Curve based on Iris Dataset

The table shown below is the analysis of the two datasets for number of parties. The analysis is done for the computation of error rate.

	Error rate		
Dataset	**2**	**3**	**4**
Concensus	0.12	0.09	0.07
Iris	0.18	0.14	0.11

Table 1. Analysis of Error Rate

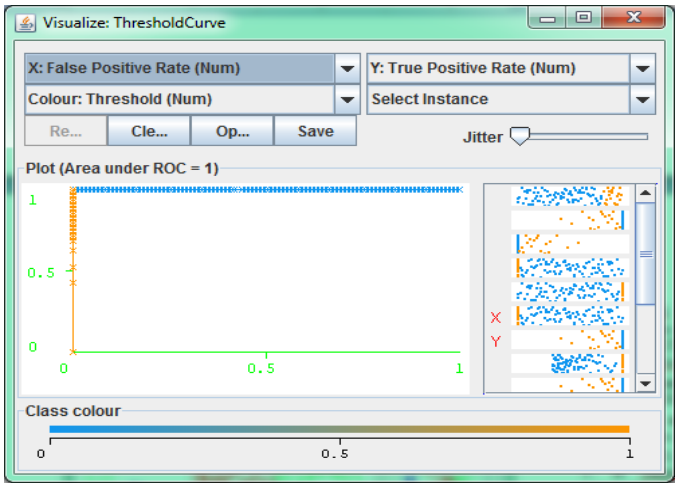

Figure 3. False Positive rate

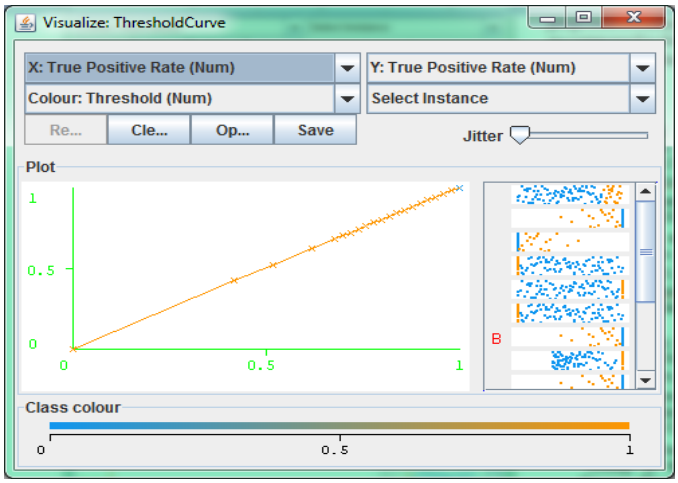

Figure 4. True Positive Rate

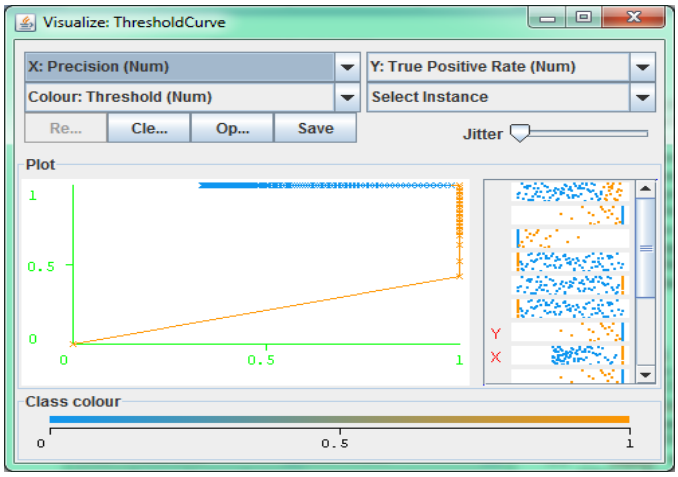

Figure 5. Precision

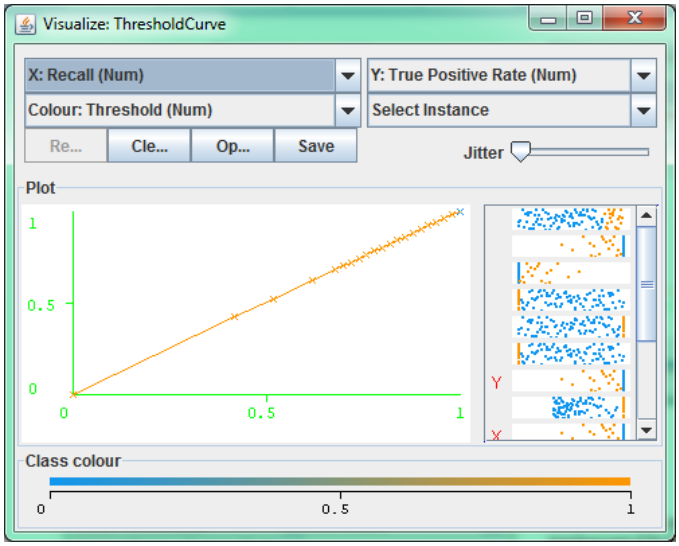

Figure 6. Recall

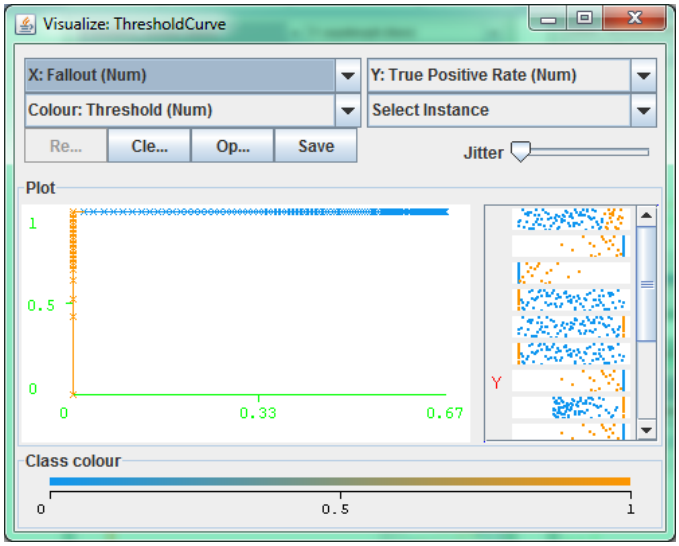

Figure 7. Fallout

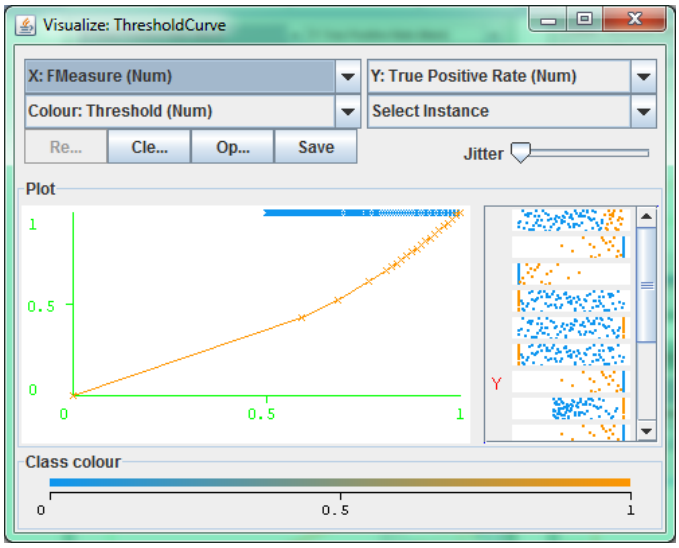

Figure 8. F-Measure

The figure shown below is the analysis and comparison of computational time for a number of parties on Conesus and iris dataset.

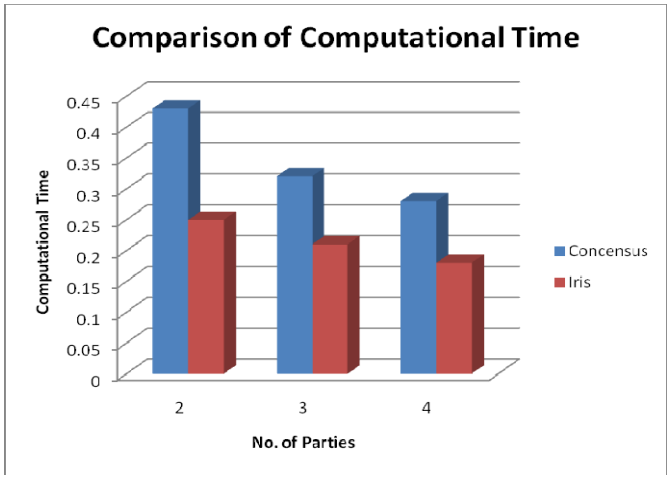

Figure 9. Comparison of Computational Time

The figure shown below is the analysis and comparison of Accuacy for a number of parties on Conesus and iris dataset.

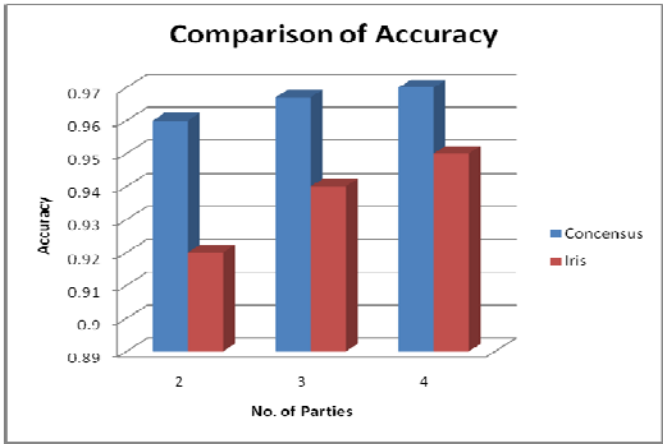

Figure 10. Comparison of Accuracy

The table shown below is the analysis of the two datasets for number of parties. The analysis is done for the computation of Kappa Statistic.

	Kappa		
Dataset	**2**	**3**	**4**
Concensus	0.94	0.95	0.96
Iris	0.93	0.94	0.96

Table 2. Analysis of Kappa Co-efficient

V. Conclusion

Here in this paper a new and efficient technique is implemented for the privacy preservation so that the extracted knowledge from the dataset can't be accessed by the external or un-authorized users. The proposed methodology implemented here is based on the concept of dividing the dataset into 'N' parties and computation of information and On demand of information gain provides the most dependent attributes on the basis of which data can be analyzed. The proposed technique implemented here provides efficient results as compared to the existing technique implemented for privacy preservation in data minig.

References

[1] D. Agrawal and C.C. Aggarwal, "On the Design and Quantification of Privacy Preserving Data Mining Algorithms," Proc. 20th ACM SIGMOD-SIGACT-SIGART Symp. Principles of DatabaseSystems (PODS '01), pp. 247-255, May 2001.

[2] R. Agrawal and R. Shrikant, "Privacy Preserving Data Mining,", Proc. ACM SIGMOD Int'l Conf. Management of Data 2000.

[3] Y. Lindell and Benny Pinkas, "Privacy Preserving Data Mining", Proc. Int'l Cryptology Conf. (CRYPTO), 2000.

[4] Verykios V.S., Bertino E., Fovino I.N., Provenza L.P., Saygin, Y. & Theodoridis Y.(2004a). State-of-the-art in privacy preserving data mining, SIGMOD Record, Vol. 33, No. 1, pp.50-57.

[5] Lindell Y. & Pinkas B.(2009). Secure Multiparty Computation for Privacy-Preserving Data Mining. Journal of Privacy and Confidentiality, Vol 1, No 1, pp.59-98.

[6] S. Papadimitriou, F. Li, G. Kollios, and P.S. Yu, "Time Series Compressibility and Privacy," Proc. 33rd Int'l Conf. Very Large Data Bases (VLDB '07), 2007.

[7] F. Li, J. Sun, S. Papadimitriou, G. Mihaila, and I. Stanoi, "Hiding in the Crowd: Privacy Preservation on Evolving Streams Through Correlation Tracking," Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), 2007.

[8] Yaping Li, Minghua Chen, Qiwei Li, and Wei Zhang , "Enabling Multilevel Trust in Privacy Preserving Data Mining" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 24, NO. 9, SEPTEMBER 2012.

[9] B.Anitha , B.Hanmanthu , B.Raghu Ram , "Enhanced Batch Generation based Multilevel Trust Privacy Preserving in Data Mining" International Journal of Computer Applications Volume 82 – No.9, November 2013.

[10] M.S. Ramya Partial Information Hiding in Multi-Level Trust Privacy Preserving Data mining, 2012.

[11] Sailaja.R.J.L, P.Dayaker Preventing Diversity Attacks in Privacy Preserving Data Mining International Journal of Computer Trends and Technology (IJCTT) – volume 4 Issue 9– Sep 2013.