ISSN:0975-9646

Privacy Preserving Data Mining Using Association Rule With Condensation Approach

Supriya S. Borhade[1] , Bipin B.Shinde[2]

Researcher, Department of Computer Engineering,
Pune University, Pune. [1]
Researcher ,Department of Computer Engineering,
RGPV,Bhopal. [2]

***Abstract*—With the rapid development data mining within various fields and security and privacy concerns come into view. In data mining while releasing micro-data or patterns from large databases individual or organizational private data may get compromise the information. The main aim behind privacy preserving data mining is to maximizing analysis outcome and minimizing disclosure of individuals or organizational private data. Association rule mining explores interesting relationship between data. This paper is based on concepts: condensation method and association rule. SMC (secure multiparty computation) securely transferring data over the network with hiding process which hide sensitive association rule which create threat to privacy. Now privacy preserving data mining has become increasingly popular because it provides sharing of private or sensitive data for analysis purposes. Most of people and organizations are afraid or hesitate to share their data or refusing to share their data or sometimes might provide wrong data. Due to rapid proliferation of private information on the internet, lot of research has been done in recent years for privacy preserving data mining. Users are unwilling to provide private or personal data unless and until privacy is assured. Sometimes automated transaction system holds or track information about individuals in day today's life. For example credit card transactions.**

***Keywords*—Association Rule, Condensation, PPDM, SMC, Perturbation.**

I. INTRODUCTION

Now it has become more and more important in recent years and even for future because day by day data is growing rapidly and even increasing ability to store that data about users and organizations personal information. It is difficult to handle and preserve such a huge data.

In recent years to provide privacy preserving data mining number of techniques have been suggested like classification, K anonymity, clustering and association rule mining algorithm to hold information. Various data mining techniques are successfully used to retrieve useful knowledge in order to provide support a variety of domains like marketing, medical diagnosis, research, weather forecasting ,military or security. But it is still challenging issue in various domains to provide privacy to certain kind of information without violating the individual's privacy. For example- In credit card transactions while purchasing frequently used items, while mining patients private data in health care or research purpose.

Data mining is spreading widely throughout an area or a group of people on internet so privacy concerns are increasing. Most of the organizations collect data about users for their own specific needs. However different branches within an organization themselves may need to share information. In this situation each organization or branch must be sure that privacy of the individual is not violated and private or sensitive business information should not get disclosed.

A wide variety of sources holds individuals private data such as banks(personal information name, birth-date , PAN), police records(name, address, birth marks, physical appearance), airports (passport number departure,destination, duration, age and gender) expenditure data while purchasing or bank transaction.

In most of countries sharing individual private data or exposing confidential information is against the law to share or make such information publicly available to others.

In this project, in order to preserve privacy of such information records can be de- identified before the information records are shared with other users without violating individual's privacy.This can be done by deleting unique identity fields item such as passport no,age. But even if this information is deleted there are still other kinds of information fields when linked with other fields available in datasets could identify the individual.

To provide security for such types of violations, I need variety of data mining algorithm. It is important problem in recent years, because of the large amount of user's data tracked by automated systems on the internet. Due to the growing market of electronic commerce on the internet has resulted holding large amounts of transactional data with personal information about users.While doing simple transaction such as using credit card results in automated system store information about users buying behavior. Sometime users are unwilling to supply such private data unless and until privacy is guaranteed.

In order to provide and ensure effective data gathering, it is important to implement methods which minimize disclosure risk and maximize the mining analysis outcome with a guarantee of privacy.

The paper is organized as follows. First section reviews the critical points of current knowledge including substantive findings as well as theoretical and methodological contributions to privacy preserving data mining technique along with comparison. Implementation section contains execution plan and in the last section concludes the paper work.

II. LITERATURE SURVEY

Several research communities contributed their work to privacy preserving data mining using variety of technique. Let first discuss privacy preserving work in the data mining community. Over the past few years, variety of approaches has been proposed in the area of privacy preserving data mining. Some of the important approaches include cryptographic approach, heuristic approach and reconstruction based approach.

The concept of the heuristic approach method is the way to hide sensitive rules which is used to be mined from the dataset while maximizing the outcome of the released data.

The second approach is Cryptography based method, This approach has been developed to solve the problem such as SMC: If Two or more parties want to perform a computation based on their private inputs, but party is unwilling to disclose its own output to any other else. Such problem is referred to as the Secure Multiparty Computation (SMC) problem.

Next approach is reconstruction based method,In this approach they first used some methods to distort or twist the values of the original data and then release these twisted data.

Another important approach is the Access control based approach. It was built over existing technologies was proposed called Multi- relational association rules (MRAR). This model has three layers those are Authenticator, checker and the database server. MRAR is the type of policy where the users are associated to mining levels which is mandatory access control. Disadvantage of MRAR is that it is not always possible to assign sensitivity levels to data in case level contains another level.

Anonymization Method: This method is used to protect user's identities while releasing micro data. The k-anonymity protects against identity disclosure. But it does not provide sufficient protection against field's disclosure and original data can be reconstructed.

Perturbation Method: Independent operation is performed on the different fields by this method. This method does not reconstruct the original data values, but only distribution,new algorithms have been developed which uses these reconstructed distributions to carry out mining of the data available.

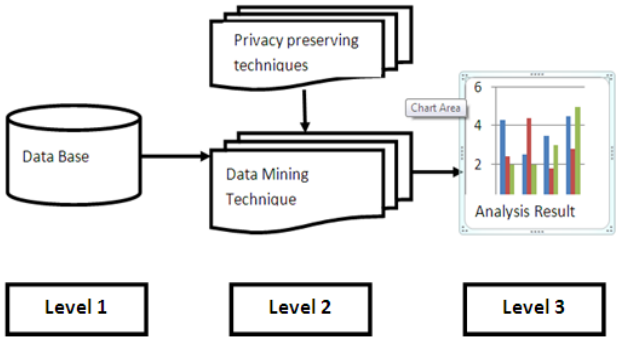
Randomized Response Method: This is very simple technique which can be easily implemented at the time of data collection. It is useful technique for hiding individual data in PPDM. This method results in high information loss. It is not suitable for multiple attribute databases.

Rather than participants input and output no more data get disclosed to a participant while performing computation. But it is important to let know that the data modification results is degrading the performance of database.

Degradation of data is measured in two dimension metrics .The first dimension measures the confidential data security and second measures the loss of functionality.

III. IMPLEMENTATION DETAIL

Privacy of all the records is not the same but can vary to a notably large extent. In various applications they may require different privacy requirements for different groups of individuals.



The heterogeneous condensation is able to handle both static and dynamic data sets. This approach creates condensed groups of records, which can be used directly with a variety data mining algorithm or directly with minor change in existing data mining algorithms.

Let consider that I have a set of N records and each of which contain $d$dimensions, also assume that associated with each data point i,with corresponding privacy level $p(i)$. The complete database is represented by D, Whereas the database corresponding to the privacy level p is represented by Dp.The data is partitioned into number of groups of records.Records within a given group cannot be distinguished from one another. For all the groups need to maintain certain summary statistics about the records. This summary statistics provides the ability to apply data mining algorithms directly to the condensed groups of records.

The size of groups may be of different size but its size must be of at least equal to the desired privacy level of each record of that group.

The size of the group must be at least equal to the maximum privacy level of any one record from that group. Each group of records is referred to as condensed group.

Let assume G be a condensed group containing the records $\{R_i..R_k\}$,also assume that each record R_i contains the d dimensions which are represented by $(R_{i1}..R_{id})$.

For each group of records has to maintain the following information.

1.I need to maintain the sum of the corresponding values for each attribute j ,which is represented by $F_{sj}(G)$.
2.I need to maintain the sum of the product of corresponding attribute values for each pair of attribute values for each pair of attributes i and j,which is represented by $S_{cij}(G)$.
3.Then I am maintaining the sum of the privacy levels other records in the group$P_s(G)$.
4.Lastly I am also maintaining the total number of records K in that group $n(G)$.

While constructing group each record must be inserted in a group which must be at least equal to maximum privacy level of any record in the group.

Firstly I have to classify the records based on their privacy levels and then create the groups for various privacy levels individually.

ALGORITHM : CreateGroup (Level: MaxPLevel, Database : D1)

Begin

$P = 2$;

$H1 = Groups\ from\ singleton\ points\ in\ D1$;

While ($p \leq MaxPLevel$) do

Begin

$H_p = segment(D_p, p)$;

$(H_{p-1}, H_p) = Cannibalize(H_{p-1}, H_p)$;

$(H_{p-1}, H_p) = Attrition(H_{p-1}, H_p)$;

$H_p = H_p\ U\ H_{p-1}$;

$p = p + 1$;

End;

End;

ALGORITHM Segment(Database: D_p, PLevel: p)

Begin

While D_p contains at least p data points

Begin

Sample a data points R from D_p;

Find the $(p-1)$ data points closest to R in D_p;

Create the group G of p data points comprising R and p -1 other closest data points;

Add G to the set of groups H ;

End;

Assign remaining data points in D_p to closest groups ;

End;

ALGORITHM Cannibalize(Groups: H_{p-1}, H_p)

Begin

For each group $G \in H_{p-1}$ do

Begin

For each point in G perform Temporary assignment to closest group in Hp;

If (SSQ of temporary assignment lower or (H_{p-1} contains Lower than $(p-1)$ members))

Then make the assignment permanent ;

Else

Keep old assignment;

End;

End;

ALGORITHM Attrition(Groups : H_{p-1}, H_p, PLevel: p)

Begin

For each data point R in H_p do

Begin

$Distc(R, p) = Distance\ of\ R\ to\ centroid\ of\ its\ current\ group\ in\ H_p$

$Disto(R, p-1) = Distance\ of\ R\ to\ centroid\ of\ its\ closest\ viable\ group\ in\ H_{p-1}$;

$Improve\ (R) = Distc(R, p) - Disto(R, p-1)$;

End;

For each group in H_p with atleast $p' > p$ points do

Begin

Find(if any)the atmost$(p' - p)$ datapoints with the largest value of Improve (.) function which is larger than 0;

Assign these atmost $(p' - p)$points to their corresponding closest group in $Hp - 1$;

End;

The privacy level of each group is calculated by the number of records present in it. The information loss is defined by the average difference of the record about their centroid. The input given to the algorithm is the database D with the maximum privacy level which is referred by *MaxPLevel*.

Dpdenotes the privacy level requirement of p in the segment of the database.

H_p is used to denote privacy level of p with their set of groups.

$D1$is the database which contains the set of points which have no privacy constraints at all.

$H1$ group consist of the singleton items from the database $D1$. Iterative algorithm is used to construct statistics of the groups in Hp. Segmentation process is first step to construct the group Hp. Once the segmentation process has been completed I have to apply the process of Attrition and cannibalize in order to reduce the information loss without compromising on the privacy requirement.

If group $G \in H_{p-1}$does not form a natural cluster then cannibalization is performed. In this case it is more effective to cannibalize the group G and group members can be reassigned to one or more cluster in Hp

For better fitment group in H_{p-1}need to move the excess points using attrition process. The quality of data representation in terms of reducing the level of information loss is improved by movement of excess points.

In first level dataset is collected from condensed group and second level will generate all association rule of collected datasets. And in the final level association rule is hidden by using hybrid algorithm which is combination of both ISL and DSR technique.

A. Association Rule Generation

Data miner applies apriori algorithm on dataset to get frequent item set and hence to find out association rules $A \rightarrow B$contains which satisfy the following two conditions:

$$Support(A \rightarrow B) = \frac{(A \cup B)}{n} \geq min\ \_support$$

$$Confidence(A \rightarrow B) = \frac{Support(A \cup B)}{Support\ (A)} \geq min\ \_confid$$

Where $n$number of record available $min\text{-}support$ and $min\text{-}confidence$ are user defined threshold values .

B. Hiding Association Rule

There are two ways for hiding association rule. First by increasing the support of A I can decrease the confidence of a rule $A \rightarrow B$. Second by decreasing the support of Y, the right hand side of the rule, this would reduce the confidence faster than simply reducing the support of $(A \cup B)$.For decreasing support of an item, I will modify one item at a time by changing from 1 to 0 or from 0 to 1 in a selected transaction. Hybrid algorithm will first hide the rules in which item Ain RHS and then hide the rules in which item A in LHS.

IV. CONCLUSION

In this paper I have proposed the system for maximizingthe privacy and minimizing information losswhile sharing the data without disclosing individual's identity and secure the database owners privacy rules. For this model condensation are used for preserving covarianceinformation and preserving privacy. Associationrule hiding with ISL and DSR technique is used whichis based on modifying the database transaction so thatthe confidence of association rules can be reduced.

REFERENCES

[1] lberto Trombetta, Wei Jiang, Elisa Bertino, "Privacy Preserving Updates To Anonymous And Confidential Databases",*IEEE Transactions On Dependable And Secure Computing*, Vol. 8, No. 4, PP. 578-587, 2011.

[2] Yi-Hung Wu, Chia-Ming Chiang, Arbee L.P. Chen, "Hiding Sensitive Association Rules With Limited Side Effects ", *IEEE Transactions On Knowledge And Data Engineering*, Vol. 19, No. 1, PP 29-41, 2007.

[3] Tamir Tassa, "Secure Mining Of Association Rules In Horizontally Distributed Databases",*IEEE Transactions On Knowledge and Data Engineering* , Vol. 1, No. 99, PP. 1-14, 2013.

[4] Murat Kantarcioglu , Wei Jiang, "Incentive Compatible Privacy-Preserving Data Analysis", *IEEE Transactions On Knowledge And Data Engineering*, Vol. 25, No. 6, PP. 1333-1335, 2013.

[5] Sowmyarani C N, Dr. G N Srinivasan, "Survey on Recent Developments in Privacy Preserving Models", *International Journal of Computer Applications*, Vol. 38, No.9, PP. 18-22,2012.

[6] Bin Zhou, Yi Han, Jian Pei, Bin Jiang, YufeiTao,YanJia,"Continuous Privacy Preserving Publishing of Data Streams", *EDBT* , P. 24,S26, 2009.

[7] Madhan Subramaniam, Senthil R," An Analysis on Preservation of Privacy in Data Mining",*(IJCSE) International Journal on Computer Science and Engineering*,Vol. 02, No. 05, PP.1696-1699, 2010.

[8] Dr.K.P.Thooyamani, Dr.V.khanaa, "Privacy-Preserving Updates to Anonymous and Confidential Database",*International Journal of Data Mining Techniques and Applications*, Vol. 01,PP. 2278-2419, 2012.

[9] Kirubhakar Gurusamy, Venkatesh Chakrapani, " An assessment of Identity Security in Data Mining", *International Journa of Science and Modern Engineering (IJISME)*, Vol. 1, No. 7, PP. 29-31, 2013.

[10] Manish Sharma, Atul Chaudhary, Manish Mathuria, Shalini Chaudhary, " A Review Study on the Privacy Preserving Data Mining Techniques and Approaches" , *International Journal of Computer Science and Telecommunications*, Vol. 4, No. 9,PP.42-46, 2013.

[11] Ekta Chauhan, Sonia Vatta," Review of Privacy Preserving in Data Mining Using Homomorphic Encryption", *International Journal of Advanced Research in Computer Science and Software Engineering* , Vol. 3, No. 5, PP. 1431-1433, 2013.

[12] Mohammad Reza Keyvanpour, Somayyeh Seifi Moradi," Classification and Evaluation the Privacy Preserving Data Mining Techniques by using a Data Modification based Framework",*International Journal on Computer Science and Engineering (IJCSE)*, Vol. 3, No. 2, PP. 862-870, 2011.

[13] K. Sathiyapriya, Dr. G. SudhaSadasivam,"A SURVEY ON PRIVACY PRESERVING ASSOCIATION RULE MINING", *International Journal of Data Mining and Knowledge Management Process*, Vol.3, No.2, PP. 119-131, 2013.

[14] Murat Kantarcioglu, Chris Clifton, "Privacy-Preserving Distributed Mining Of Association Rules On Horizontally Partitioned Data", *Knowledge And Data Engineering,IEEE Transactions*,Vol. 16, No. 9, 2004.

[15] BENJAMIN C. M. FUNG, KE WANG, RUI CHEN, PHILIP S. YU,"Privacy-Preserving Data Publishing: A Survey of Recent Developments " , *ACM Computing Surveys*, Vol. 42,No. 4, 2010.