PRESERVING THE PRIVACY OF USER BY USING ANONYMIZATION TECHNIQUE

Tanaya Gajanan Marathe, [2]Madhura Nitin Raverkar, [3]Sneha Gopalkrishna Suryawanshi, [4]Minakshi Fakira More

Students, Department of Computer Engineering
SSBT COET Bambhori, Jalgaon

***Abstract*: Today we are living in the complex world in which sensitive information privacy is the main issue. Many algorithms are used to protect sensitive information in mined data which is not efficient because resulted output can be easily linked with public data so it reveals user identity. Many techniques are used to protect privacy in data mining. Data mining approaches aim to avoid direct use of sensitive data. Proposed system uses anonymization techniques which help to eliminate privacy risk in data preparation. In the proposed work, an anonymization technique is given that is a combination of the benefits of anatomization, and enhanced slicing approach adhering to the principle of k-anonymity for the purpose of dealing with high dimensional data with multiple sensitive data. The anatomization approach dissociates the correlation observed between the quasi identifier attributes and sensitive attributes and yields two separate tables with non-overlapping attributes. In the slicing algorithm, vertical partitioning does the grouping of the correlated sensitive attributes in sensitive table together and thereby minimizes the dimensionality. The anatomization technique minimizes the loss of information and slicing algorithm helps in the preservation of correlation and utility which in turn results in reducing the data dimensionality and data loss.**

***Keywords*- Privacy, anonymization approaches, mined data, anatomization, slicing.**

Introduction:-

Today's health care providers store a huge amount of sensitive data as a content of their business. The sensitive information can be personally recognizable information from the clients. So in this any kind of misuse of this information creates a critical risk in their business. When making the sensitive information available to the public, it is necessary for them to protect it from any abuse and risk. Data anonymization is the one and only popularly used approach. It modifies and changes information. This approach tries to ensure the identity along with the sensitive information of the data subjects when data is shared for diverse purposes. There are multiple anonymization techniques that prevail for retaining privacy. They are generalization, suppression, anatomization, bucketization, permutation, and perturbation. The majority of the strategies above focuses on anonymizing the micro data with only single SA. As they are not suitable for functional usage, the current challenge is to preserve the multiple SA efficiently in the high dimensional data.

Motivation:-

The generalization for k-anonymity and bucketization for l-diversity are popularly understanding privacy preservation strategies. Hence K-anonymity used in the anonymization technique for data preparation. Generalization for k-anonymity ignores a huge amount of high dimensional data. In order to get over imperfection in generalization, an inventive anatomization technique is brought into use, it reduces the loss of data, although it is efficient for preserving privacy for single sensitive data. The anatomization preserves privacy as it is not illustrative of the sensitive value corresponding to any tuple, which might be assumed randomly from sensitive table. Motivated by these works, proposed system focuses on the preservation of the privacy of data with numerous sensitive attributes with lesser information loss and better data utility. In this work an anatomization approach is employed to minimize the loss of information by releasing the quasi-identifier attributes directly. On the contradictory, slicing maintains the correlation in the column and then carries out the break of correlation across the columns by means of vertical and horizontal partitioning. Every attribute in a column can be considered in the form of a sub table. This removes the dimensionality with respect to the data. Additionally, the research work functions in according with the principle of k-anonymity and l-diversity that does not impact the quasi-identifier values which are directly released by means of anatomization. Also uses reconstruction and modification approaches for preventing the privacy of data unwanted discovery.

Literature Survey:-

1] A Framework for Categorizing and Applying Privacy-Preservation Techniques in Big Data Mining:- Lei Xu et al.[1] surveyed on the Rampart framework categorizes protection approaches and encouraged interdisciplinary solutions to the growing variety of privacy problems associated with knowledge discovery from data. In this, anonymization technique is first stage which is used to protect the privacy of data. A key issue of data anonymization was to choose an appropriate privacy model to quantify how much privacy can be preserved. So they focussed on K-anonimity which is used to modify the values of QI so that every tuple in the anonymized table is indistinguishable from at least k − 1 other tuples.

2] Slicing – A New Approach for Privacy Preserving Data Publishing .

Zang J. Molloy et al. [6] presented work on slicing. Several anonymization techniques, such as bucketization and generalization, have been designed for privacy preservation. Recent work has shown that generalization loses considerable amount of information mainly for high-dimensional data. Bucketization does not prevent membership disclosure. In the paper, they present a

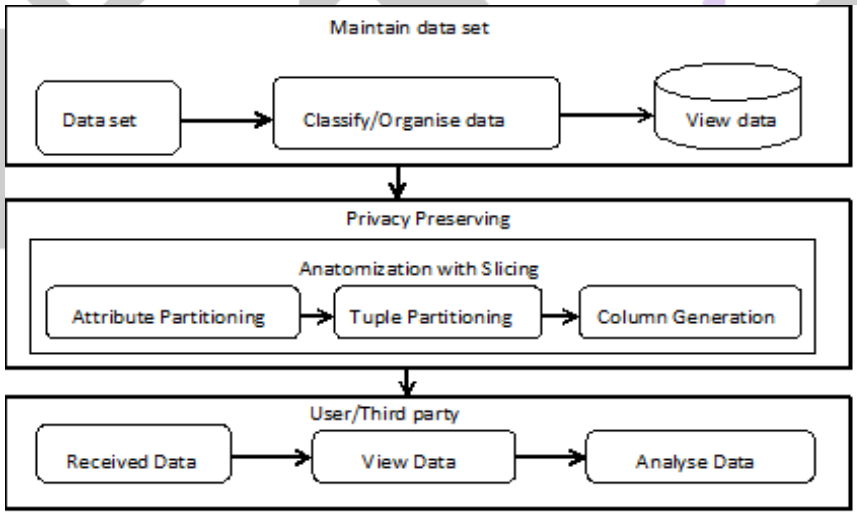
novel technique called slicing, which partitions the data horizontally as well as vertically. The paper shows that slicing preserves better data utility than generalization and can be used for membership disclosure protection.

3] Analysis of Data Security by using Anonymization Techniques:-
Preet Chandan Kaur et al.[3] proposed the work on for preserving the privacy of personal information for that they used anonymization technique which were applied to avoid retrieval of data from database. Thus, privacy preservation means to protect the data value and it is used for data mining in order to get the valid and accurate results. These were discussed and successfully analyzed with different parameters such as revealed co-relation quality, loss of information, type of data, security and membership disclosure in the paper.

Proposed System:-
Proposed system introduces anonymization technique that is anatomization with slicing to eliminate the privacy risk in data preparation. It also uses reconstruction and modification approaches to protect sensitive data from unwanted discovery by hiding rules, encryption and mining algorithm.

Problem Definition: The data collector and data miner are two different entities, with the data collector processing information from its original owners and then making it available to the data miner. Processing must be done in such a way that makes it impossible for the data miner to identify the identities of data owners and other sensitive information, yet still produce data that the data miner finds useful. This is the main need of the project. For that purpose, anonymization approach, reconstruction and modification approaches are useful.

1]Anatomization with slicing-
The data collector and data miner are two different entities, with the data collector processing information from its original owners and then making it available to the data miner. Processing must be done in such a way that makes it impossible for the data miner to identify the identities of data owners and other sensitive information, yet still produce data that the data miner finds useful. This is the main need of the project. For that purpose, anonymization approach, reconstruction and modification approaches are useful.



- **Data owner:** In this, owner gives information by using registration field and create login and fill important data for analysis purpose.
- **Data Analyst:** In this, analyst login allows analyst to view data and performs various operations on it. So that privacy can be maintained. Hence by applying anatomization, slicing and hiding rules, new data generation takes place.
- **Data Consumer:** In this, consumer means third party user creates his login and works as a data miner. He will able to view data or mined data which he fined useful. Here mining techniques and encryption is used for unwanted disclosure of data.

Implementation:- Implementation involves the environment in which system is implemented and overall system development. Overall system development requires suitable environment and proper resources for successful completion. Following algorithms are used to perform attribute and tuple partitioning for slicing the data. [2]

1] Algorithm:

Input : Sensitive data{S_1…Sm},the number of cluster R
Output : Multiple sensitivity attribute tables
Procedure:
Begin
Describe multiple sub-set { S_1…S_M} from the sensitive datasets.
Repeat 3 for m=1 to N.
In each sub-set, let the middle point be the initial centroid
For each sensitive attribute calculate the nearest centroids and assign to nearby cluster.
Choose minimal of minimal distance from the cluster the cluster's center.
Repeat the calculation for the dataset S for R clusters.
Merge the two nearest clusters into a cluster.
Recalculate the new cluster center for the collective cluster until the number of clusters is reduces into R.
For i-1 to R
Release tables ST_1……ST_R
End
End

2] Algorithm:-

Input : Data set QIT, the parameter K
Output : Sliced QIT
Procedure :
Begin
Q-QIT, Sliced bucket SB = ø
While Q is not empty {
Remove the first bucket into completely two different buckets using firefly algorithm
Check the tuple from Q
Set the objective function for Q
Compute the Intensity function by the objective function
Find attractiveness by the Minimum distance between the tuples
Using the intensity and attractiveness from the bucket Q
Check K anonymity
Q=Q U {B1, B2}
Else SB= SB U {B}
Return SB
}
End while
End

1] Flow of system development:-

Flow of system development first starts with data owner sign-up where registration is done. Then data owner is able to login and filling needed information. Then collected information is followed by data analyst. He logged in for analysis where all data of data owners are displaying. Then anatomization technique is applied. Then it is followed by slicing technique where data is sliced into buckets by using attribute partitioning, tuple partitioning and column generalization. After that by using rule hiding method data will be hidden and secured. At the time of analysis and viewing of information third party that is consumer will able to see the data by doing user login and after that he will able to see data by giving key and support, so that only wanted data will be display. In such a way flow of system development is displayed here.

- In Fig, firstly original data is displayed.

ID	Name	Resting BP	Hypertension	Severity Index	Treatment	Pulse Rate	Medicine	Education	Profession	Salary	Address	PinCode	Gender
1	Tanaya	84	108	3	Diabets	72	Diafix	B.E	Manager	58000	Pune	41108038	Female
2	Madhura	97	95	1	BP	73	Lisinopril o	M.E	Lecturer	45000	Mumbai	400007	Female
3	Sneha	112	90	3	Fewer	71	Nice	MS	Project Manager	80000	Kotharud	41108038	Female
4	Minak	100	111	2	Cold	72	Synus77	B.E	Bank	560	Pune	41108	Fem

	shi								Manager	00		038	ale
5	Amit	88	140	5	Cancer	70	Chemotherapy	BA	Teacher	36000	Jalgaon	425001	Male

- After that, anatomization technique is applying on original data and generate anoymized data. For example in anonymized data table, resting BP and hypertension are combining to generate new column1. Similarly severity index and treatment generates column2, pulse rate and medicine are in column3, education, profession and salary are in column4, and finally address, pin code and gender are in column3.

ID	Name	Colum1	Column2	Column3	Column4	Column5
5	Amit	88,140	5,Cancer	70,Chemotherapy	BA,Teacher,36000	Jalgaon, 425001, Male
1	Tanaya	84,108	3,Diabets	72,Diafix	B.E,Manager,58000	Pune,41108038, Female
3	Sheha	112,90	3,Fewer	71,Nice	MS,Project Manager,80000	Kotharud,41108038, Female
4	Minakshi	100,111	2,Cold	72,Synus77	B.E,Bank Manager,56000	Pune,41108038, Female
2	Madhura	97,95	1,BP	73,Lisinoprilo	M.E,Lecturer,45000	Mumbai,400007, Female

- After that slicing is applying on anonymized data to generate sliced data. In this sliced data, column1 is generated by combining resting BP, hypertension and pulse rate. Similarly severity index, treatment and medicine in column2. Also education, profession and salary are in column3. And remaining columns which are insensitive data are removing in this stage.

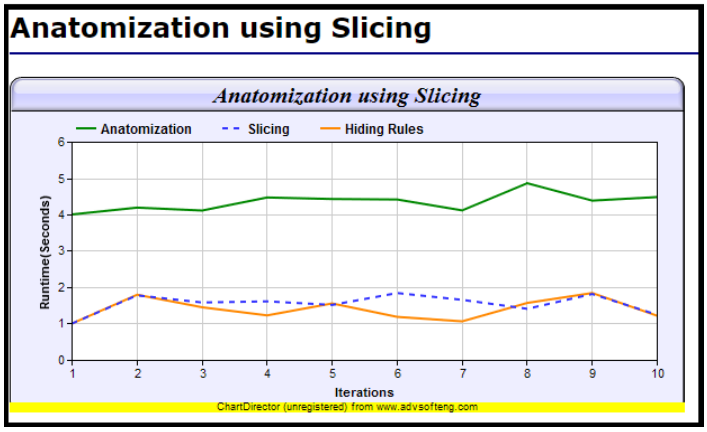
ID	Name	Column1	Column2	Column3
5	Amit	88,140,70	5,Cancer,Chemotherapy	BA,Teacher,36000
1	Tanaya	84,108,72	3,Diabets,Diafix	B.E,Manager,58000
3	Sneha	112,90,71	3,Fewer,Nice	MS,Project Manager,80000
4	Minakshi	100,111,72	2,Cold,Synus77	B.E,Bank Manager,56000
2	Madhura	97,95,73	1,BP,Lisinoprilo	M.E,Lecturer,45000

- Hiding data is generating by applying hiding rules. In this technique, attributes like treatment, medicine and salary are hiding by using asterisk (*) which are highly sensitive data. In such a way, various techniques are applying on original data, so that privacy risk can be eliminated.

ID	Name	Column1	Column2	Column3
5	Amit	88,82,70	5,C*****,C**********	BA,Teacher,3****
1	Tanaya	84,80,72	3,D******,D*****	B.E,Manager,5****
3	Sneha	112,90,71	3,F****,N***	MS,Project Manager,8****
4	Minakshi	100,85,72	2,C***,S******	B.E,Bank Manager,5****
2	Madhura	97,78,73	1,B*,L**********	M.E,Lecturer,4****

The next operation is modification and reconstruction. In this, data mining techniques are used for unwanted discovery of data. So that firstly 'Naïve Byes' theorem' is applied on the data and creates frequent sets of data according to the priorities and performs data mining. Also the generated output not reveals the data immediately as it is in encrypted form and hence ask the data consumer for a unique key and decrypted data gives exact mining result. This operation is performed by using 'data encryption standard algorithm'.

Result and Analysis: In this section, results are generated according to the time of execution of various techniques such as anatomization, slicing and then hiding rules. Hence graph is generated according to the time of execution.



Anatomization with slicing approach reduces the execution time of each iterations. The anatomization approach eliminates generalization and reduces the execution time by direct release of the QI attributes. The proposed work shows the attributes partitioning into columns that reduce the necessity for the reassignment of the data point multiple number of times during each iteration. This can efficiently assist in enhancing the clustering speed and thereby reduce the complexity that is seen in the computation. Additionally, in the proposed work, the single cluster group name is only utilized for referring to several datasets in the samples which also minimizes the complexity of the work. This way, the proposed technique consumes lesser execution time and utility by a significant factor for any number of SA in a patients' record. In first perspective, anatomization technique shows it's time of execution by dissociating the QI and SA. So that QI and SA have no overlap. In second perspective, enhanced slicing approach shows it's time by measuring the time required to generate sliced data by attribute and tuple partitioning. In the third perspective, hiding data shows it's execution.

Conclusion:-

Proposed work introduces anonymization technique that is anatomization with slicing which preserves the multiple sensitive attributes and improve the utility of the data. This method used to operate for any number of sensitive attribute and also avoids too much information loss.

Future Scope:-

In future, the anatomization with slicing algorithm can be applied simultaneously to both quasi-identifiers and sensitive attributes to reduce the time further through increased processor speed and memory. In future, continually expand scope to keep pace with the broadening view of privacy issues related to data mining.

Abbreviations:-

QI Quasi-Identifier
SA Sensitive Attributes
ST Sensitive Tables
QIT Quasi Identifier Table
SB Sliced Buckets

References:-

[1] Lei Xu, Chunxiao Jiang, Yan Chen, Jian Wang and Yong Ren, " A Framework for Categorizing and Applying Privacy-Preservation Techniques in Big Data Mining". Feb.2016.
[2] V. Shyamala Susan and T. Christopher, "Anatomisation with slicing: a new privacy preservation approach for multiple sensitive attributes". 2016.
[3] Preet Chandan Kaur, Tushar Ghorpade and Vanita Mane, "Analysis of Data Security by using Anonymization Techniques." Cloud System and Big Data Engineering, IEEE, 6th international conference 2016.
[4] Abid Mehmood, Iynkaran Natgunanathan, Yong Xiang, Guang Hua and Song Guo, "Protection of Big Data Privacy", IEEE Access 2016.
[5] Kinjal Parmar and Vinita Shah "A Review on Data Anonymization in Privacy Preserving Data Mining" Vol. 5, Issue 2, February 2016.
[6] Zang J. Molloy, Li N , "Slicing A New Approach for Privacy Preserving Data Publishing" 2016.
[7] S.Renuka Devi, A.C. Sumathi, "Enhancement of Privacy Preserving Technique using Slicing with Entity Resolution" Volume 2 Issue 3, May June 2016.