International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)
Web Site: www.ijettcs.org Email: editor@ijettcs.org
Volume 5, Issue 5, September - October 2016 **ISSN 2278-6856**

Preserving Privacy for Sensitive Data Items by Utilizing Data Mining Techniques

[1]M Ramesh, [2]Y Naga Sowmya, [3]T Lakshmi Surekha

[1] Assistant Professor Department of Information Technology
VR Siddhartha Engineering College, Vijayawada, Andhra Pradesh, India

[2]Department of Information Technology
VR Siddhartha Engineering College, Vijayawada, Andhra Pradesh, India

[3]Assistant Professor Department of Computer Science and Technology
VR Siddhartha Engineering College, Vijayawada, Andhra Pradesh, India

Abstract
The enhancement of data mining technologies leads to the threat for privacy of individual's private information. In recent years Privacy Preserving Data Mining (PPDM) topic in data mining research is studied extensively. The motive of PPDM is to maintain data secure by performing data mining algorithms effectively. Current investigations of PPDM basically approaches to the very proficient method to lessen the protection attacks brought by data mining operations, while actually, undesirable revelation of sensitive data may also happen during the time spent data collecting, data distributing, and data (i.e., the information mining results) conveying. In this paper, mainly concentrate about the privacy threatening issues that were detected with information mining extracted from a huge data repositories and research different methodologies that can ensure sensitive data. The knowledge derived from their local data repositories is insufficient to meet their projected outcomes. Hence there exists a need for sharing data for effective data mining and better analysis. While performing data analysis there exists a chance for intruder to know the individuals sensitive information.

I. INTRODUCTION

Data mining was created to give devices to naturally/brilliantly change huge information learning applicable to users.

The separated information,communicated as affiliation guidelines, choice trees or groups, grants finding designs/regularities covered in information yet intended to encourage basic leadership. This knowledge disclosure process returns sensitive data about individuals, bargaining their entitlement to protection. Data mining methods additionally uncover basic data about business, trading off free rivalry, thus exposures of private/individual data ought to be avoided not withstanding learning considered delicate in a given connection. Hence, research was devoted to addressing privacy preservation in data mining bringing about numerous data mining procedures which included privacy protection systems taking into account different methodologies. Different disinfection procedures were proposed to shroud sensitive items/designs in view of expelling saved data/embedding's clamor in information. Privacy preserving classification techniques keep mineworkers from constructing classifiers capable of predicting sensitive data. Also, recently proposed privacy preserving clustering techniques distort sensitive numerical attributes but preserve general features for cluster analysis. Data mining needs correct input for meaningful results, but privacy concerns influence users to provide wrong information. To preserve client privacy in data mining procedures, various random perturbation of data records based techniques were proposed. Randomization/Distortion are two methods that preserve privacy. Randomization modifies transactions through replacing some items with non-existing items and also through the addition of fake items to ensure privacy preservation. Distortion operates on a transaction database through probabilistically changing items in every transaction.

II.BACKGROUND

Privacy Preserving Data Mining Statement

Privacy Preserving Data mining Analysis is blending the data of different users and conceal user's private information.

2.1. Problem Statement

Stipulation of a comprehensible however supported methodology for early analysis of privacy preserving in the milieu of component based software advancement, keeping in mind the end goal to assess and contrast and each and every one of the Techniques in a general stage and to devise, develop and execute functionalities like a

framework that can be easily understand by user, portability etc.

2.2. Classification of Privacy Preserving Techniques. Many methods have been developed to preserve privacy using data mining techniques. They were classified based on the following dimensions:

- Data distribution
- Data modification
- Data mining algorithm
- Data or rule hiding
- Privacy preservation

The first dimension alludes to the sharing of information. A portion of the methodologies have been produced for brought together information. The second dimension refers to the alteration of original information. All in all, information alteration suggests changing the first values with copy values when information is should be discharged in broad daylight.

- Perturbation, which is consummate by rotating the estimation of a quality with another worth (i.e., changing a 1-worth to a 0-esteem, or including commotion).
- Blocking is changing the estimation of a property with a "?".
- Aggregation or consolidating which is the mix of a few values into a coarser classification.
- Swapping alludes to exchanging the estimations of individual records.
- Sampling, which alludes to discharging information for just a specimen of a populace?

The third measurement infers to the data mining algorithm, for which the information adjustment is occuring. This is really something that is not known before, but rather it encourages the examination and model of the data concealing algorithm.

The fourth dimension implies suggests to which sort of data ought to be kept cover up whether crude data or accumulated data. The level of multifaceted nature raises for covering up collected data as guidelines, and hence, for the most part heuristics have been created.

The last measurement, which is the most basic, insinuates the security defending strategy used for the specific modification of the data. Particular modification is needed in order to reach quality utility for the modified data given that the security is not imperiled.

For this reason these are the techniques that have been applied:

- Heuristic-based techniques like adaptive modification which changes specific values that minimize the loss in utility rather than all accessible values.
- Cryptography- based strategies like secure multiparty calculation where a computation is secure toward the end of the computation, no gathering knows anything with the exception of its own information and the outcomes.
- Reconstruction-based procedures where the first appropriation of the information is reproduced from randomized information.

III. RELATED WORK

3.1 Hiding Association Rules by Using Confidence and Support

In this paper, proposed few guidelines for concealing affectability by morphing the support and the certainty of the association rule or frequent item set as data mining for the most part makes do with era of association rules. To hide an association rule a new method of 'not adjusting the support' for the delicate item(s) has been proposed in this process.

Advantages

- First advantage of proposed algorithm is that support for the sensitive item will not be changed. Rather, just the position of the sensitive item set is changed.
- This proposed method uses various approach for changing the database trades so that the certainty of the sensitive principles can be lessened however not modifying the support of the delicate item.

Disadvantage

- One of the main disadvantages of the existing approaches is that the approach in tries to conceal all without checking on the off chance that a portion of the standards could be pruned after change of couple of exchanges from the arrangement of all exchanges.

3.2 Privacy Preserving Clustering By Data Transformation

When individual's information is needed to share for clustering, providing security is a complex issue. The test was the way to ensure the fundamental information values subjected to clustering without risking the comparability between articles under investigation. Stanley R. M. Oliveira, and Osmar R. Zaiane [5] returned to a group of geometric information transformation strategies that contort numerical properties by scaling, revolutions, interpretations or by the blend of every single above change. This strategy was intended to determine privacy-preserving clustering, in context where data owners must meet privacy requirements as well as guarantee valid clustering results. Authors also provided a particularized, wide and advanced picture of methods for privacy-preserving clustering by data transformation.

Advantages

- The geometric data transformation methods (GDTMs) that distort confidential numerical ascribes in order to reach privacy protection in clustering analysis.
- By utilizing data transformation own tools can be used by owners so that the condition for privacy has to be applied earlier the mining process on the data.
- Data owners must meet security prerequisites as well as certification legitimate clustering results.

Disadvantages

- One major disadvantage is that the preserving privacy for individual's information when data is shared for clustering is very complex.
- The protection of the underlying data values subjected to clustering without imperiling the probability between objects under analysis will be complicated to complete.

IV. PROPOSED SYSTEM

The proposed work uses k-anonymity's granularity reduction technique for privacy preserving during data mining. GA optimizes feature selection.

4.1. GENETIC ALGORITHM

GA is a family of advancement motivated computational models. For a particular problem on chromosome-like data structure these models conceal a potential solution by applying recombination operators to these structures to safeguard critical information. GAs are viewed as function optimizers, however issue reach to which GA is applicable is broad.

4.1.1 GA Operators

A simple GA involves 3 operators: selection, crossover (single point), and mutation.

4.1.1.1 Selection: This operator chooses chromosomes for reproduction in a population. The fitter the chromosome, the more times it will be chosen to reproduce in equation (3) [4].

$$P_s(i)=f(i)/\sum^{n}_{j=1}f(j)$$

Where $p_s(i)$ and f(i) are selection and fitness value probabilities for ith chromosome respectively. Roulette wheel selection is implemented as follows:

1. For each individual in a population fitness f(i) is to be evaluated.
2. Compute likelihood (slot size), p(i), of selecting each population member as in equation:
   $$P_i=f_i / \sum^{n}_{j=1} f(j)$$
   Where n is population size.
3. For each individual calculate cumulative probability, qi, as in equation:
   $qi = \sum^{i}_{j=1} P_j$
4. A fixed random number should be produce, r (0, 1].
5. If $r < q1$ then first chromosome x1 must be selected, otherwise individual xi will be chosen such that $q_{i-1} < r < q_i$
6. To create n candidates in mating pool steps 4–5 need to be repeated n times.

4.1.2 Crossover operates individually. A crossover point is arbitrarily decided for 2 randomly chosen individuals (parents). The point is between 2 bits isolating every individual into left and right sections. Crossover swaps left (or right) segment of both individuals. A crossover example: consider two parents:

Parent 1: 1010101010
Parent 2: **1000010000**

Every new child receives one portion of the parent's bits if crossover point randomly occurs after fifth bit.

Child 1: 10101**10000**
Child 2: **10000**01010

4.1.3 Mutations are global searches. A mutation likelihood is foreordained before starting the algorithm and applied to every individual bit of each posterity chromosome for determining if it is to be inverted [5].

4.2. PPDM

The privacy-preserving data mining (PPDM) has transformed into a crucial issue starting late. Tzung-Pei Hong et al. [8] proposed a paper, a greedy-based methodology for concealing delicate item sets by embedding sham exchanges. That registers the superlative count of exchanges to be embedded into the original database for altogether concealing private item sets. Trial results were likewise performed to survey the execution of this approach. As of late, the wide accessibility of individual information has made the issue of Privacy Preserving Data Mining a key one.

The expanding ability to track and collect bulk information with the utilization of current equipment innovation has taken to an enthusiasm for the improvement of data mining calculations which save client security. Various methods have as of late been proposed for protection safeguarding data mining of multidimensional information records.

Advantages

- PPDM is exceptionally invaluable being developed of different data mining strategies.

- It permits sharing of substantial measure of security touchy information for examination purposes.
- It has a capacity to track and accumulate immense measure of information with the use of current equipment innovation.

Disadvantage

- One of the huge issue of security saving information mining is the inexhaustible availability of individual information.

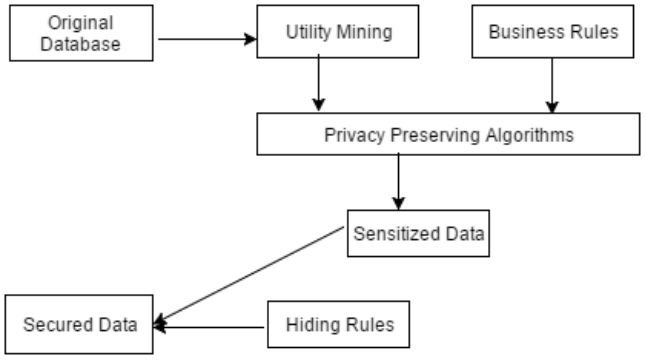


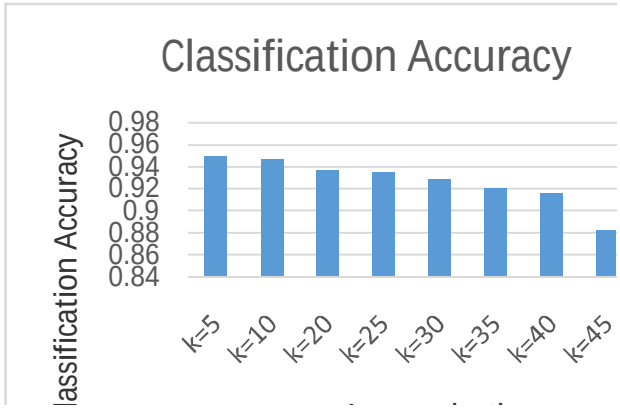
Fig. 1: Block diagram of Privacy Preserving Data Mining Technique

V. RESULTS AND DISCUSSION

Mushroom data set is used, and proposed algorithm tested. Results reveal the algorithm being capable of finding an optimal/near optimal solution for varying k-anonymity model levels. Performance metrics used include average accuracy, precision and recall.

Table1 Shows the Results Obtained by the Genetic Algorithm Method

Anonymization	Classification
No Anonymization	0.958271787
k=5	0.954455933
k=10	0.947193501
k=20	0.937715411
K=25	0.93525357
K=30	0.929837518
K=35	0.921221073
K=40	0.916789759
K=45	0.882912899
K=50	0.91211226



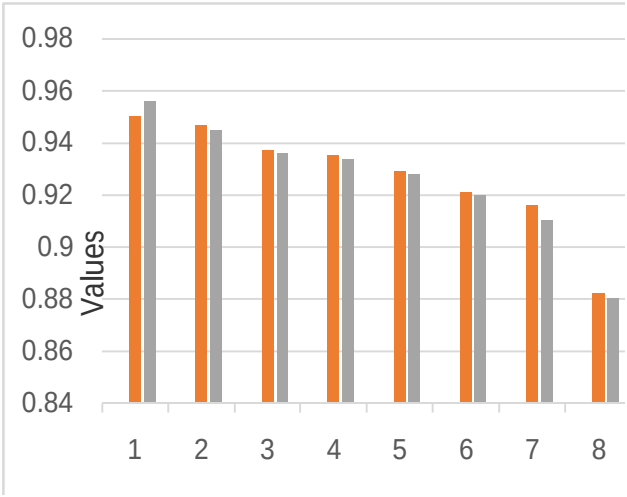
(Graph 1)

Graph1 defines the plot between classification accuracy to anonymization.

It is seen from graph 1 that classification accuracy declines with increase in k-levels. Between no anonymization and k==30 it decreases by 2.96% and between k=5 and k=50 it decreases by 4.43%.

Table 2: Precision and Recall Achieved

Anonymization	Classification	Recall
No Anonymization	0.961365139	0.956999557
k=5	0.957571805	0.953156209
k=10	0.950324439	0.945861951
k=20	0.94034287	0.936437867
K=25	0.937621654	0.934028091
K=30	0.930809935	0.928840408
K=35	0.92196315	0.920388942
K=40	0.917708964	0.915892426
K=45	0.886578431	0.882385574
K=50	0.912653555	0.911294989



Graph 2 defines the plot between classification accuracy, anonymization and recall.

VI. CONCLUSION

The above analyzed privacy preserving information mining strategies are strikingly great, yet there is dependably degree for more upgrades. This study paper on PPDM can be useful for finding the escape clauses and inconveniences of existing data mining strategies. This overview guarantees capable protection safeguarding of information. The use of existing algorithm works towards the heading to diminish the effect of PPDM on the source database. A similar concentrate every one of these frameworks would help in building another framework that consolidates every one of the favorable circumstances and conquers the disadvantages of these frameworks.

REFERENCES

[1]. Yehuda Lindell and Benny Pinkas, "Secure Multiparty Computation for Privacy-Preserving Data Mining," The Journal of Privacy and Con¯dentiality, vol. 1, no. 1, pp. 59-98, 2009.

[2]. Marina Blanton, "Achieving Full Security in Privacy-Preserving Data Mining," In Proc. of the 2011 ieee third international conference on social computing (socialcom) Privacy, security, risk and trust (passat), Dame, IN, pp. 925-934, Oct 2011.

[3]. Bin Yang, Hiroshi Nakagawa, Issei Sato, and Jun Sakuma, "Collusion-Resistant Privacy-Preserving Data Mining," In Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), pp. 483–492, 2010.

[4]. Li Liu, Murat Kantarcioglu, and Bhavani Thuraisingham, "The applicability of the perturbation based privacy preserving data mining for real-world data," Journal of Data & Knowledge Engineering, vol. 65, pp. 5–21, 2008

[5]. Stanley R. M. Oliveira, and Osmar R. Zaiane, "Revisiting Privacy Preserving Clustering by Data Transformation," Journal of Information and Data Management, vol. 1, no. 1, 2010.

[6]. Yuhong Guo, 2007, "Reconstruction-Based Association Rule Hiding", Proceedings of SIGMOD2007 Ph.D. Workshop on Innovative Database Research 2007(IDAR2007), 51-56.

[7]. Dr. Duraiswamy. K, Dr. Manjula. D, and Maheswari. N "A New Approach to Sensitive Rule Hiding", ccsenet journal, vol 1, No. 3, August, 107-111.

[8]. Mohammad Naderi Dehkordi, Kambiz Badie, Ahmad Khadem Zadeh, "A Novel Method for Privacy Preserving in Association Rule Mining Based on Genetic Algorithms", Journal of software, vol. 4, no. 6, August 2009 .

[9]. T.Berberoglu and M. Kaya, "Hiding Fuzzy Association Rules in Quantitative Data", The 3rd International Conference on Grid and Pervasive Computing Workshops, May 2008, pp. 387-392