Improved Accuracy-Constrained Privacy Preserving Access Control Mechanism in Data Mining

Seena Patel[1], Ms. Ripal Patel[2]
Department of Computer Engineering
[1]PG Scholar,Silver Oak College of Engineering & Technology
[2]Assistant Professor,Silver Oak College of Engineering & Technology

Abstract*-The paper demonstrate on accuracy constrained privacy-preserving access control mechanism for relation data framework with multilevel anonymization techniques. Access control policy which define selection predicate on sensitive data and privacy requirement deals with anonymity. As privacy protection mechanism (PPM) provides less privacy protection and the data is shared so the user should compromise the with the privacy of data. The goal of the paper is to provide more security to the sensitive data along with minimal level of precision. The concept of accuracy constraints for permissions can be applied to any privacy-preserving security policy. Our goal is to solve problem of K-anonymity algorithm and provide solution by improving l-diversity algorithm.*

Keywords-Access control, privacy, k-anonymity, l-diversity, query evaluation

I. INTRODUCTION

Organizations collect and analyze consumer data to improve their services. Access Control Mechanisms (ACM) are used to ensure that only authorized information is available to users. However, sensitive information can still be misused by authorized users to compromise the privacy of consumers[1]. So we have to protect sensitive information from the misuse. Privacy preserving mechanism used to protect sensitive data. Organizations implement access control mechanism to assure that only sensitive information is available to authorized users. Sometimes confidential information is misused by authorized users to adjust the privacy of the customer. Organizations collect and analyze the data to improve the services [2].After removing the primary keys from the database of particular users ,the sensitive data may suffer from linking attacks from authorized users [6]. To improve the protection against identity discloser and support the privacy policy ,the concept of privacy preservation of sensitive data is introduced by satisfying some privacy requirements [7]. Every database have to maintain the sensitive information from privacy mechanisms, then also there is possibility that they suffer from linking attacks from authorized users. This problem has been studied in micro data publishing and privacy definitions like k-anonymity [6], l-diversity [2] and variance diversity [2].

In literature survey [1] they proposed an accuracy-constrained privacy-preserving access control framework for relational data. The framework is a combination of access control and privacy protection mechanisms. The access control mechanism allows only authorized query predicates on sensitive data. The privacy preserving module anonymizes the data to meet privacy requirements and imprecision constraints on predicates set by the access control mechanism. They formulate this interaction as the problem of k-anonymous Partitioning with Imprecision Bounds (k-PIB). They give hardness results for the k-PIB problem and present heuristics for partitioning the data to the satisfy the privacy constraints and the imprecision bounds. For this current work, they assumed static access control and relational data model.

Two types of information disclosures i) identity disclosure and ii) attribute disclosure. Identity disclosure occurs when an individual is linked to a particular record in the released table. Attribute disclosure happens when the new information about some individuals is revealed, i.e., the released data makes it possible to infer the characteristics of an individual more accurately than it would be possible before releasing the data. To counter identity disclosure, Samarati and Sweeney proposed a privacy model called k-anonymity [8]. The model works by ensuring that each record of a table is identical to at least k – 1 other record with respect to quasi-identifiers (QI), which could be potentially used to identify individuals by linking these attributes to external data sets [9]. Although k-anonymity protects against identity disclosure, it is insufficient to prevent attribute disclosure. So, l-diversity model [10] was proposed.

Intuitively, l-diversity means that an adversary needs l-1 pieces of background knowledge to eliminate l-1 possible values of a sensitive attribute in order to breach privacy. Specifically, if a table is l-diverse, in each QI group, at most 1/l of the tuples possesses the most frequent sensitive value. However, depending on the nature of the sensitive attributes, even these enhanced properties still permit the information to be disclosed. The existing methods for l-diversity only consider l "well represent" sensitive value, but omit the size of every QI-group, so the information loss of the published data sets is much larger, that lead to released data set useless. Most of k-anonymity and l-diversity methods rely on generalizations to preserve privacy, that is, attribute values are replaced with less specific information (for example, state may be replaced with region and age may be replaced with age range). Thereby, the utility of these anonymized data sets should be taken into account when constructing the privacy protection [11]. The less the information distortion in the anonymity protected table makes, the larger the table usability is. Therefore, an anonymity model must minimize the information distortion in terms of its original table. Unfortunately, the computational complexity of finding optimal solutions for both k-anonymity and l-diversity models are NP-hard [12-14].

The concept of privacy-preservation for sensitive data uses anonymization techniques. Anonymization algorithm uses suppression or generalization of records to satisfy the privacy requirement with minimal distortion of micro data. This techniques can be used to ensure security and privacy of the sensitive information. The privacy is achieved at the cost of accuracy and imprecision is introduced in the authorized information under an access control policy [1].

II. BACKGROUND

Let DT be the initial data table and PT be th released anonymous table. The attributes in an original unprotected dataset DT can be classified in four categories, which are identifiers, quasi-identifiers, sensitive attributes, non-sensitive attributes. In what follows we assume that the quasi-identifier specified by the administrator based on the background knowledge, and the values of the sensitive attributes are not available from any external source.

Definition 1(Identifier) Attributes, e.g., name and social security that can uniquely identify an individual. These attributes are completely removed from the anonymized relation.

Definition 2(Quasi-identifier (QI)) Attributes, e.g., gender, zip code, birth date, that can potentially identify an individual based on other information available to an adversary. QI attributes are generalized to satisfy the anonymity requirements.

Definition 3(Sensitive attribute) Attributes, e.g., disease or salary, that if associated to a unique individual will cause a privacy breach.

Definition 4(k-anonymity) PT is said to satisfy k-anonymity if and only if each combination of quasi-identifier attributes (QI-group) in PT occurs at least k times.

For example, patient diagnosis records without conducting the k-anonymity model is shown in Table 1, where the attributes Age, Country, and Zip Code are regarded as quasi identifiers. If the hospital simply publishes the table to other organizations directly, those organizations might extract patients' disease histories by joining this table with other tables. By contrast, Table 2 is a 4-anonymity version of the original table.

Table 1: Original Data

ID	**Age**	**Country**	**Zip code**	**Disease**
1	25	USA	480120	HIV
2	27	Canada	421020	Cancer
3	22	China	446085	Asthma
4	41	USA	480126	Flu
5	44	India	380061	Flu
6	32	Canada	421006	HIV
7	36	India	380025	Cancer

Table 2: 4-Anonymous Data

ID	**Age**	**Country**	**Zip code**	**Disease**
1	15-30	America	4*****	HIV
2	20-30	America	4*****	Cancer
3	20-30	Asia	4*****	Asthma
4	>40	America	4*****	Flu
5	>40	Asia	3*****	Flu
6	30-40	America	4*****	HIV
7	30-40	Asia	3*****	Cancer

The *k*-anonymity property ensures protection against identity disclosure. However, it does not protect the data against attribute disclosure, which occurs when the intruder finds a target entity.

Definition 5(*l*-Diversity) A QI-group satisfies *l*-diversity if there are at least *l* distinct values for the sensitive attribute. A modified table satisfies *l*-diversity if every cluster of the table satisfies *l*-diversity.

For instance, Table 2 is also 2-diverse because, in each cluster, at most 50% of the tuples have the same ***Disease*** value. Although the *l*-diversity principle represents an important step beyond *k*-anonymity in protecting sensitive attribute disclosures, it still has some shortcomings.

Definition 6(Partition) A Partition consists of several subsets of DT, such that each tuple in DT belongs to exactly one subset.

III. METHODOLOGY

In the context of k-anonymization problems, a database is a table with n rows and m columns. Each row of the table represents a record relating to a specific member of a population and the entries in the various rows need not be unique. The values in the various columns are the values of attributes associated with the members of the population.

There are two common methods for achieving k-anonymity for some value of k.

.

- Generalization
- Suppression

Generalization

Generalization consists in replacing attribute values with a generalized version of them. Generalization should be applied on the data which are repeated in nature. Generalization can be applied at the level of single cell (substituting the cell value with a generalized version of it) or at the level of attribute (generalizing all the cells in the corresponding column).

Table 3: 2-anonymized Generalization Table

	QI$_1$	QI$_2$	S$_1$
ID	Age	Zip	Disease
1	5	15	Flu
2	12	25	Fever
3	22	22	Cancer
4	35	35	Diarrhea
5	40	26	Flu
6	28	40	Fever

(a) Sensitive table

	QI$_1$	QI$_2$	S$_1$
ID	Age	Zip	Disease
1	0-20	10-30	Flu
2	0-20	20-40	Fever
3	20-30	20-40	Cancer
4	30-40	30-50	Diarrhea
5	30-40	20-40	Flu
6	20-30	30-50	Fever

(b) 2-anonymous table

Table (a) contain original data value and does not satisfy k-anonymity because knowing the age and zip code of a person allows associating a disease to that person. So prevent the k-anonymity linking attack the Generalization method is applied. In this technique the sensitive value is replaced with some value duration. In table (b) the ID attribute is removed in the anonymized table and is shown only for identification of tuples. Here, for any combination of selection predicates on the zip code and age attributes, there are at least two tuples in each equivalence class.

Suppression

Suppression consists in protecting sensitive information by removing it. Suppression can be applied at the level of single cell, entire tuple, or entire column. It allows to reduce the amount of generalization to be enforced to achieve k-anonymity. Intuitively, if a limited number of outliers would force a large amount of generalization to satisfy a k-anonymity constraint, then such outliers can be removed from the table thus satisfying the k-anonymity with less generalization (an therefore, reducing the loss of information). Suppression applied when data is large in nature. It allows to reduce the amount of generalization to be enforced to achieve k-anonymity. In this method, mask the Quasi-Identifiers value using a special symbol like *.

Table 4: Patients detail

	Zip code	Age	Nationality	Condition
1	13053	28	Russian	Heart disease
2	13068	29	American	Cancer
3	14853	50	Indian	Viral infection
4	14850	47	American	Cancer
5	13053	31	Indian	Cancer
6	13068	36	Japanese	Heart disease
7	13068	35	Russian	Viral infection

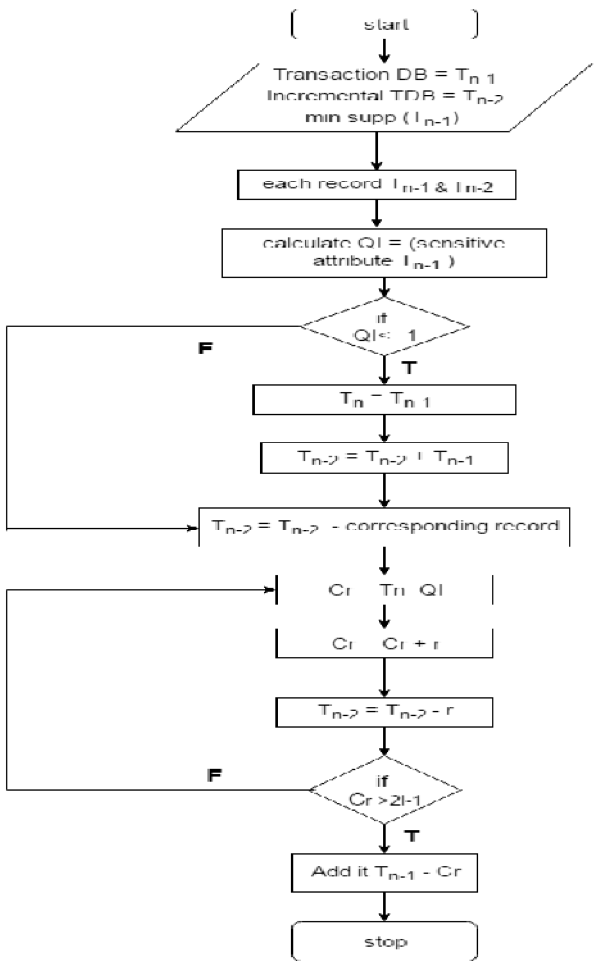
Table 5: 3-anonymous patient data

	Zip code	Age	Nationality	Condition
1	1****	<30	*	Heart disease
2	1****	<30	*	Cancer
3	1****	>=40	*	Viral infection
4	1****	>=40	*	Cancer
5	1****	3*	*	Cancer
6	1****	3*	*	Heart disease
7	1****	3*	*	Viral infection

IV. PROPOSED WORK

```
Proposed Algorithm
Input: Transaction DB, Incremental TDB, Qi, Corresponding record cr
Output: Tn-1
1) Read DB=Tn-1 and TDB=Tn-2
2) Calculate Qi=(sensitive attribute Tn-1)
3) if Qi<=1
4) takeTn=Tn-1
5) then, Tn-2 = Tn-2 + Tn-1
6) else value of Tn-2 = Tn-2 – cr
7) filter records as per attribute, cr = Tn-Qi
8) thencr = cr + r
9) add Tn-2 = Tn-2 – r
10) ifcr>2l-1
11) add it in Tn-1 – cr                    //new Tn-1 generate
12) else go to step 7
```

Proposed Framework



V. EXPERIMENTS AND RESULT

1) Comparison for 100 attributes

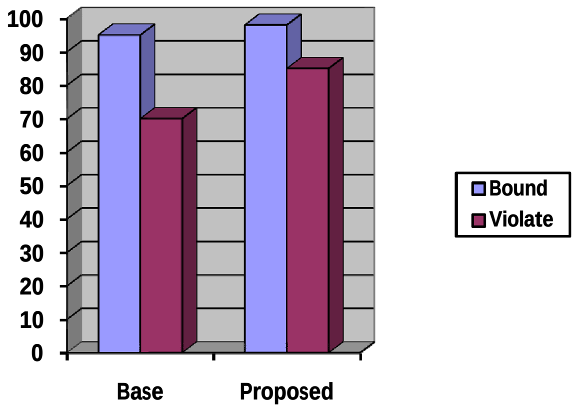


Fig 1: comparison for variance at 100 attributes

The experiments have been carried out on two data sets for the empirical evaluation of the proposed algorithm. The first data set is the Adult data set from UCI Machine Learing Repository[22] having 46k tuples and it is useful for k-anonymity research. The attributes in Adult data set are: Age, Work class, Education, Country, Education, Marital status, Occupation, Gender and etc.

The second data set is German Credit Data set also from UCI Machine Learning Repository having about 1.3 million tuples. The attributes are: Age, Occupation, Gender, sex, income, marital status, language, Birthdate and etc. Here, the comparison is between base algorithm and proposed algorithm. We took different range of attribute and compare with the base algorithm result. Here, we took 100 attributes. Among them there are 95 attribute as a bound. From that the result is 70 attributes with the help of base algorithm [1]. As well we took 98 attributes as bound and by using the proposed algorithm we get the 85 attributes as violate. As well as to ensure the result we took more data. For 200 attributes, by using base algorithm we get 94 attributes from 126 bounds. But by using the proposed algorithm we get 139 anonymized attributes from 148 attributes. So, it is clearly that by using the proposed algorithm we can get better result.

The existing literature is based on the imprecision bound for each query in a given query workload. In that for k-anonymity we have to check for all queries separately. It will take more time and the efficiency is decrease. We introduced proposed algorithm in which no of queries can be calculate within a time. There is no requirement to check *l*-Diversity and k-anonymity separately. The accuracy is better compared to previous algorithm.

2) Comparison for 200 Attributes

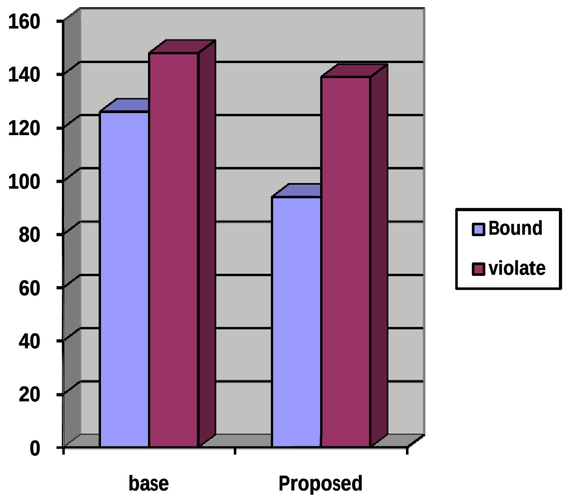


Fig 2: comparison at variance of 200 attributes

VI. CONCLUSION

Here we present the proposed algorithm for remove disadvantage of anonymity or diversity. Previous, there are many algorithms based on k-anonymity or l-diversity. But by using those algorithms, it will carry some disadvantages. So we try to merge the disadvantage of both anonymity and diversity within a proposed algorithm. And we try to solve the problem of both with proposed algorithm. Also the efficiency of the system is better comparative previous result.

REFERENCES

[1] Pervaiz, Z., Aref, W. G., Ghafoor, A., &Prabhu, N. (2014). Accuracy-constrained privacy-preserving access control mechanism for relational data. IEEE Transactions On Knowledge And Data Engineering, 26(4), 795-807.

[2] Suhasini Gurappa .Metri PG Student, CSE Dept Cambridge institute of technology, Bangalore ,India ,International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 3 Issue: 4

[3] Pratik Bhingardeve1, 2 Pune University, Smt. KashibaiNavale College of Engineering, Vadgaon (BK), Pune-411041, India, International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064

[4] Yang, G., Li, J., Zhang, S., & Yu, L. (2013, July). An enhanced l-diversity privacy preservation. In Fuzzy Systems and Knowledge Discovery (FSKD), 2013 10th International Conference on (pp. 1115-1120). IEEE.

[5] Ebin P.M, Brilley Batley. C, AMIE, Assistant Professor Department of Computer Science & Engineering, International Journal of Science and Research (IJSR) , India Online ISSN: 2319□7064

[6] Samarati, P. (2001). Protecting respondents identities in microdata release. IEEE transactions on Knowledge and Data Engineering, 13(6), 1010-1027.

[7] Rask, A., Rubin, D., & Neumann, B. (2005). Implementing row-and cell-level security in classified databases using SQL Server 2005. MS SQL Server Technical Center.

[8] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.

[9] Benjamin C. M. Fung, Ke Wang, Rui Chen, Philip S. Yu, "Privacy preserving data publishing: A survey of recent developments," ACM Computing Surveys, 2010, pp. 1-53.

[10] Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1), 3.

[11] Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., & Fu, A. W. C. (2006, August). Utility-based anonymization using local recoding. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 785-790). ACM.

[12] Meyerson, A., & Williams, R. (2004, June). On the complexity of optimal k-anonymity. In Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 223-228). ACM.

[13] Tian, H., & Zhang, W. (2011). Extending ℓ-diversity to generalize sensitive data. Data & Knowledge Engineering, 70(1), 101-126.

[14] Xiao, X., Yi, K., & Tao, Y. (2010, March). The hardness and approximation algorithms for l-diversity. In Proceedings of the 13th International Conference on Extending Database Technology (pp. 135-146). ACM.

[15] JaydipSen,Innovation Lab, Tata Consultancy Services, Kolkata, India, ‘Privacy Preserving Data Mining’, 2nd ICCCT, MNNIT, Allahabad Sep 16, 2011

[16] Spiliopoulou, M., &Faulstich, L. C. (1998, March). WUM: a tool for web utilization analysis. In International Workshop on the World Wide Web and Databases (pp. 184-203). Springer Berlin Heidelberg.

[17] K. LeFevre, D. DeWitt, and R. Ramakrishnan, “Workload Anonymization Techniques for Large-Scale Datasets,”ACMTrans. Database Systems, vol. 33, no. 3, pp. 1-47,2008.

[18] Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007). l-diversity: Privacy

beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD), 1(1), 3..

[19] Olumofin, F., & Goldberg, I. (2010). Preserving access privacy over large databases. Technical Report 33, University of Waterloo.

[20] Sweeney, L. (2002). k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05), 557-570.

[21] S.V.G.REDDY,Associate professor, ,Dept.of CSE, GIT, GITAM UNIVERSITY, ‘Introduction to Data Mining

[22] A Frank and A. Asuncion, “UCI Machine Learning Repository,” 2010