OPEN ACCESS Freely available online

International Journal of Biomedical Data Mining

Research Article

K-AMSOA: Privacy-Preserving Technique for Multiple Co-Related Sensitive Numeric Attributes using Dynamic Data Set

Nidhi M. Chourey*, **Bhakti M. Thombre**

Department of Computer Science and Engineering, Vikrant Institute of Technology and Management, Indore, India

ABSTRACT

Privacy Preservation of released data is a serious issue nowadays. If a dataset contains multiple correlated attributes then privacy concerns related to data set increases, because data publisher has to release data set in such a way that data is secure from disclosures as well as their co-relation is maintained as well. If attributes whose co-relations should be maintained contain a set of multiple numerical sensitive attributes then the privacy-preserving becomes a very tough task. So, there is a need for a new privacy preservation technique which is specially focused on this direction. This work is focused on privacy preservation for the dynamic dataset. Here, a dynamic data set means data publishers can add or remove attributes at the time of applying the privacy preservation technique and release results. Data set attributes may vary as per dataset release objective.

Keywords: Dynamic datasets; Clustering; Multiple sensitive numerical attributes; k-anonymity technique; Generalization; Attributes correlations

INTRODUCTION

There are several models already proposed during this area to stop data set against information disclosure and linking attack. But there's a requirement to propose new models that protect data set against membership disclosure within the presence of multiple co-related sensitive attributes are present.

Even when multiple sensitive numerical attributes are present in data set chances of proximity breach attack increase. Which may be a privacy threat specific to numerical sensitive attributes in data publication happens. To take care of co-relation among attributes the simplest way is to place these attributes during a single cell which is understood as overlapped attributes. But this system increases membership disclosure and proximity breach if co-related overlapped attributes contain sensitive numerical values.

Almost all proposed models are focused on privacy preservation techniques for static data set but several new challenges occur for dynamic data set because attackers can study all published versions and find some clue associated with actual sensitive values. Therefore, there is a significant requirement of the latest model that protects the dataset against any sort of co-related attributes for multiple published versions. the large amount of knowledge that is collected by various resources and used for a few kinds of deciding or knowledge discovery is mined by employing a data processing algorithm. There are various algorithms present for data processing task and people are often modified as per user requirements.

Data are of two types sensitive and non-sensitive. The samples of sensitive data are record and non-sensitive data contains the overall information ex: DOB, Zip code, and Gender because they need knowledge mining increases privacy problems with data have also increased. the most objective of knowledge mining is to seek out the hidden information which previously might not be visible but that information may contain sensitive information and should cause severe attacks.

So, the info publishers need to release information in such how that unauthorized person won't be ready to track the sensitive information. There are several techniques present for privacy-preserving data publishing named K-anonymity, perturbation, swapping, etc. These techniques are wont to convert data in such a format that unauthorized one that isn't intended to trace the info cannot find the

Correspondence to: Nidhi M. Chourey, Department of Computer Science and Engineering, Vikrant Institute of Technology and Management, Indore, India, E-mail: nvchourey@gmail.com

Received date: March 10, 2020; **Accepted date:** May 13, 2020; **Published date:** May 23, 2020

Citation: Chourey NM, Thombre BM (2020) K-AMSOA: Privacy-Preserving Technique for Multiple Co-Related Sensitive Numeric Attributes using Dynamic Data Set. Int J Biomed Data Min 9:134. doi: 10.4172/2090-4924.2020.9.134

first data also as an associated user can find the limited required data for his use.

There are several limitations and assumption for PPDP techniques as per studied papers:

- We assume that data publishers trust worthy and have full knowledge of the system
- Data publisher decide attribute type i.e. sensitive or non-sensitive
- The value of K is predefined
- Results are static data set i.e. result analysis performed on the stored or available data set. If the data set is dynamic new challenges will we explore?
- If multiple numerical sensitive attributes are present in data set chances of proximity breach attack increase
- If multiple co-related sensitive attributes are present then chances of membership disclosure high very tough

LITERATURE STUDY

Privacy means to hide some information which is private or belongs to individuals. for instance, India's current population is around 133.92 crores as per the 2017 survey and it's belonging to each Indian, therefore there's no got to hide this information with others. Rather than this cancer patient's name is sensitive and must be hidden. When data is released for a few government or semi-government use, the concept of privacy-preserving data publishing concept is available to focus. Privacy-preserving data publishing supported the definition of privacy, it ensures what information is sensitive within the original data and performs operations to guard it against direct or indirect disclosure [1].

Data publishers need to release datasets for public interest but in such how that sensitive information about individuals isn't disclosing. To stop these sorts of disclosure many privacy-preserving data publishing techniques are used. These techniques are wont to hide or protect sensitive attributes from the adversary.

There are three sorts of attributes that are present in any dataset named as Key attribute, sensitive attribute, and non-sensitive attribute [2].

- Key attributes or identifiers are wont to uniquely identify a private. Examples: Name, SSN, PAN [3]
- Some attributes contains some very personal information which individual wants to cover from others. Example: medical history, Salary, Increments [3]
- Those neither attributes which aren't a key nor sensitive attributes referred to as non-sensitive attributes

These sorts of attributes also are referred to as quasi-identifiers. Quasi-identifiers are easily available publicly so, results in linking attack. Example: DOB, Sex, Nationality, Zip Code.

When sensitive information is disclosed thanks to the linking of quasi-identifiers of public and personal tables. As quasi-identifiers are general values or attributes which are present in both datasets. When the adversary matches these quasi-identifiers values of both tables, sensitive values of the personal table is revealed [4]. This sort of linking is understood as a linking attack. Table 1 shows a public table and Table 2 contains a private table when both tables are linked it discloses that tom is affected by a heart condition.

Table 1: Public table.

S. No.	Name	DOB	Sex	Zip Code
1	Tom	1/21/76	Male	50023
2	Beena	4/13/86	Female	50023
3	Carry	2/28/76	Male	50002
4	Daniel	1/21/76	Male	50002
5	Esha	4/13/86	Female	53706
6	Sam	2/28/76	Female	53706

Table 2: Private table (hospital).

S. No.	DOB	Sex	Zip Code	Disease
1	1/21/76	Male	50023	Heart Issue
2	4/13/86	Female	50023	Hepatitis
3	2/28/76	Male	50002	Bronchitis

4	1/21/76	Male	50002	Broken Arm
5	4/13/86	Female	53706	Flu
6	2/28/76	Female	53706	Hang Nail

K-anonymity model and various techniques

K-anonymity model was proposed to regulate linking attacks. K-anonymous mining applied on data in two ways, one is before mining applied on private data, second is on mined data [5]. The definition of K-anonymity as per Samrati is "K-anonymity requires that if a mixture of values of quasi-identifying attributes appears within the table, then it appears with a minimum of k occurrence. In other words, if a tuple has K occurrences then any of its sub tuples must have a minimum of K-occurrences [6]. The word anonymity means a nameless state, where the identity of an object is hidden.

K-anonymity model emphasizes that every released record has a minimum of (K-1) other records within the release whose values remain the same over those fields belonging to external data. Therefore, every combination of values of a quasi-identifier in the table that satisfies k-anonymity contains a minimum of k-records sharing those values. K-anonymous model was proposed to guard private data, which are released for a few users from linking attack which causes the knowledge disclosure. There are three sorts of information disclosure present [7].

- When is linked to record published data as Identity disclosure [8]
- When sensitive information of is leaked linking attack, attribute disclosure [8]
- When additional sensitive information regarding is disclosed from set while trying for sensitive attribute, membership disclosure [8]

Generalization and Suppression are the hottest technique to realize k-anonymity where generalization emphasizes replacing the quasi-identifier value with a less specific value [9]. For example, postcode 405028 may be a quasi-identifier value. When generalization is applied thereon, the resultant value is going to be 40502. This sort of generalization is understood as Level-I Generalization because one value is hidden. If the adversary wants to understand the particular value for the postcode, he must try 0-9 possible values to urge its original value. because the generalization level increases, the overhead of the adversary also increases. On the other hand, suppression suggests suppressing the entire attribute value. So, for postcode attribute value 405028 are going to be [10]. In this case, the adversary is blank about the postcode value. But the main drawback of suppression is it decreases data utility and results in information loss because sometimes, the postcode is additionally used for special sort of survey [11]. The rationale for the unpopularity of suppression overgeneralization is it disobeys the principle of minimality, which emphasizes on applying the minimum level of generalization. Because the target of Generalization and Suppression techniques to guard data from the adversary and to not hide from others [12].

Drawbacks of generalization

When generalization is applied on data set, it loses high dimensional data. Because in high dimensional data, most of the info shows almost similar values. So, to realize k-anonymity, generalization level is going to be high [2]. The generalization technique reduces data utility. When generalized data is employed by data analyst for analysis, he must assume values for generalized data. So overhead of knowledge analysts increases [2]. Correlations between attributes are lost. When the generalization technique is applied. For example, some diseases are supported age, but when generalization is applied on the age that's quasi-identifier, age value is replaced or suppressed. Therefore, the important point age-disease relation is lost [2]. The main reason behind releasing private tables as the public is a few kinds of surveys. But once the generalization is applied, it leads information hiding of this significant fact. Therefore, the analysis of the released dataset is a smaller amount accurate. Our proposed model overcomes the disadvantage of generalization [13]. To take care of the correlation between attributes slicing approach is employed and the proposed paper is an advanced version of the slicing technique. K-anonymity technique is straightforward and straightforward for the cover of sensitive data. Suppose any table ABC which satisfies K-anonymity for K value=4 referred to as a 4-anonymous table. When an adversary who only knows quasi-identifier values of that table cannot identify the worth regarding individual confidently greater than ¼ [14].

L-diversity and bucketization

In the previous model associated with K-anonymity justified that K-anonymity fails to supply sufficient protection against attribute disclosure. to deal with this limitation of K-anonymity a replacement technique l-diversity is proposed which applied to anonymous data. The l-diversity principle state that "An equivalence class is claimed to possess l-diversity if there is a minimum of l-" well-represented " values for the sensitive attribute [15]. A table is claimed to possess l-diverse if it follows diversity measures for equivalence sets to realize the l-diversity Bucketization technique is employed.

In Bucketization sensitive values are permuting in between the buckets, where the table is split into buckets on the idea of quasi-identifier values. Therefore, the resultant anonymized data consists of a group of buckets with permuted sensitive attribute values [14]. Bucketization has better data utility and protection then generalization but even has some limitations.

Bucketization technique fails to guard against membership disclosure because it publishes quasi-identifier original form and

only applies permutation among sensitive attributes, the adversary can directly make sure his required information about the individual is present in the dataset or not [2]. Clear separation between sensitive attributes and quasi-identifiers aren't mentioned. Like generalization, the Bucketization technique also breaks correlations between attributes [15]. Table 3 contains the hospital dataset. Table 4 shows 3-diverse 4-anonymous table.

Table 3: Hospital data set.

S. No.	Sex	Age	Zip	Disease	Weight	Occupation	Qualification	Salary
1	M	21	432157	Cancer	78	Manager	MBA	10000
2	M	22	123456	Heart Attack	44	Clark	BA	3000
3	F	34	345678	HIV	57	Student	UG	5400
4	F	45	678920	Brain Tumor	68	Student	UG	56000
5	F	67	234567	Flu	66	BE	BCA	42000
6	F	63	234589	Headache	66	CA	BCA	48000
7	M	45	234567	Cancer	66	Developer	BCA	36000
8	F	23	678901	Gastric problem	68	CA	CA	56000
9	F	41	345678	Cancer	57	Manager	ME	54000
10	F	45	234567	Kidney Failure	46	Developer	BE	45000
11	M	46	123456	Headache	40	Clark	BA	30000
12	M	21	432156	Flu	78	Manager	MBA	1000000

Table 4: 3-diverse 4-anonymous table.

S. No.	Zip Code	Age	Nationality	Medical Status
1	1305*	≤ 40	*	Heart disease
2	1305*	≤ 40	*	Viral infection
3	1305*	≤ 40	*	Cancer
4	1305*	≤ 40	*	Cancer
5	1485*	≤ 40	*	Cancer
6	1485*	≤ 40	*	Heart disease
7	1485*	≤ 40	*	Viral infection
8	1485*	≤ 40	*	Viral infection
9	1306*	≤ 40	*	Heart disease
10	1306*	≤ 40	*	Viral infection
11	1306*	≤ 40	*	Cancer
12	1306*	≤ 40	*	Cancer

* is used for Generalization Technique. Generalization is used to protect sensitive data from information disclosure

Slicing technique

As generalization and Bucketization techniques break co-relations between attribute slicing techniques were proposed to beat these drawbacks. The slicing technique directly applied on the dataset without prior knowledge of sensitive and non-sensitive attributes. So, there's no got to mention this sort of separation by Data publisher.

In slicing dataset is split into both directions horizontally and vertically. In horizontal partition tuples groups into buckets, whereas in vertical partition columns are divided on the idea of correlations among attributes, where new attribute contains a subset of the highly correlated attribute. Vertical Partitioning reduces the dimensionality of the dataset, because multiple attributes store in a single attribute [2]. Table 5 shows results when maintaining the correlations of dataset and the results of it.

Table 5: Results of maintaining attribute correlations.

S. No.	Age, Sex, Disease	Zip, Disease	Qualification, Salary
1	21, M, Cancer	Cancer, 432157	MBA, 10000
2	22, M, Heart Attack	Heart Attack 123456	BA, 3000
3	34, F, HIV	HIV, 345678	UG, 5400
4	45, F, Brain Tumor	Tumor, 678920	UG,56000
5	67, F, Flu	Flu, 234567	BCA, 42000
6	63, F, Headache	Headache, 234589	BCA, 48000
7	45, M, Cancer	Cancer, 234567	BCA, 36000
8	23, F, Gastric problem	Gastric problem,678901	CA, 56000
9	41, F, Cancer	Cancer, 345678	ME, 54000
10	45, F, Kidney failure	Kidney failure, 234567	BE, 45000
11	46, M, Headache	Headache, 123456	BA, 30000
12	21, M, Flu	Flu, 432156	MBA, 1000000

Proposed model K-AMSOA overcomes limitations of slicing are as follows:

- Slicing technique-based model was proposed for a single overlapped attribute. So, if single attributes overlapped in multiple columns may show challenges for security
- Slicing fails to guard against membership disclosure in some cases
- Slicing may face challenges against strong background attack
- All attribute set values are true. Only values of set are interchanged
- High possibility of homogeneity attack
- System cannot prevent against background attack
- The values are randomly permuted within bucket. So, if adversaries have some knowledge for distribution then he can easily access the info
- If during a bucket only 4 records are present and three of them shows same disease then the resultant combination shows lack of diversity
- System randomly generates the associations between column values of a bucket, this might lose data utility
- Possibility of unsorted matching attack
- This system isn't feasible for dynamic datasets

There should be some proper mechanism present within the case of overlapped attributes and for dynamic Datasets. Overlapped attributes are those attributes that show correlations to quite one attributes or appended to quite one attribute and dynamic data set means data publisher can add or remove attributes at the time of applying privacy preservation technique and release results. Because each one attributes don't need to be compulsory for release. Data publishers can modify his release to guard against disclosure.

K-AMOA

K-Anonymity model for multiple overlapped attribute:
The main objective of K-AMOA is to style and develop A Model for privacy-preserving data publishing where correlations between attributes are maintained within the presence of overlapped attributes. Attribute correlation means if any attribute from dataset reflects some relations with another attribute set, i.e. age and disease, mostly below 18 age boys do not suffer from cancer. This information must be forwarded to the medical health department because it shows some important statics about cancer and the chances of cancer after 18 [16].

In given example age and disease show some important relations that shouldn't be generalized or suppressed. it's also possible that occupation and disease show some relation but age and occupation aren't related [17]. We can't store the attribute in one set i.e. (Age, Occupation, Disease). As disease attribute is overlapped in two groups (Age, Disease) and (Occupation, Disease). When an attribute is present in an attribute set quite once referred to as an overlapped attribute.

Proximity breach

Previously proposed Anonymization principles aren't effective at preventing the "proximity breach", which may be a privacy threat specific to numerical sensitive attributes (such as salary). Intuitively, a proximity breach occurs when an adversary has high confidence about his belief for sensitive information range even, he doesn't know the particular value but supported his belief he predicts the range of sensitive information.

PROPOSED OUTCOME

Proposed Model Privacy-preserving technique for multiple co-related sensitive attributes should produce following outcomes:

- choice for data publisher attribute type i.e. sensitive or non-sensitive
- If data publisher has less knowledge about system even then he acts by selecting options i.e. select attribute type, co-related attributes, overlapped attributes as per his requirement
- Value of K may vary as per requirement
- System work effectively for multiple published versions over dynamic data set
- Protection from proximity breach attack and membership disclosure

CONCLUSION

Although many models are proposed for privacy preservation models, but as data quantity and formats change the need of the latest model is increases. Every proposed model supported one aspect of security because it very hard to secure the whole dataset contains multiple attributes. When scene changes and data set are considered as correlated attributes and also contains sensitive attributes system need a replacement model. During this paper advantages and drawbacks of previous models are proposed and an overview of our proposed system and its background and wish is proposed.

REFERENCES

1. Maheshwarkar N, Maheshwarkar B, Patidar P, Rawat MK. K-AMOA: K-anonymity model for multiple overlapped attributes, conference ICTCS 2016.
2. Liu Q, Shen H, Sang Y. Privacy-preserving data publishing for multiple numerical sensitive attributes. Tsinghua Sci Technol. 2015;20:246-254.
3. Li T, Li N, Zhang J, Member, Ian M. Slicing: A new approach for privacy preserving data publishing. IEEE Transactions on KDE. 2012;24:3.
4. Maheshwarkar N, Pathak K, Chourey V. N-SA K-anonymity model: a model exclusive of tuple suppression technique. Third Global Congr Intell Syst. 2012;229-232.
5. Maheshwarkar N, Pathak K, Chourey V. Performance evaluation of various K-anonymity techniques, proc SPIE 8350. Fourth International Conference on Machine Vision. 2011.
6. Maheshwarkar N, Pathak K, Chourey V. Privacy issues for K-anonymity model. Int J Eng Res Apl. 2011;1:1857-1861.
7. Maheshwarkar N, Pathak K, Chourey V. Performance issues of various K-anonymity strategies. Int J Comp Tech Electr Eng (IJCTEE). 2011:2249-6343.
8. Aggarwal C. On k-anonymity and the curse of dimensionality, proc int'l conf. Very Large Data Bases (VLDB). 2005:901-909.
9. Blum A, Dwork C, McSherry F, Nissim K. Practical privacy: the SULQ framework. Proc ACM Symp. Principles Database Syst. 2005:128-138.
10. Brickell J, Shmatikov V. The cost of privacy: Destruction of data-mining utility in anonymized data publishing, Proc ACM SIGKDD Intl Conf Knowledge Discovery and Data Mining (KDD). 2008.
11. Chen BC, LeFevre K, Ramakrishnan R. Privacy skyline: privacy with multidimensional adversarial knowledge, Proc Intl Conf. Very Large Data Bases (VLDB). 2007:770-781.
12. Cramt'er H. Mathematical methods of statistics. Princeton University Press. 1948.
13. Dinur I, Nissim K. Revealing information while preserving privacy, Proc ACM Symp. Principles Database Syst. 2003:202-210.
14. Dwork C. Differential privacy, Proc Int'l Colloquium Automata, Languages and Programming (ICALP). 2006:1-12.
15. Dwork C. Differential privacy: A survey of results, Proc Fifth Intl Conf. Theor Appl Models of Comput. 2008:1-19.
16. Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. Theor Cryptography. 2006:265-284.
17. Friedman JH, Bentley JL, Finkel RA. An algorithm for finding best matches in logarithmic expected time. ACM Trans Math Software. 1977;3:209-226.