Journal of Advances and Scholarly Researches in Allied Education
Vol. XV, Issue No. 5, July-2018, ISSN 2230-7540

An Overview on Data Anonymization and Encryption in Data Mining

Amit Kumar[1]* Dr. Manoj Kumar[2]

[1] Research Scholar, Shri Venketashwara University, Gajraula, Uttar Pradesh

[2] Department of Computer Science, Shri Venketashwara University, Gajraula, Uttar Pradesh

Abstract – Anonymization is a term explained in oxford dictionary as 'unknown'. Anonymization makes a protest indifferent from other items. It tends to be done by removing personally identifying information (PII) like Name, Social Security number, Phone number, Email, Address and so forth. De-identification is the way toward removing or obscuring any personally identifiable information from individual records in a way that minimizes the risk of unintended disclosure of the character of individuals and information about them. Anonymization of data alludes to the procedure of data de-identification that produces data where individual records can't be linked back to an original as they don't include the required translation variables to do as such. General data anonymization is a huge research region spanning numerous decades. In any case, the most generally utilized procedures for anonymization of data content are at present k-anonymity, L-Diversity and T-Closeness for privacy-preserving microdata discharge. In this paper we discuss about data anonymization and encryption in data mining.

- X -

INTRODUCTION

The disclosure of personal information one of a kind personal identifiers like personal numbers, social security number or some other interesting numbers can without much of a stretch be erased from datasets before releasing them publicly. To distribute these data to be utilized by researchers or examiners. In spite of the fact that data anonymization and encryption are connected topics and are both valuable techniques for securing cloud-based data from privacy and security ruptures, they are not a similar thing. Data anonymization is the way toward transforming data with the goal that it very well may be handled in a valuable manner, while preventing that data from being linked to individual personalities of individuals, articles, or organizations. Encryption involves transforming data to render it indistinguishable to the individuals who don't have the key to unscramble it. Encryption can be a valuable instrument for doing anonymization especially when hiding identifying information in an arrangement of data. Notwithstanding, encryption while helpful is neither fundamental nor adequate for doing anonymization. Data can be effectively anonymized without encryption and scrambled data isn't really anonymized.

DATA ANONYMIZATION AND ENCRYPTION

Data encryption is an anonymization technique that replaces touchy data with scrambled data. The procedure gives powerful data confidentiality yet in addition changes data into a mixed up arrangement. For instance, once data encryption is connected to the fields containing usernames, "John Doe" may progress toward becoming "data". Data encryption is reasonable from an anonymization viewpoint, however it's often not as appropriate for commonsense utilize. Other business prerequisites, for example, data input validation or application testing may require a particular data type, for example, numbers, cost, dates or compensation and when the scrambled data is put to utilize, it might have all the earmarks of being the wrong data type to the framework trying to utilize it.

Researchers and experts require data which is consistent and sound, encrypting these data with cryptographic algorithms won't give them the data with their culmination/honesty. Other than the exceptional personal identifiers, datasets could hold properties which could be a danger in the wake of being linked with other publicly accessible data (alluded as semi identifiers). This data should be legitimately examined on how much information could be found by linking this data with other publicly accessible information. A superior method for hiding those one of a kind and joined qualities (semi identifiers) from identifying individuals is to utilize anonymization techniques.

DATA ANONYMIZATION CONCEPTS AND TECHNIQUES

There are various data anonymization techniques that can be utilized including data encryption, substitution, shuffling, number and date difference and nulling out particular fields or data sets.

Substitution consists of replacing the contents of a database segment with data from a predefined rundown of dissident however comparable data types. So it can't be followed to the original subject.

Shuffling is like substitution, with the exception of the anonymized data is gotten from the section itself. The two techniques have their advantages and disadvantages, depending on the measure of the database in utilize. For instance, in the substitution procedure, the integrity of the information remains intact. In any case, substitution can represent a test if the records consist of a million usernames that require substitution. A viable substitution requires a rundown that is equivalent to or longer than the measure of data that requires substitution. In the shuffling procedure, the integrity of the data additionally remains intact and is anything but difficult to obtain, since data is gotten from the existing section itself. Yet, shuffling can be an issue if the quantity of records is little.

Number and date fluctuation are helpful data anonymization techniques for numeric and date segments. The algorithm involves modifying each value in a section by some random level of its real value to altogether change the data to an untraceable point.

Nulling out consists of just removing touchy data by deleting it from the mutual data set. While this is a basic technique, it may not be appropriate if an evaluation should be performed on the data or the imaginary type of the data. For instance, it is hard to question client accounts if fundamental information, for example, client name, address and other contact subtle elements are invalid values.

Anonymizing the data will consist of recognizing which variables are potential identifiers and modifying the dimension of precision of these variables to reduce the risk of re-identification to an adequate dimension. The key test is to expand the security while minimizing the resulting data misfortune.

Anonymized data can be put on cloud without worrying about others and can be mapped to original data in secure and confided in zone. Following are the some anonymization techniques which will assist us with providing security to data over cloud :K-anonymity, L-Diversity, T-closeness, Anatomy, randomization and so on. Anonymization using other techniques, for example, hashing, hiding and permutation can likewise be utilized to secure cloud data.

ANONYMITY BASED PRIVACY PROCTECTION MODELS

Diverse anonymization practices and techniques exist with variable degrees of heartiness. It deliver the main points to be considered by data controllers in applying them by having respect, specifically, to the certification attainable by the given technique taking into record the current condition of technology to consider three risks which are fundamental to anonymization.

1. Singling out corresponds to the likelihood to detach a few or all records which recognize an individual in the dataset;

2. Linkability is the capacity to link, no less than, two records concerning similar data subject or a group of data subjects either in a similar database or in two unique databases.

3. Inference is the likelihood to conclude with huge probability the value of a trait from the values of an arrangement of other properties. The two most commonly utilized protection models are K-Anonymity and L-Diversity

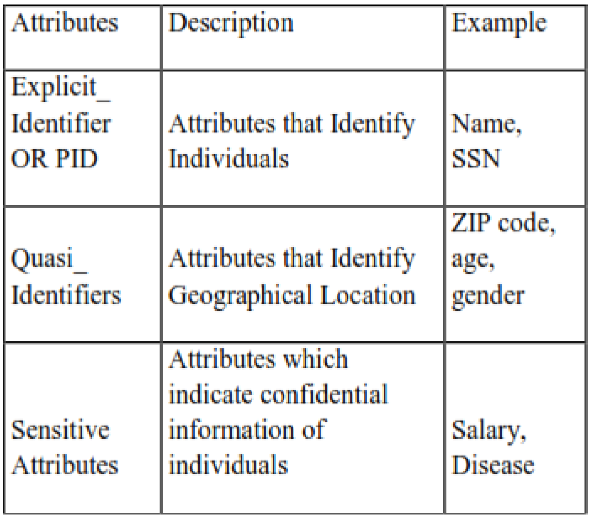
K-Anonymity

Anonymity word is gotten from a Greek word whose meaning is Nameless state or without name. As the name recommends, k-anonymity model converts properties in such a state where their personalities are avoided outside world Nidhi et al (2012), Maheshwarkar et al (2012)

K-Anonymity is useful for deterministic anonymization algorithm. It isn't adequate because of weaker than original k-anonimity which requires strong parametric assumption to be noteworthy Which thinks about privacy dimension of probabilistic smaller scale data discharge algorithm with deterministic one .No one can limit a person's record to not as much as k records.

The concept of K-anonymity was first introduced by Latanya Sweeney et al (2002). She had proposed this way to deal with take care of the problem of identifying individual from person particular organized data discharged .Data discharged said to be k-anonymous when each record can't be distinguished from in any event k-1 individuals whose information is distributed in that data. Commonly, such data is put away in a table, and each record (push) corresponds to one individual. Each record/dataset in smaller scale data has various qualities, which can be separated into the following three classes.

Table 1 K-Anonymity Attributes



Attributes	Description	Example
Explicit_ Identifier OR PID	Attributes that Identify Individuals	Name, SSN
Quasi_ Identifiers	Attributes that Identify Geographical Location	ZIP code, age, gender
Sensitive Attributes	Attributes which indicate confidential information of individuals	Salary, Disease

(1) Personal identification (PI) are properties that recognize individuals, for example, Name and SSN.

(2) Quasi-identifier (QI) are properties which include Demographic characteristics, for example, ZIP code, age, sexual orientation.

(3) Sensitive identifier (SI) Attributes which indicate confidential information of individuals, for example, Salary, **Dise**

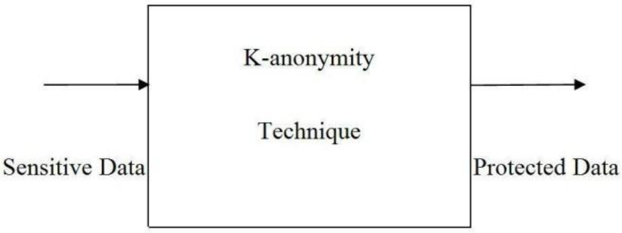


Figure 1: K-Anonymity Privacy Model

K-Anonymization methods

There are two methods followed in K-anonymization

(i) **Suppression:** Suppression consists in preventing touchy data by removing it. Suppression can be connected at the dimension of single cell, whole tuple, or whole segment, enables reducing the measure of generalization to be forced to accomplish k-Anonymity. The following are the kinds of suppression

(a) Tuple Suppression (TS) Suppression is performed at column level suppression operation evacuates entire tuple.

(b) Attribute Suppression (AS) is performed at section level suppression operation shrouds every one of the values of a segment.

(c) Cell Suppression (CS) is performed at single cell level

Finally k-anonymized table may wipe out only certain cells of a given tuple/trait. In suppression some property values are supplanted by *(asterisk)

(ii) **Generalization:** Generalization is the way toward converting a value into a less particular general term. For Example, "Male" and "Female" can be summed up to "Person".

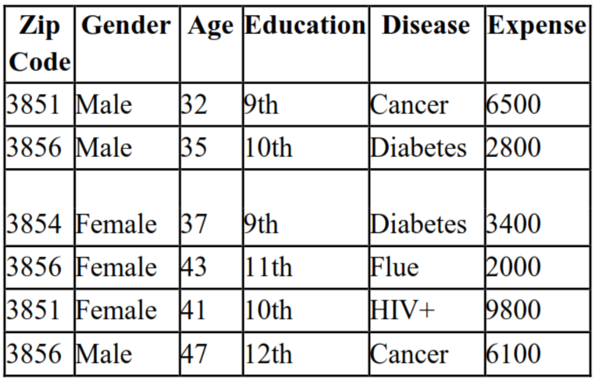
For instance: if an organization gathers data on individual travel developments, the individual travel designs at occasion level would in any case qualify as personal data for any gathering, as long as the data controller (or some other gathering) still approaches the original crude data, regardless of whether coordinate identifiers have been expelled from the set gave to outsiders. Be that as it may, if the data controller would erase the crude data and only give total measurements to outsiders on an abnormal state, for example, 'on Mondays on direction X there are 160% a greater number of travelers than on Tuesdays', that would qualify as anonymous data.

(a) Attribute Generalization (AG) is performed at the segment level. Every one of the values in the segment are summed up at a generalization step.

(b) Cell Generalization (CG) can be performed on a single cell, finally a summed up table may contain for a particular segment and values at various dimensions of generalization.

In generalization, individual values of traits are supplanted by more extensive classification (ex: Age=30 can be supplanted by age:<25-35>) as appeared in Table 3.3.

Table 2 Sample set of Records to be Anonymized

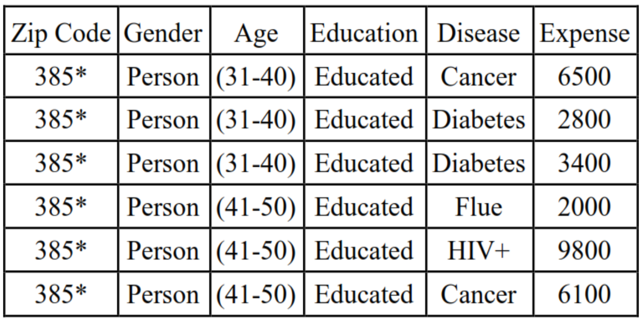


Zip Code	Gender	Age	Education	Disease	Expense
3851	Male	32	9th	Cancer	6500
3856	Male	35	10th	Diabetes	2800
3854	Female	37	9th	Diabetes	3400
3856	Female	43	11th	Flue	2000
3851	Female	41	10th	HIV+	9800
3856	Male	47	12th	Cancer	6100

The Table 2 demonstrates an example set of records taken and the properties in the records are postal division, sexual orientation, age, education, sickness and costs. Here the traits postal division, sex, age and education are only taken as semi identifiers. For Example: Table 2 is to be anonymized with Anonymization Level (AL) set to 3 and the arrangement of Quasi identifiers as QI = {ZIP CODE, GENDER, AGE, EDUCATION}.Sensitive characteristic = {DISEASE}. The semi identifiers and touchy qualities are identified by the organization according to their tenets and regulations.

The Table 3.3 demonstrates an example anonymized table based on the semi identifier traits identified by organization according to their tenets and regulation.

Table 3 Sample Anonymized Records (K=3)



Zip Code	Gender	Age	Education	Disease	Expense
385*	Person	(31-40)	Educated	Cancer	6500
385*	Person	(31-40)	Educated	Diabetes	2800
385*	Person	(31-40)	Educated	Diabetes	3400
385*	Person	(41-50)	Educated	Flue	2000
385*	Person	(41-50)	Educated	HIV+	9800
385*	Person	(41-50)	Educated	Cancer	6100

The date of birth of an individual can be summed up, in the type of month and year or only year. So this contains some original values and in addition increase confusion to enemy to infer touchy data. Suppression isn't performed dependably. For the most part the data distributers disregard to perform suppression since it causes data misfortune. For the situation of numerous delicate traits, suppression causes the distortion and information misfortune. Be that as it may, if suppression is disregarded then it can reduce the distortion.

Generalization technique is connected on Domain and value. At the point when generalization is connected on Domain set, it is called Domain Generalization Hierarchy and when it is connected on single values, it is considered as Value Generalization Hierarchy.

The K-anonymity model: pros and cons

K-Anonymity is one of the most acknowledged models for privacy in real-life applications gives the theoretical premise to privacy related legislation. Aside from this few essential reasons as pursues

i. The k-anonymity model defines the privacy of the yield of a procedure and not of the procedure itself. This is in sharp contrast to by far most of privacy models that were proposed before, and it is in this feeling of privacy that customers are generally interested.

ii. It is a basic, intuitive, and surely knew model. In this way, it advances to the non-master who is the end customer of the model.

iii. Although the way toward computing a k-anonymous table might be very hard, it is anything but difficult to approve that a result is indeed k-anonymous. Thus, non-master data proprietors are effectively guaranteed that they are using the model legitimately.

iv. The assumptions regarding separation of semi identifiers, method of attack, and inconstancy of private data have so far withstood the trial of real-life situations.

The limitations of the k-anonymity model is based on two assumptions. To begin with, it might be hard for the proprietor of a database to determine which of the attributes are not accessible in outer tables. This limitation can be overwhelmed by adopting a strict methodology that expect a great part of the data is public.

The second limitation is a lot harsher. The k-anonymity model expect a certain strategy for attack, while in real situations there is no reason why the attacker ought not attempt other strategies, for example, injecting false columns into the database. Obviously, it tends to be guaranteed that other acknowledged models present comparative limitations. For instance, the very much acknowledged model of semi-honest attackers in cryptography likewise limits the actions of the attacker.

A third limitation of the k-anonymity model distributed as of late in the literature is its certain assumption that tuples with comparative public characteristic values will have diverse private property values. Regardless of whether the attacker knows the arrangement of private trait values that coordinate an arrangement of k individuals, the assumption remains that he doesn't know which value coordinates any individual specifically. Since there is no express restriction forbidding it, the value of a private characteristic will be the equivalent for an identifiable group of k individuals. All things considered, the k-anonymity model would allow the attacker to find the value of an individual's private property.

PROBLEM OF DATA WITH MULTIPLE SENSITIVE ATTRIBUTES

The group of the initial four patients has three distinct values on the two attributes. In any case, if the attacker knows that a patient does not have coronary illness, he can choose that this patient has IV (intravenous therapy) as treatment in light of the fact that only patients with heart infections got the other two kinds of treatment. Along these lines this group isn't 3-various. The underlying driver of this problem is that the elimination of lines containing one sensitive property value may eliminate multiple values of other sensitive attributes. For this situation, the elimination of lines containing the value heart maladies additionally eliminates values medicine and medical procedure. Therefore, preserving L-diversity on every individual sensitive quality won't safeguard L-diversity for multiple sensitive attributes.

CONCLUSION

Before publishing the small scale data on cloud for security reason, the incoming data is first bunched using incremental clustering strategy as the data on cloud is enormous and circulated. Distinctive clustering algorithms are examined. To protect privacy on cloud, anonymization based privacy preservation techniques is received. Diverse anonymization techniques were examined.

REFERENCES

[1] R. Vernica, M. J. Carey, C. Li (2010). Efficient parallel set-similarity joins using Map Reduce, in: Proceedings of the 2010 ACMSIGMOD International Conference on Management of Data, IGMOD'10, 2010, pp. 495-506.

[2] L. Wang, J. Zhan, W. Shi, Y. Liang (2012). In cloud, can scientific communities benefit from the economies of scale?, IEEE Trans.Parallel Distrib.Syst.23(2) pp. 296-303.

[3] Guang Dong, Luping Jiang and Chunmei Liu (2012). "Based on Cloud Computing Inventory and Distribution Management Platforms", vol.7, no.6(Part A), pp. 2837-2842.

[4] Michael Armbrust, Armando Fox, and et. al. (2009). "Above the clouds: A Berkeley view of cloud computing," Technical Report UCB-EECS-2009-28, University of California, Berkeley.

[5] Siani, P., S. Yun and M. Miranda (2009). "A privacy manager for cloud computing," Cloud Computing, vol.5931, pp. 90-106.

[6] S. Chaudhuri (2012). What next? : A half-dozen data management research goals for big data and the cloud, in: Proceedings of the 31st Symposium on Principles of Database Systems, PODS'12, 2012, pp.1-4.

[7] H. Takabi, J. B. D. Joshi, G. Ahn (2010). Security and privacy challenges in cloud computing environments, IEEE Secur.Priv. 8(6) pp. 24-31.

[8] Dean, J., Ghemawat, S. (2010). Map Reduce: A flexible data processing tool. Commun. ACM 53(1) pp. 72–77.

[9] X. Xiao, Y. Tao (2006). Anatomy: Simple and effective privacy preservation, in: Proceedings of 32nd International Conference on Very Large Data Bases, VLDB'06, pp. 139-150.

[10] T. Li, N. Li, J. Zhang, I. Molloy (2012). Slicing: A new approach for privacy preserving data publishing, IEEE Trans. Knowl. Data Eng. 24(3) pp. 561-574.

[11] M. Terrovitis, J. Liagouris, N. Mamoulis, S. Skiadopolous (2012). Privacy preservation by disassociation, Proc.VLDBEndow. 5(10) pp. 944-955.

Corresponding Author

Amit Kumar*

Research Scholar, Shri Venketashwara University, Gajraula, Uttar Pradesh