International Journal of Technical Innovation in Modern Engineering & Science (IJTIMES)
Impact Factor: 3.45 (SJIF-2015), e-ISSN: 2455-2585
Volume 4, Issue 5, May-2018

MedPub: A new Clustering-based approach for grading methods for Publishing Medical Data

Shalini Bhaskar Bajaj
Department of Computer Science and Engineering
Amity University Haryana
shalinivimal@gmail.com

***Abstract*— Most medical data publishing studies does not partition the semantics of the values of the sensitive attribute. This leaves the data prone to attacks such as homogeneity. This problem has been discussed and resolved in this paper by presenting a model that uses clustering based approach to partition the attributes values that are disease sensitive by grading methods for publishing medical data. The effectiveness of the proposed graded medical publishing method has been experimentally shown. Results shows that the new proposed model is nearly same as L-diversity in released information quality**

***Keywords*— k-means algorithm; homogeneity; data publishing; semantics**

I. INTRODUCTION

Identifiers are deleted from the tables for storing data before publishing any sensitive data. Identifiers are usually used to differentiate one data point from another. Most commonly used identifier is Social Security Number to distinguish data points. Even though the identifiers are removed from the data table before making it public on internet still the data thieves can use link attack [1] in order to fetch private information of different users from the data tables made public. Thus, to protect the data tables they need to be further processed. This can be achieved by using K-anonymity[1] which divides the table storing data into groups in such a way that every group is having at least K records and thus generalizes similar records in one group. K-anonymity does not take into account the diversity requirement of sensitive attribute values and thus it is unsafe. L-Diversity was presented in order to resolve this problem. L-Diversity states that each group should have minimum of L different sensitive values. Let Table I represents initial (original) Reference table and Table II represents L-Diversity for L value equal to two. From Table II we can observe that each group has at least two different attribute sensitive values. Here, age of a person and its region cannot identify individuals uniquely but still has some information about the individuals. Thus they can be called quasi identifies. Usually attribute Disease associated with an individual gives sensitive information thus it comes under the category of sensitive attribute and thus it should be protected from getting published. Though L-Diversity ensures privacy of the sensitive data but still suffers from the problem of disclosure of privacy. It can be explained with the help of an example. Ram has a neighbour Shyam. Shyam's data is stored in the data table (refer Table II). Shyam lives in Gurgaon and is 25 years of age. Ram can find out the disease Shyam is suffering from refering Table II since age and Region are quasi identifiers. Ram can find out that Shyam is in group 1. Although Ram cannot find out details of Shyam but still he can relate that Shyam belongs to group 1 and is sure that Shyam is suffering from HIV or cancer. Thus, Shyam's privacy is disclosed on Ram. The above situation is a perfect example of attack refereed here in this paper i.e. homogeneity.

Since L-Diversity has not taken into account the semantics thus it is prone to homogeneity attack. A number of data publishing models has been proposed in literature that talks about K-anonymity and L-Diversity [3, 4, 5]. The paper presented in [3] discusses partitioning of the sensitive attribute in order to resist homogeneity attack but since the partitioning method proposed in [3] is based on common sense therefore it is difficult to ensure that the attribute picked as sensitive is correct one or not.In this paper disease attribute has been picked as the sensitive attribute who's values are divided into four different risk levels discussed in the next section. Clustering algorithm is applied on the disease attribute value to divide the data points into different groups.

Table I
Initial (Original) Reference Table

ID	Age	Region	Diseases
1	35	Panipat	HIV
2	44	Delhi	Cold
3	45	Chandigarh	Fever
4	31	Gurgaon	Cancer

Table II
L-Diversity for L = 2

Group ID	ID	Age	Region	Diseases
1	1	31-35	(Panipat, Gurgaon)	HIV
1	4	31-35	(Panipat, Gurgaon)	Cancer
2	3	44-45	(Delhi, Chandigarh)	Fever
2	2	44-45	(Delhi, Chandigarh)	Cold

II. Model on (C, L)-Diversity

A. Disease sensitivity degree details

Sensitivity degree of disease can be measured from two different aspects:

i). from the point of view of health;

ii). from the moral issues.

Disease are divided into four different levels from the health point of view (refer Table III):

L1: symptoms are mild wherein hospitalization is not needed;

L2: hospitalized is must for recovery;

L3: serious illnesses which may need surgery for recovery;

L4: diseases that cannot be cured and finally leads to death of the patient comes under the category of fatal diseases

Table III

Sensitivity Degree Associated with Diseases

Level X	order	Associated weight values	Level Y	order	Associate weight values
1	1	0	1	1	0
2	2	1/3	2	2	1
3	3	2/3			
4	4	1			

The Level X has four different levels and for each given level j ($1 \leq j \leq 4$) assign an order riX[6] such that {r1X=1,r2X=2,r3X=3,r4X=4} and assign a weight ziX to each order riX, we can get {z1X,z2X,z3X,z4X}, each ziX∈ {z1X,z2X,z3X,z4X} satisfies [6]:

ZiX = (riX - 1)/(r4X - 1) ,

where ZiX belongs to [0,1]

Similarly, we can divide diseases into two levels from the point of view of moral issues:

L1: diseases not having moral issues associated with it.

L2: diseases having moral issues associated with it such as HIV

In the similar way, weights can be assigned with Level Y. For diseases m and n, m.weight(X) and n.weight (X) represents the weight of m and n in X respectively, if m.weight (X)> n.weight (X), we can say that sensitivity of disease m is more than disease n as it is assigned more weight value. Weight values as assigned to different diseases by taking advice from experts in medical field. E.g. consider that diseases in Table I have some weight values associated with them (refer Table IV), K-means [6] can be applied to cluster four different diseases into two classes (C = 2). Thus each group of diseases should have two different L values and at the same time should have sensitive attribute values from at least C different classes (for details please refer Table V). The example discussed above is a erfect example for (C, L)-Diversity. In the following text definition of (C, L)-Diversity and its implementation are discussed in detail.

Definition 1 If in a given data table T, each group contains at least L different sensitive attributes then T satisfies L-Diversity.

Definition 2 If in a given data table, each group not only contains L different sensitive values but at the same time contains values from different C classes such that C <= L, then we can say that it satisfies (C, L)-Diversity.

(C,L)-Diversity shows that not only the semantic diversity of the data tables is maintained but at the same time highly sensitive information of the data table. Proposed algorithm is given below:

Input: released table RT

Output: table RTO that satisfies (C,L)-Diversity, SA represents sensitive attribute; x represents value of sensitive attribute, x.number represents the number of record y; y satisfy represents value of y in SA is x .i.e. y.SA = x.

Values of SA is divided into classes as {class 1, ..., class n}

1. When RT satisfies the condition of grouping repeat step 2 to 8
2. Create a new group G1;
3. For j= 1 to n, j++ , repeat 4;
4. Choose x ϵ classj; x satisfies: the x.number is largest in classj. G1∪ y, y stisfes: y.SA = x.RT-y, x.number--; for each x ~ϵ classj && x ~≠ x, Q ∪ v ~;
5. Choose {x1, x2, ... xL-n} from Q, x1, x2, ... xL-n have largest number in Q;
6. For each x ~ϵ{x1, x2, ... xL-n}
7. Choose y~ : y~ satisfies; y~ SA = x~. RT-y~, G1∪ y~, x~num--;
8. RT~∪G1, delete all values in Q;
9. For each yϵRT, find G1ϵT ~, G1 satisfies after adding y to G1, G1 satisfies (C,L)-Diversity, G1∪ y, RT-y

Table IV

Coefficient of Weights associated with different diseases for Level X and Level Y

Disease	Weight assigned to X	Weight assigned to Y
HIV	1	1
Cold	0	0
Fever	0	0
Cancer	1	0

Table V

(C,L)-Diversity for C=2 and L=2

ID	Age	Region	Disease
1	35	Panipat	(HIV,Cancer)
4	31	Gurgaon	(HIV,Cancer)
3	45	Chandigarh	(Fever,Cold)
2	44	Delhi	(Fever,Cold)

1. III. EXPERIMENTS

Adult [7] dataset is used for conducting experiments. In Adult dataset relationship attribute is used as a sensitive attribute since the dataset does not have Disease attribute. Weights are assigned to relationship attribute as follows; husband is given weight as 1,1; own child as 1/3, 0; unmarried as 0,1; wife as 1,0; not in family as 0,0; other relatives as 1/3,1. Randomly records have been chosen for each group for testing. From the experimental analysis it is shown that L-Diversity and (C,L)-Diversity [8] is improved in security.

2. IV. CONCLUSIONS

A graded method has been presented in this paper for medical data publishing that can counter homogeneity attack. Experiments conducted on Adult dataset shows that loss of information in case of (C, L)-Diversity is same as L-Diversity.

REFERENCES

1. [1] Sweeney, "K-anonymity:A model for protecting privacy", International Journal on Uncertainty,Fuzziness and Knowledge Based Systems,2002,10(5):557-570
2. [2] A. Machanavajjhala, J. Gehrke,D. Kifer, and M.Venkitasubramaniam, "L-diversity:privacy beyond k-anonymity," in ICDE, 2006, pp. 24-26.
3. [3] Wang Qian, Zeng Zi-ping, "(p,a)-sensitive k-anonymity: privacy protection mode", Application Research of Computers, 2009,26(6):2177-2183.
4. [4] Long Qi, "A k-anonymity Study of the Student—Score Publishing", Journal of Yunnan University of Nationalities (Natural Science Edition). 2011, 20 (2):144-148
5. [5] LIU Ming YE Xiao-jun, "Personalized K-anonymity", Computer Engineering and Design,2008, 29 (2):282-286
6. [6] Han JW, Kamber M., "Data Mining Concepts and Techniques", 2011.
7. [7] Hettich S, Blake C L, Merz C J. UCI repository of machine learning databases [EB/OL]. (1998). http://www.ics uci edu/mlearn/ MLRepository.html.
8. [8] Yang Xiao-Chun, Wang Ya-Zhe, WANG Bin ,YU Ge, "Privacy Preserving Approaches for Multiple Sensitive Attributes in Data Publishing", Chinese Journal of computers (China), 2008, 31(4),574-586