International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6S3, September 2019

Privacy Preservation using (L, D) Inference Model Based on Dependency Identification Information Gain

R. Deepika, V. Divya, C. Yamini, P. Sobiyaa

Abstract: ***The improvement of an information processing and Memory capacity, the vast amount of data is collected for various data analyses purposes. Data mining techniques are used to get knowledgeable information. The process of extraction of data by using data mining techniques the data get discovered publically and this leads to breaches of specific privacy data. Privacy-preserving data mining is used to provide to protection of sensitive information from unwanted or unsanctioned disclosure. In this paper, we analysis the problem of discovering similarity checks for functional dependencies from a given dataset such that application of algorithm (l, d) inference with generalization can anonymised the micro data without loss in utility. [8] This work has presented Functional dependency based perturbation approach which hides sensitive information from the user, by applying (l, d) inference model on the dependency attributes based on Information Gain. This approach works on both categorical and numerical attributes. The perturbed data set does not affects the original dataset it maintains the same or very comparable patterns as the original data set. Hence the utility of the application is always high, when compared to other data mining techniques. The accuracy of the original and perturbed datasets is compared and analysed using tools, data mining classification algorithm.***

Keywords: Anonymization, Classification, Functional Dependency Attributes, Gain Ratio Index,(l,d) Inference model, Privacy Preservation Data mining, ,Perturbation.

I. INTRODUCTION

Privacy preserving publishes of micro data have been considered briefly in previous years. Government agency and other organization regularly require distributing data, For Example: Healthcare data, survey data, for investigate and other determination. Normally, these report are save in the form of table, and each row in a table is called tuple or row corresponds to one individual person. Micro data that contain records of each individual entity which contains information a person, a household, or a society. Numerous micro data anonymization methods have been proposed during earlier work. K-anonymity for generalization and L - diversity method for bucketization. In both approach, these attributes are divided into three types: [21]

1) The attributes are called identifiers that can absolutely identify a person or thing, such as name or Security Number like aadhar card number;

2) The Identifiers Birth date, gender, and Zip code; these may previously be familiar and can potentially identify an individual is called Quasi Identifiers

3) Some attributes are Sensitive Attributes (SAs), which are undisclosed to the opposition and are considered perceptive, such as Disease and Salary.

II. DISCOVERING DEPENDENCY USING INFORMATION GAIN

Information gain (IG) ratio was proposed by Ross Quinlan.It mainly used to construct a decision tree and it is based on the idea of entropy from information theory.

2.1 SHANNON ENTROPY

Claude Shannon introduces the Shannon entropy in 1948. It quantifies "uncertainty". Using equation (2.1) on the values of attribute then Lower value indicates less insecurity and higher value indicates more insecurity [15]

Entropy (D) = - $(\sum_{i=1}^{m} p_i * log_2 p_i)$ ---------- > (2.1.1)

2.2 INFORMATION GAIN

The Information gain(IG) is calculated using the entropy value, the eq (2) is used.[15]

Information Gain (D) =Entropy (D) -$\sum_{i=1}^{n} \frac{|D_i|}{|D|} * Entropy(D_i)$ -----> (2.2.2)

Assumption 1: To find information gain Using eq(2.2) on the attribute in the dataset. The highest information gain based on class attribute is consider as FD1 (L) (Functional Dependency LEFT) and next highest information based on FD1 (L) is FD1 (R) (Functional Dependency RIGHT) and so on.

FD1 (L) -> FD1 (R), FD2 (L) -> FD2 (R)……FDn (L) ->FDn (R).In mandate to secure the data privacy by applying (l, d) inference model on functional dependency dataset

2.3 Finding Information Gain

Manuscript published on 30 September 2019.
***** Correspondence Author (s)
R. Deepika, Assistant Professor, Bannari Amman Institute of Technology, Tamilnadu,India.(Email: deepikar@bitsathy.ac.in)
V.Divya, Assistant Professor, Bannari Amman Institute of Technology, Tamilnadu,India.(Email: divyav@bitsathy.ac.in)
C.Yamini, Assistant Professor, Bannari Amman Institute of Technology, Tamilnadu,India.(Email: yamini@bitsathy.ac.in)
P. Sobiyaa, Assistant Professor, Bannari Amman Institute of Technology, Tamilnadu,India.(Email: sobiyaa@bitsathy.ac.in).

S.NO	Built	Height	Weight	Diagnosis	Breathing and Non_Breathing
1	Moderate Low	142	41	Disease-A	Non_Breathing_Ailments
2	Moderate Low	152	59	Disease-N	Non_Breathing_Ailments
3	Moderate Low	152	50	Disease-A	Breathing_Ailments
4	Moderate Medium	152	60	Disease-A	Breathing_Ailments
5	Moderate High	142	42	Disease-A	Breathing_Ailments
6	Moderate Medium	165	61	Disease-A	Non_Breathing_Ailments
7	Moderate High	148	47	Disease-A	Non_Breathing_Ailments
8	Moderate High	132	24	Disease-A	Non_Breathing_Ailments
9	Moderate Medium	93	11	Disease-A	Breathing_Ailments
10	Moderate High	140	27	Disease-U	Breathing_Ailments
11	Moderate Low	162	82	Disease-O	Non_Breathing_Ailments
12	Moderate Low	145	45	Disease-O	Non_Breathing_Ailments
13	Moderate Medium	176	76	Disease-O	Non_Breathing_Ailments
14	Moderate Low	160	61	Disease-C	Breathing_Ailments
15	Moderate Medium	158	61	Disease-G	Non_Breathing_Ailments

Table 2.1 Input Tables

Classification Entropy for Breathing and Non-Breathing attribute:

6/15 Breathing, 9/15 Non breathing

Using equation (2.2.2): IE= - (6/15) log2 (6/15)-(9/15) log2 (9/15) = ~0.971

Calculation:

Build: 6 moderate-low, 5 moderate-medium, 4 moderate-high. There are three values for attribute build, so we need three entropy calculations

Moderate-low: 5 no, 1 yes	$I_{low} = -(5/6)\log_2(5/6)-(1/6)\log_2(1/6) = \sim 0.65$
Moderate-medium: 3 no, 2 yes	$I_{medium} = -(3/5)\log_2(3/5)-(2/5)\log_2(2/5) = \sim 0.97$
Moderate-high: 2 no, 2 yes	$I_{high} = 1$ (evenly distributed subset)

Table 2.3.2 Entropy calculation

IGBuild = IE(S) – [(6/15)*Ilow + (5/15)*Imedium + (4/15)*IHigh]

IGBuild = 0.971 – 0.85 = 0.121

Using above example calculates the information Gain for all attribute based on class attribute.

SNO	ATTRIBUTE	INFORMATION GAIN VALUE
1	Complaints	0.34889
2	Diagnosis	0.18527
3	Taluk	0.03931
4	Appetite	0.02855
5	Mental_Generals	0.02833
6	Occupation	0.02232
7	Sex	0.00411
8	Height	0
9	Weight	0
10	Age	0

Table 2.3.3 Information gain value

III. PRIVACY PRESERVATION TECHNIQUES:(L, D) INFERENCE MODEL

It is a combat of L-diversity and T-closeness.

3.1 L-Diversity

L-DIVERSITY means privacy attributes must be "dissimilar" within each quasi-identifier equivalence class. It can be implemented using entropy based diversity.[10] It can be implemented using diverse privacy technique like L-Diversity, Probabilistic based l-diversity, Entropy based l-diversity and Recursive based (c,l)-diversity. In our paper we implement using L-diversity using Entropy class E is defined to be: Refer equation (2.1.1)

3.1.1 Algorithm For L-Diversity

INPUT: D is dataset, (A1, A2…An) attribute in dataset, (V1, V2…Vn) values in dataset.

OUTPUT: L-diversity on information gain.
Find dependency attribute based on information gain.
Formalizing the functional dependency attribute as A->B.
Ai(V1, V2…Vn)
Find information gain of (V1, V2…Vn) on class attribute
If IG (V1) is equal to or close to IG (V2)
Then Replace V1 with V2

3.2 T-Closeness

The requirement of T-CLOSENESS is distribution of a sensitive attribute in overall table is similar to the distribution of a sensitive attribute in tany equivalent class. The information gain of an observer measured the privacy.[16] Information Gain = Posterior Belief – Prior Belief where P = the portion of the sensitive attribute in equivalent class and Q = the allotment of the sensitive attribute in the complete table. The threshold t level should be greater then only the class is said to have t-closeness if the space between the allotments of a sensitive attribute in this class and the allocation of the attribute in the entire table. T-closeness is achieved by using EMD (Earth Moving Distance) formula. [1]

Calculation of EARTH MOVER DISTANCE: Calculating the EMD between two distributions lead to use privacy preservation techniques with EMD.

Attribute domain will be {U1, U2...Un}, where Ui is the ith smallest value. Ordered Distance: The distance between two values of is based on the number of values between them in the total order of the following formula:

Ordered dist(Ui, Uj) = |i−j| / n−1 .

EMD for Categorical Attributes: A full arrange often does not exist for categorical attributes. Equal Distance: The ground space between any two values of a categorical attribute is set to be 1. It is effortless to verify that this is a metric. As the space between any two values is 1, if the calculated value is Ai − Bi > 0, then it is move to some other points. Thus the formula used for: [8]

T [P, Q]= ∑_(i=1)^m|Ai-Bi|-----□(4.3)

3.2.1 Algorithm For T-Closeness

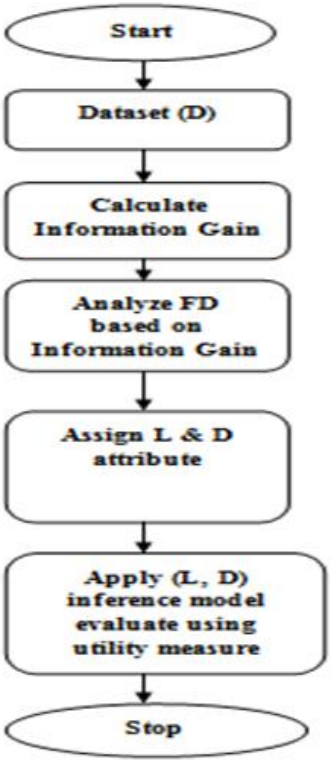
INPUT: D is dataset, (A1, A2…An) attribute in dataset, (V1, V2…Vn) values in dataset.

OUTPUT: T-closeness on Earth Moving Distance.

Calculate_EMD(Ai(Vi)) on FDLeft attribute //FD-Functional Dependency

If EMD(Ai(V1)) is equal to or close to EMD(Ai(V2))

Then Replace V1 with V2



3.2.1.1Flow chart Representation

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

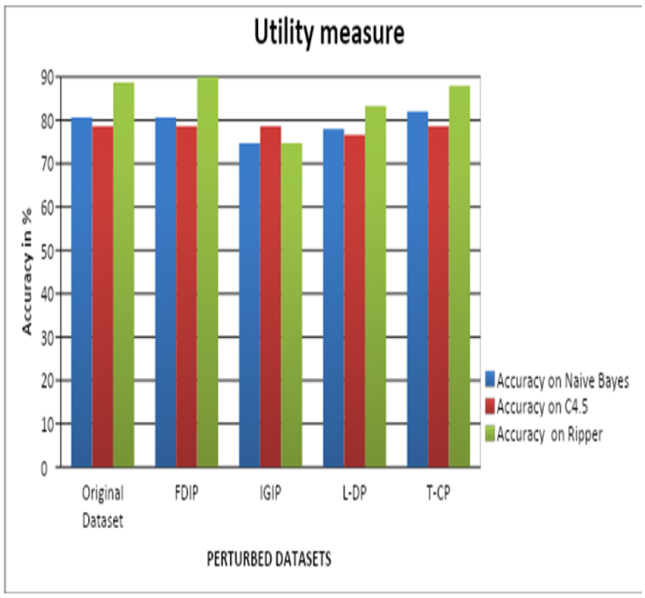
Hence in this paper we have perturbed data using (l,d) inference model. We are finding information gain for based functional dependency and applying (l,d) inference model obtain the perturbed dataset. The accuracy of the perturbed dataset is depicted in the graph as follows: where Functional Dependency based Inference Perturbation (FDIP), Information Gain Based Inference Perturbation (IGIP), L-Diversity Perturbed Dataset (L-DP), T-Closeness Perturbed Dataset (T-CP)

ID	COMPLAINTS	DIAGNOSIS
C60001	Pain_Related_Complaints	Disease-A
C60002	Allegric_Complaints	Disease-N
C60003	Itching_Complaints	Disease-A
C60004	Fever_And_Cold_Relaed_Complains	Disease-A
C60005	Skin_Related_Complaints	Disease-A

Table 4.1 Table before perturbation

ID	COMPLAINTS	DIAGNOSIS
C6 001	Others	Disease-A
C6 002	Ent_Complaints	Disease-A
C6 003	Ent_Complaints	Disease-A
C6 004	Irritation_In_Throat_Sneezingrunning _Nose	Disease-A
C6 005	Digestive_Complains	Disease-A

Table 4.2 Perturbed data



Figures 4.3 Information Gain based Perturbation

FDIP- FUNCTRIONAL DEPENDENCY BASED PERTURBATION

IGIP- INFORMATION GAIN BASED PERTURBATION

L-DP- L DIVERSITY BASED PERTURBATION

T-CP T-CLOSENESS PERTURBATION

V. CONCLUSION

The Figures 4.3 Information Gain based Perturbation shows the accuracy of all the datasets based on three data mining classifiers like Naïve bayes, C4.5, Ripper. The functional dependency of the attribute is considered based by decrease in the entropy or information gain for the attribute. The IG value calculated dataset is perturbed by privacy preservation techniques like L-diversity, T-Closeness. The scope of the work is to maintain the data mining accuracy even after applying privacy preservation techniques at the same time the sensitive data of individual should not breech. We attain the data preservation with good level of privacy without affecting the data mining performance metrics

VI. FUTURE WORK

The future work will be considering problem like non optimal solution result in over fitting while taking the attribute with large number of values. In order to solve the issue in future we apply depicting dependency of the attribute by using some other techniques like Gain ratio Index in order to reduce bias so level of privacy is improved than the current one in the case of attribute having large number of values.

REFERENCES

1. Charu C. Aggarwal, Philip S. Yu "Privacy-Preserving Data Mining:ModelsAnd Algorithms" University of Illinois at Chicago,Springer, 2008.
2. Agrawal R., Srikant R. Privacy-Preserving Data Mining. ACM SIGMOD Conference, 2000.
3. Aggarwal C. C., Yu P. S.: A Condensation approach to privacy preserving data mining. EDBT Conference, 2004.
4. Aggarwal C. C., Yu P. S. On Variable Constraints in Privacy Preserving Data Mining. ACM SIAM Data Mining Conference, 2005.
5. Agrawal D. Aggarwal C. C.On the Design and Quantification of Privacy Preserving Data Mining Algorithms. ACM PODS Conference,2002.
6. Aggarwal C. C. On k-anonymity and the curse of dimensionality. VLDB Conference, 2004.
7. Xinjing Ge and Jianming Zhu," Privacy Preserving Data Mining" School of Information, Central University of Finance and Economics Beijing, China
8. Shaoxu Song et al "Efficient discovery of similarity constraints for matching dependencies", Data and Knowledge Engineering ,2013, pp- 146-166
9. Hui Wang and RuilinLui,"Privacy Preserving Publishing Microdata with Full Functional Dependencies",Data and Knowledge Engineering 2011, pp 249-268
10. AshwinMachanavajjhala, ohannesGehrke, Daniel Kifer,MuthuramakrishnanVenkitasubramaniam , "L-Diversity: Privacy Beyond K-Anonymity"
11. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, "k-Anonymity" UniversitµadegliStudi di Milano, 26013 Crema
12. Meyerson A., Williams R. "On the complexity of optimal k-anonymity", 2004
13. Osmar R. Za¨ane et al., "Privacy-Preserving Data Mining on the Web Foundations and Techniques" Department of Computing Science University of Alberta Edmonton, AB, Canada.
14. Dip. di MatematicaPuraedApplicata F. Aiolli "Entropy and Information Gain" istemiInformativi 2007/2008
15. Michael Crawford, USRA; "ID3 Algorithm"
16. Murat Kantarcioglu "Other Privacy Definitions: l-diversity and t-closeness" Erik Jonsson School of Engineering & Computer Science, UT DALLAS
17. Krishna.V#, Santhana Lakshmi. S "Secured Medical Data Publication & Measure the Privacy "Closeness" Using Earth Mover Distance (EMD)" Coimbatore Institute of Engineering and Technology Coimbatore, India
18. Pranav Khaitan, KorraSathyaBabu, et al "Approximation algorithms for optimizing privacy and utility" International Conference on Computer Science and its Applications CSA 2009
19. Eibe Frank and Ian H. Witten "Data Mining- Gain ratio Computing the gain ratio" March 2004.
20. Gregory Piatetsky; "Classification: Decision Trees".