International Journal of Modern Trends in Engineering and Research
www.ijmter.com
e-ISSN No.:2349-9745, Date: 2-4 July, 2015

Privacy Preservation Data Mining Using GSlicing Approach

Mr. Ghanshyam P. Dhomse[1], Prof. Sonal Patil[2]

[1]*Department of C.S.E, GHRIEM Jalgoan, dhomase.ghanshyam@gmail.com*
[2]*H.O.D.Department of I.T., GHRIEM Jalgoan, sonalpatil3@gmail.com*

Abstract— Data mining is one of the most interested research area for many people. After mining the data it is necessary to publishing the data in proper way so that to keep Privacy and data utility of micro data, because it contains the sensitive information about individual. There are various anonymization techniques are used to fulfill requirement of balancing between privacy and data utility such as Generalization, Bucketization and Slicing approach but when compare this technique it is found some drawback like in generalization it is not suitable for large size of database due there is loss of information if size of data is increased. in case bucketization there is lots of tuple having same value in same bucket due to this it cannot prevent the membership disclosure. Another problem is difficult to identify the quasi-identifying attributes and sensitive attributes because of no separation between them. Slicing is better than that of both provides the membership and attributes disclosure protection but it contains some issue like it takes more time to search the fake tuple in original data set, it displays the record as it is while publishing the data that reduces the privacy. In case of slicing, for attribute clustering correlation is applied on discrete attributes, hence equal width Discretization is applied, but this method is very time consuming because it requires lot of sorting of data. So for solving this issue a new GSlicing Approach has been proposed .it does not require sorting of data for Discretization and it is better than previous slicing technique.

Keywords- Data Mining, Privacy Preservation, Generalization, Bucketization, Slicing

I. INTRODUCTION

In few recent years, privacy-preserving data mining has been popular area for the most of the Researcher so they studied extensively, because of the huge amount of information available on the internet and due to increase in variety of data collections containing person-specific information and sensitive type of information as growth in computer and Information technology, inter network connectivity become increasingly affordable. Data collection is the process of collecting information from data receiptant. Data collection is the process of publishing information for use to any user. Today most of the organizations require publish microdata. Microdata is records in which each of entry contains information about an individual entity, eg. Household.

It is the necessity of study important aspects in the area of data mining how to provide protection for sensitive attribute and reduce the disclosure of information on the basis of correlation between them. Bucket mainly contains permuted sensitive values but there is no separation between the Quasi and Sensitive attribute. Slicing better approach compare to both but it takes more time for execution and display the original record as it is while release the information so it is again causes the difficulty while provide the protection.

Therefore in this paper propose new approach for privacy preservation for adult dataset which partition the attribute in column on the basis of compute the correlation between each pair of attribute and form them in cluster. The following table shows the micro table which is used as our

dataset. Main motivation behind about work is First compute the correlation between the pairs of attribute and then form the cluster of attribute based on their correlation. To measure the correlation between two pair here use mean square contingency coefficient because most of our attribute categorical. For the Continuous Attribute first apply Discretization to partition the domain of continues value into interval because real-valued data ie continues data is often used as a pre-processing step in many data mining algorithms. in this paper Unsupervised Discretization methods based on clustering. Discretization method based on the k-means clustering algorithm which avoids the sorting time.

II. EXISTING APPROACH

Figure 1 shows the existing system architecture for privacy preservation data mining. Which include the different levels such as first of all select the dataset then perform the various anonymization operation on the dataset such as generalization, bucketization, Multiset based generalization, one attribute per column generalization then Slicing approach finally the sliced data is displayed. But till some issues related selection of categorical and continues attribute selection in slicing approach so to solve the problem we need some novel method for same.

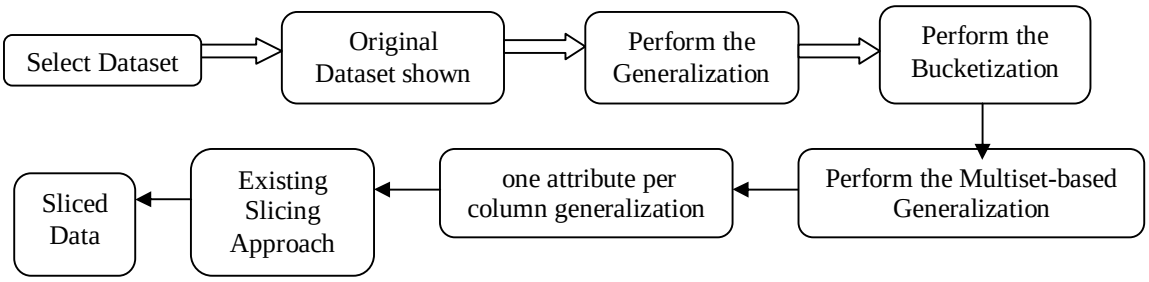


Figure 1. Existing System Architecture for Privacy Preservation

III. PROPOSED GSLICING APPROACH

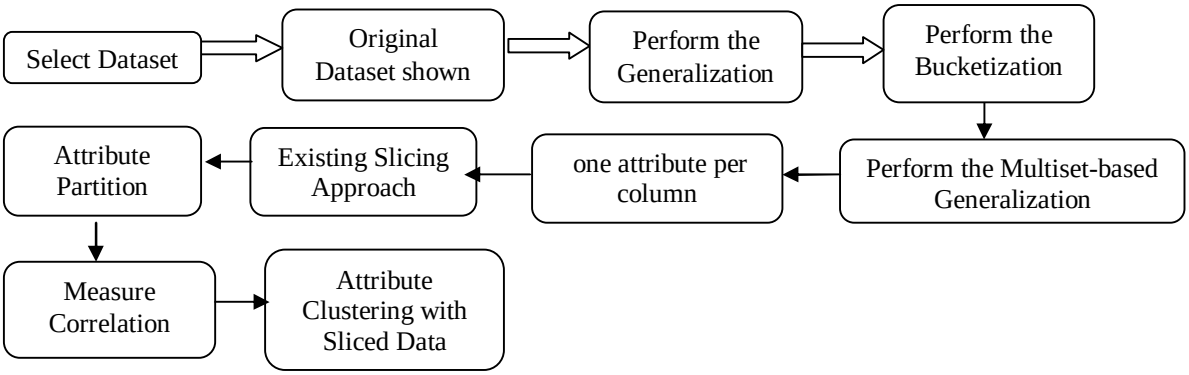


Figure 2. Proposed frameworks for GSlicing Approach

Figure 2 all the phase upto the Slicing Approach already deploy in existing model the main changes done at the time of Slicing Process in proposed method or Approach. In simple term our algorithm partitions attributes on the basis of correlation between them while in slicing approach highly correlated attributes are available in one Colum. This provides privacy better than that of recent approach. Therefore, it is better to break the associations between the attribute which are not related for better result. We measure correlation using above formula.

$$\Phi^2(A1,A2)=\frac{1}{\min\{d1,d2\}-1}\sum_{i=1}^{d1}\sum_{j=1}^{d2}\frac{\{fij-fi*fj\}^2}{fi*fj}$$

Evaluate the quality of the anonym zed data for classifier learning, which has been used in the Naive Bayes Classification which give the better accuracy as compare to other classification technique like Decision Tree. In our experiments, choosing one attribute as the target attributes (the attribute on which the classifier is built) and all other attributes serve as the predictor attributes.

IV. GSLICING APPROACH ALGORITHM

Steps-

1. Compute Maximum of Attribute a {a1; a2;…an} and minimum of a{a1; a2;… an}
2. Now Initialization of centers of the clusters.
3. Then choose the centers as the first k distinct values of the attribute A.
4: Arrange them in increasing order i. e. such that C[1] < C[2] < : : : < C[k]:
5: Define boundary points for b0.
6. for loop i= 1 to n
7. Now find the closet cluster to ai.
8. Recomputed the centers of the clusters as the average of the values in each cluster.
9. Find the closest cluster to *ai* from the possible clusters.
10. Determination of the cut points t0.
11. Count the Fake Tuple and Original Tuple.
12. Count Matching Buckets for Fake Tuple.

Output-

More Protected Sliced data.

Some Important Formulae used in Implementation Process

Calculation of l-diversity:

1. Calculate the k-anonymity (K) = Take average of all unique content of attributes / total no of attributes. i.e. K = (a + b + : : : + d)=n:
2. L-diversity= Total no of tuples / k-anonymity.

V. RESULT

We use OCC 15 UCI Repository Adult Data Set for experimental purpose available on website which consist of total 15 attribute includes both continuous and categorical type.

Table 1: UCI Repository Adult Data Set.

	Attribute	Type	# of values
1	Age	Continuous	74
2	Workclass	Categorical	8
3	Final-Weight	Continuous	NA
4	Education	Categorical	16
5	Education-Num	Continuous	16
6	Marital-Status	Categorical	7
7	Occupation	Categorical	14
8	Relationship	Categorical	6
9	Race	Categorical	5
10	Sex	Categorical	2
11	Capital-Gain	Continuous	NA
12	Capital-Loss	Continuous	NA
13	Hours-Per-Week	Continuous	NA
14	Country	Categorical	41
15	Salary	Categorical	2

In this paper proposed work focuses on the Privacy Preservation Data mining approach using the cluster based attribute Preservation. Typically focused on the distance between two attribute on the basis on the correlation the attribute having smaller distance between them are correlated into the cluster mainly form total 9 clusters in our proposed work.

another important aspect of the work is focused on multi attribute based correlation calculation. The distance from one attribute to another attribute let us consider example suppose attribute like A1 to A15 now first calculate the distance between two attribute like A1 to A2 then similarly. The distance from same A1 attribute to the A3, A4, A5, A6….A15 Attribute then same process apply from again A2 to A15 attribute. The distance from the each attribute is calculated by the contingency formula so mainly used the mean square contingency formula. Then calculate the phi square distance between each attribute.

Second important evaluation of our approach is calculation of the fake tuple from each dataset are used. In our experiment the ADULT Dataset is used which consist of the total near about 45000 attribute out of this many tuple are fake tuple this is mainly added into dataset because to preserve the original dataset privacy. So it is an important that to find out the more fake tuple because to preserve the privacy it is necessary. When find the more fake tuple that can reduces the ration to matching bucket. So more fake tuple indicate less matching bucket relationship. The result achieved by new proposed GSlicing Algorithm.

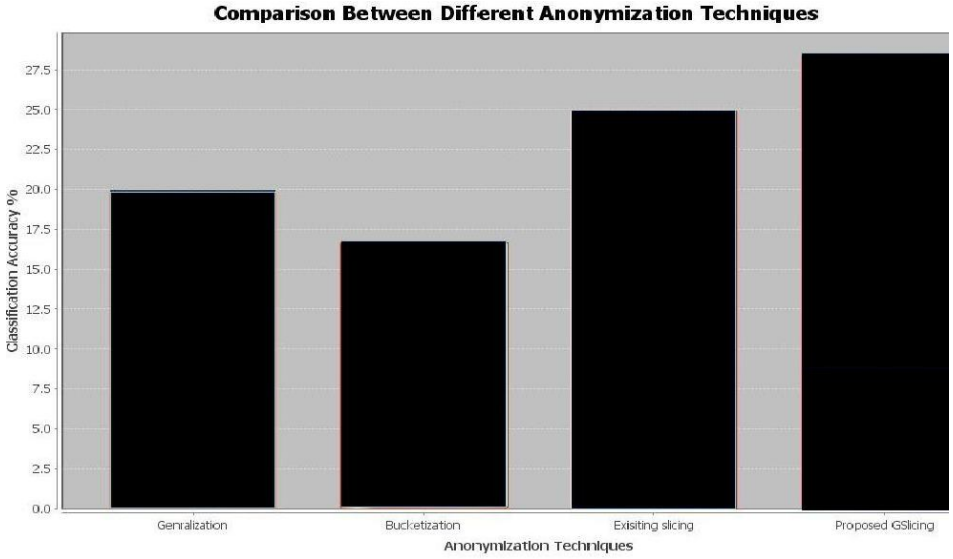
Third important aspects here using Discretization method by type unsupervised learning that can reduce the time complexity and no need of sorting the data ie attribute value in each column mainly used k means clustering algorithm that can provide the better result as compare to the other result of different algorithm. In K-means clustering the centroid used to find the distance between each data point near to cluster, then also calculate the cut point distance.

Therefore Membership discloser Protection is possible by using the Discretization method the advantage of this method is it dynamically searches the different attribute for the purpose of clustering. It is faster method due to unsupervised type of leaning it also provide the more privacy due to labeling of column value the another important aspect is it shows the attribute matching combination then labeling the column the certain value changed in that column also shown the value that change with-in that column. The following details shown about the dataset used.

age:61

Workclass:7

Finalweight:518
Education:16
Educationnumber:16
Maritalstatus:7
Relationship:6
Race:5
sex:2
capitalgain:20
capitalloss:21
hrsperweek:46
country:27
salary:2
Occupation:13
Total Number of tuples:520
Existing slicing: number of fake tuples::459
Existing slicing: number of matching tuples::436
Proposed GSlicing number of fake tuples::486
Proposed GSlicing: number of matching tuples::462



CONCLUSION

In the last several years, many Anonymization techniques have been proposed by various authors. Here, to analyze and classify many previous Anonymization methods for the knowledge of them and a help for new researchers in native areas. After classified the previous works from the various aspects like Membership Disclosure Protection, Attribute Data Privacy for the high dimensional of dataset such as Adult data set. So, there is need to develop a new method which solve all these problems with keeping data utility protection. After reviewing the number of techniques for Privacy Preservation data mining. Generalization grater information loss for high dimensional data while bucketization not taken separate Sensitive and Non sensitive attribute. As per security concern it is important that matching bucket is less for the fake tuple we show this work by our novel algorithm.
Here mainly used the l-diversity method to measure the privacy but there is another method like t-

closeness is subject to future work for measure the privacy. Our focused is on unsupervised learning algorithm but there is also need focus on the various options available in Supervised Learning algorithm this work is also subject to future work.

REFERENCES

[1] A.Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam , l-Diversity: Privacy beyond k-anonymity. Proceedings of the 22nd International Conference on Data Engineering. Pp:24-35,Atlanta , Georgia , 2006

[2] Aggarwal C. C., Yu P. S.: On Variable Constraints in Privacy-Preserving Data mining. SIAM Conference, 2005

[3] B.C.M. Fung, K. Wang, and P.S. Yu, "Top-Down Specialization for Information and Privacy Preservation," Proc. Intl Conf. Data Eng. (ICDE), pp. 205-216, 2005.

[4] C. C. AGGARWAL, PHILIP S. YU: Privacy-preserving data mining: Models and Algorithms, Kluwer Academic Publishers Boston/Dordrecht/London.

[5] Clifton C., Kantarcioglou M., Lin X., Zhu M.: Tools for privacy-preserving distributed data mining. ACM SIGKDD Explorations, 4(2), 2002.

[6] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy,DThomas and A. Zhu.Anonymizing tables. In Proc. of the 10th International Conference on Database Theory(ICDT05), pp. 246-258Edinburgh, Scotland. 2005

[7] G. Ghinita, Y. Tao, and P. Kalnis, "On the Anonymization of Sparse High-Dimensional Data," Proc. IEEE 24th Intl Conf. Data Eng. (ICDE), pp. 715-724,2008.

[8] J.H. Friedman, J.L. Bentley, and R.A. Finkel, "An Algorithm for Finding Best Matches in Logarithmic Expected Time," ACM Trans. Math. Software, vol. 3, no. 3, pp. 209-226, 1977.

[9] K LeFevre, D DeWitt , R Ramakrishnan. Incognito:Efficient full domain k-anonymity Proceedings of the ACM SIGMOD International Conference on Management of Data. Baltimore , Maryland , 2005: 49-60

[10] Kifer D., Gehrke J.: Injecting utility into anonymized datasets. SIGMOD Conference, pp. 217-228, 2006.

[11]Koudas, D. Srivastava, T. Yu, and Q. Zhang,"Aggregate Query Answering on Anonymized Tables," Proc. IEEE 23rd Intl Conf. Data Eng. (ICDE), pp. 116-125, 2007.

[12] L. Sweeney. k-Anonymity: A Model for Protecting Privacy. International Journal on Uncertainty Fuzziness Knowledge based Systems,10(5), pp 557-570. 2002

