www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 2. Issue 10 October 2013 Page No. 3004-3007

PRIVACY PRESERVATION OF DATA PUBLISHING BASED ON NOISE ENABLED SLICING APPROACH

N.Sathya[1], C.Grace Padma[2]
[1]Research Scholar, Department of Computer Science, RVS college of Arts and Science
Coimbatore, Tamil Nadu, India
Sathya29natarajan@gmail.com
[2]Associate Professor & HOD, Department of Computer Application (MCA), RVS College of Arts and Science
Coimbatore, Tamil Nadu 641402, India

Abstract

The basic idea of slicing is to break the association cross columns, but to preserve the association within each column. This reduces the dimensionality of the data and preserves better utility than generalization and bucketization. Slicing preserves utility because it groups highly correlated attributes together, and preserves the correlations between such attributes. Proposed system efficient slicing algorithm to achieve ℓ-diverse slicing. Given a microdata table T and two parameters c and ℓ , the algorithm computes the sliced table that consists of c columns and satisfies the privacy requirement of ℓ -diversity. For measuring the correlation coefficient using pearson and chi squared correlation coefficient in attribute partitioning step for ℓ -diversity slicing. Slicing protects privacy because it breaks the associations between uncorrelated attributes, which are infrequent and thus identifying. Proposed system work in the following manner: attribute partitioning, attribute clustering, tuple partitioning and Analyzing the slicing using Noise enabled slicing. In first step for performing the attribute partitioning ,First compute the correlations between pairs of attributes and sensitive attributes on their correlations using the Chi squared and Pearson based correlation coefficient and then cluster attributes based on their correlations using the Chi squared and Pearson based correlation coefficient .It improves the accuracy of the system for partitioning the result, After these steps finished we perform ,By evaluation of the result by adding the noise data to sensitive attributes for both Chi squared and Pearson based L-diversity slicing. Experimental results shows that the proposed system improves the data utility and privacythentheexistingslicingmethods

INTRODUCTION

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified.[12]Data mining, *the extraction of hidden predictive information from large databases*, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses.[6] Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions.

Clustering is a mathematical tool that attempts to discover structures or certain patterns in a data set, where the objects inside each cluster show a certain degree of similarity. Clustering is a collection of data objects, similar to one another within the same cluster and are dissimilar to objects in the other clusters. Similar to classification, clustering is the organization of data in classes. However, unlike classification, in clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes.

Proposed data anonymizationtechnique called slicing to improve the current state of threat. Slicing partitions the data set both vertically and horizontally. Vertical partitioning is done by grouping attributes into columns based on the correlations among the attributes. Each column contains a subset of attributes that are highly correlated. Horizontal partitioning is done by grouping tuples into buckets. Finally, within each bucket, values in each column are randomly permutated (or sorted)to break the linking between different columns.

EXISTING SYSTEM

This system presents a new approach called slicing to privacy preserving micro data publishing. First, they introduced slicing as a new technique for privacy preserving data publishing. Slicing has several advantages when compared with generalization and bucketization. It preserves better data utility than generalization also preserves more attribute correlations with the SAs than bucketization. It can also handle high-dimensional data and data without a clear separation of QIs and SAs. Second, they showed that slicing can be effectively used for preventing attribute disclosure, based on the privacy requirement of ℓ-diversity. They introduced a notion called ℓ- diverse slicing, which ensures that the adversary cannot learn the sensitive value of any individual with a probability greater than 1/ℓ. Slicing overcomes the limitations of generalization and bucketization and preserves better utility while protecting against privacy threats. Slicing partitions the data set both vertically and horizontally. Vertical partitioning is done by grouping attributes into columns based on the correlations among the attributes. Each column contains a subset of attributes that are highly correlated. Horizontal partitioning is done by grouping tuples into buckets. Finally, within each bucket, values in each column are randomly permutated (or sorted) to break the linking between different columns.

DISADVANTAGES

- Space and time complexity
- Have to reduce the memory space to store the data

PROPOSED SYSTEM

We consider slicing where each attribute is in exactly one column. An extension is the notion of overlapping slicing, which duplicates an attribute in more than one column. This release more attributes correlations. This could provide better data utility, but the privacy implications need to be carefully studied and understood. In this phase, tuples are generalized to satisfy some minimal frequency requirement. We want to point out that column generalization is not an indispensable phase in our algorithm. As shown by Xiao and Tao, bucketization provides the same level of privacy protection as generalization, with respect to attribute disclosure. Although column generalization is not a required phase, it can be useful in several aspects. First, column generalization may be required for identity/membership disclosure protection. If a column value is unique in a column (i.e., the column value appears only once in the column), a tuple with this unique column value can only have one matching bucket. This is not good for privacy protection, as in the case of generalization/bucketization where each tuple can belong to only one equivalence-class/bucket. The main problem is that this unique column value can be identifying. In this case, it would be useful to apply column generalization to ensure that each column value appears with at least some frequency. So we mainly focus on the tuple partitioning algorithm. The trade-off between column generalization and tuple partitioning is implemented effectively. Existing anonymization algorithms can be used for column generalization, e.g., Mondrian. The algorithms can be applied on the sub table containing only attributes in one column to ensure the anonymity requirement.

Several changes made to improve the accuracy of the system.

In first step for performing the attribute partitioning ,First compute the correlations between pairs of attributes and sensitive attributes on their correlations using the Chi squared and Pearson based correlation coefficient and then cluster attributes based on their correlations using the Chi squared and Pearson based correlation coefficient .It improves the accuracy of the system for partitioning the result, After these steps finished we perform ,By evaluation of the result by adding the noise data to sensitive attributes for both Chi squared and Pearson based L-diversity slicing. Computed the correlations for each pair of attributes, we use clustering to partition attributes into columns. In our algorithm, each attribute is a point in the clustering space. The distance between two attributes in the clustering space done by using Pearson and Chi squared based correlation coefficient and then the original attribute correlation coefficient is based on the attribute coefficient .

ADVANTAGES

- The proposed system ensure the anonymity requirement.
- Improves the accuracy of the system and performance.
- Tradeoff between column generalization and tuple partitioning is implemented effectively.

RESULTS AND DISCUSSION

TABLE 1: Adult Data Set

Data Set Characteristics:	Multivariate	**Number of Instances:**	48842	**Area:**	Social
Attribute Characteristics:	Categorical, Integer	**Number of Attributes:**	14	**Date Donated**	1996-05-01
Associated Tasks:	Classification	**Missing Values?**	Yes	**Number of Web Hits:**	279908

Adult Data Set
Dataset Information

Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0)) Prediction task is to determine whether a person makes over 50K a year.

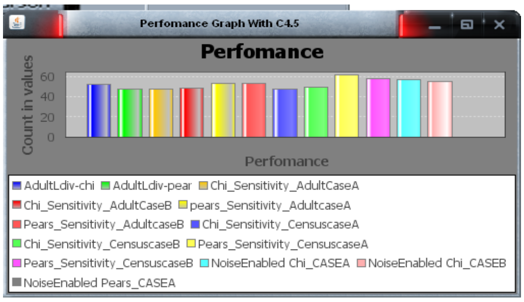

FIGURE 1: Performance Graph For Value

Not-in-family, Other-relative, Unmarried.
Race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
Sex: Female, Male.
Capital-gain: continuous.
Capital-loss: continuous.
Hours-per-week: continuous.
native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad& Tobago, Peru, Hong, Holand-Netherlands.

In this graph measure the performance of the proposed and existing system in performance with C4.5 decision tree based algorithms.

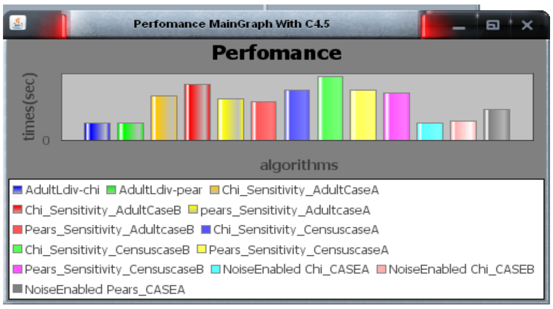

FIGURE 2: Performance For Graph Time

In this graph we measure the performance of the system with time taken to complete process than the existing system and proposed system .The tabulated values are given below.

In this graph we measure the performance of the Chi_senstivity AdultCaseA , Chi_senstivity AdultCaseB, Census_senstivity AdultCaseA and Census_senstivity AdultCaseB system with time taken to complete process than the existing system and proposed system .The tabulated values are given below.

Name	Performance
Chi_senstivity AdultCaseA	80
Chi_senstivity AdultCaseB	79
Census_senstivity AdultCaseA	78
Census_senstivity AdultCaseB	75

TABLE 2: CHI SQUARE COMPARISION FOR DATA SETS

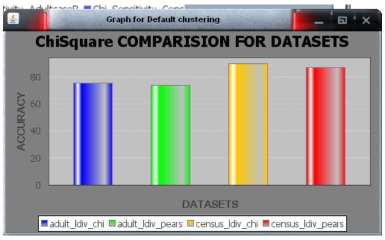

FIGURE 3: CHI SQUARE COMPARISION FOR DATA SETS

In this graph we measure the performance of the Chi_ldiv AdultCaseA,Pears_ldiv AdultCaseA, Census_ldivAdultCaseA and Census_ldiv AdultCaseA system with accuracy result to complete process than the existing system and proposed system .The tabulated values are given below.

CONCLUSION

Conclusion of this phase provides we proposed a new approach called slicing to privacy preserving microdata publishing. Computed the correlations for each pair of attributes using the Pearson and Chi squared based correlation coefficient for attribute partitioning .Clustering to partition attributes into columns. In our algorithm, each attribute is a point in the clustering space. The distance between two attributes in the clustering space done by using Pearson and Chi squared based correlation coefficient and then the original attribute correlation coefficient is based on the attribute coefficient. Proposed system Slicing overcomes the limitations of generalization and bucketization and preserves better utility while protecting against privacy threats. It prevents attribute disclosure and membership disclosure. Attribute correlations can be used for privacy attacks. Our experiments show that slicing preserves better data utility than generalization and is more effective than bucketization in workloads involving the sensitive attribute with noise data was added to the sensitive attributes in both Pearson and Chi squared based l –diversity slicing . The Proposed slicing framework can design better data anonymization techniques to know the data better.

FUTURE ENHANCEMENT

In future work we introduce an extension is the notion of overlapping slicing, which duplicates an attribute in more than one column. This could provide better data utility, but the privacy implications need to be carefully studied and study membership disclosure protection in more details.

REFERENCES

[1] C. Aggarwal, “On k-Anonymity and the Curse of Dimensionality,” Proc. Int’l Conf. Very Large Data Bases (VLDB), pp. 901-909, 2005.

[2] A. Blum, C. Dwork, F. McSherry, and K. Nissim, “Practical Privacy: The SULQ Framework,” Proc. ACM Symp. Principles of Database Systems (PODS), pp. 128-138, 2005.

[3] J. Brickell and V. Shmatikov, “The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing,” Proc. ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD), pp. 70-78, 2008.

[4] B.-C. Chen, K. LeFevre, and R. Ramakrishnan, “Privacy Skyline: Privacy with Multidimensional Adversarial Knowledge,” Proc. Int’l Conf. Very Large Data Bases (VLDB), pp. 770-781, 2007.

[5] H. Cramt’er, Mathematical Methods of Statistics. Princeton Univ. Press, 1948.

[6] I. Dinur and K. Nissim, “Revealing Information while Preserving Privacy,” Proc. ACM Symp. Principles of Database Systems (PODS), pp. 202-210, 2003.

[7] C. Dwork, “Differential Privacy,” Proc. Int’l Colloquium Automata, Languages and Programming (ICALP), pp. 1-12, 2006.

[8] C. Dwork, “Differential Privacy: A Survey of Results,” Proc. Fifth Int’l Conf. Theory and Applications of Models of Computation (TAMC),pp. 1-19, 2008.

[9]Noise to Sensitivity in Private Data Analysis,” Proc. Theory of Cryptography Conf. (TCC), pp. 265-284, 2006.

[10] 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.