Data Mining: Data Utility and Privacy Preservation: A Review

Tulesh Kumar Sahu[1], Rahul Kumar Chawda[2]
[1]*Student of MCA,* [2]*Assistant Professor*
Department of Computer Science, Kalinga University, Raipur.

Abstract **-** Data Mining plays a vital role in today's information world where it has been widely applied in various organizations. The current trend needs to share data for mutual benefit. However, there has been a lot of concern over privacy in the recent years .It has also raised a potential threat of revealing sensitive data of an individual when the data is released publically. Various methods have been proposed to tackle the privacy preservation problem like anonymization and perturbation. But the natural consequence of privacy preservation is information loss. The loss of specific information about certain individuals may affect the data quality and in extreme case the data may become completely useless. There are methods like cryptography which completely anonymize the dataset and which renders the dataset useless. So the utility of the data is completely lost. We need to protect the private information and preserve the data utility as much as possible. So the objective of the thesis is to find an optimum balance between privacy and utility while publishing dataset of any organization. Privacy preservation is hard requirement that must be satisfied and utility is the measure to be optimized.

Keyword **-** Data Utility, Privacy Preservation, Data Mining

I. INTRODUCTION

The amount of data that need to be processed to extract some useful information is increasing. Therefore different data mining methods are adopted to get optimum result with respect to time and utility of data. The amount of personal data that can be collected and analyzed has also increased. Data mining tools are increasingly being used to infer trends and patterns. In many scenarios, access to large amounts of personal data is essential in order for accurate inferences to be drawn. However, publishing of data containing personal information has to be restricted so that individual privacy is not hampered. One possible solution is that instead of releasing the entire database, only a part of it is released which can answer the adequate queries and do not reveal sensitive information. Only those queries are answered which do not reveal sensitive information. Sometimes original data is perturbed and the database owner provides a perturbed answer to each query. These methods require the researchers to formulate their queries without access to any data. Sanitization approach can be used to anonymize the data in order to hide the exact values of the data. But conclusion can't be drawn with surety. Another approach is to suppress some of the data values, while releasing the remaining data values exactly. But suppressing the data may hamper the utility. A lot of research work has been done to protect privacy and many models have been proposed to protect databases. Out of them, k-anonymity has received considerable attention from computer scientist. Under k-anonymity, each piece of disclosed data is equivalent to at least k-1 other pieces of disclosed data over a set of attributes that are deemed to be privacy sensitive.

II. CONCEPT ON DATA MINING

A. What is Data Mining? - Data mining is a technique that helps to extract useful information from a large database. It is the process of extracting relevant information from large databases through the use of certain data mining algorithms. As the amount of data doubles every three years, data mining is becoming an increasingly important tool to transform this data into information. Data mining techniques takes a long time which requires long process of research and product development. This evolution started with storing of business data on computers, continued with improvements in data access, and more recently, generated technologies that allow users to search their data in real time. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

B. Methods of Data Mining - The Amount of data that need to be processed to extract some useful information is increasing. So the methods used for extracting information from huge amount of data must be optimum. As described in the various data mining algorithms can be classified into two broad categories.

i). Heuristic-based approaches
- additive noise
- multiplicative noise
- k-anonymization
- statistical disclosure control based approaches

ii). Cryptography -based approaches

C. Samarati's Algorithm for K-anonymization - Samarati proposed an algorithm for k-anonymization in 2001. This algorithm uses generalization and tuple suppression over quasi-identifiers to obtain a k-anonymized table with maximum suppression of MaxSup tuples. This algorithm uses binary search on the generalization hierarchy

to save time. It assumes that a table PT with more than k attributes is present which is to be k-anonymized.

Given a table PT and a generalization hierarchy, different possible generalizations exist. Not all generalizations, however, can be considered equally satisfactory. For instance, the trivial generalization bringing each attribute to the highest possible level of generalization, thus collapsing all tuples in T to the same list of values, provides k-anonymity at the price of a strong generalization of the data. Such extreme generalization is not needed if a more specific table (i.e., containing more specific values) exists which satisfies kanonymity. A naïve approach to compute a k-minimal generalization would then consist in following each generalization strategy (path) in the domain generalization hierarchy stopping the process at the first generalization that satisfies k-anonymity. However this approach becomes impractical when number of paths increase. A better approach to find k-minimal generalization is proposed in [4]. In this approach concept of distance vector is induced and exploited. Let PT be a table and x,y Є PT be two tuples such that x =(v1.....vn) and y=(v1‟……vn‟) where vi and vi‟ are values in domain Di The distance vector between x and y is the vector Vx,y = [d1……..dn] where di is the (equal) length of the two paths from vi and vi‟ to their closest common ancestor in the value generalization hierarchy VGHDi (or, in other words, the distance from the domain of vi and vi‟ to the domain at which they generalize to the same value vi). Assume Dept, C.G., Age and Roll No. to be a quasi-identifier. The distance vector between (CIV, 7.5, 20, 10601012) and (CIV, 8.6, 21, 10601026) is [0,1,1,1], at which they both generalize to (CIV,>7,>20,106010**).

III. CONCLUSION

In order to improve the privacy offered by the dataset, utility of the data suffers. On conducting the experiments we found that the balancing point between utility and privacy depends on the dataset and value of k cannot be generalized for all datasets such that utility and privacy are balanced. On varying the number of sensitive attributes in a dataset the balancing point varies. We found that if number of quasi-identifiers increases balancing point moves down and balance between utility and privacy occurs at a higher value of k. Thus if a dataset contains more number of quasi-identifiers then the utility as well as privacy attained at balancing point will be less than the dataset having fewer quasi-identifiers.

IV. REFERENCES

[1]. Bertino, D. Lin, W. Jiang (2008). A Survey of Quantification of Privacy. In: Privacy-Preserving Data Mining. Springer US, Vol 34, pp. 183-205.

[2]. R. J. Bayardo, R. Agrawal (2005). Data privacy through optimal k-anonymization. In: Proc. of the 21st International Conference on Data Engineering, IEEE Computer Society, pp. 217-228.

[3]. K. Liu, H. Kargupta, J. Ryan (2006). Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. IEEE Transactions on Knowledge and Data Engineering, Vol 18(1), pp. 92–106

[4]. P. Samarati (2001). Protecting respondents‟ identities in microdata release. IEEE Transactions on Knowledge and Data Engineering, VOl 13(6), pp. 1010–1027

[5]. L. Sweeney (2002). Achieving k-anonymity privacy protection using generalization and suppression. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems, Vol 10(5), pp. 571–588.

[6]. V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati (2007). kAnonymity. In: Secure Data Management in Decentralized Systems. Springer US, Vol 33, pp. 323-353.

[7]. V. S. Verykios, E. Bertino, I. N. F. L. P. Provenza, Y. Saygin, and Y. Theodoridis (2004). State-of-the-art in Privacy Preserving Data Mining. ACM SIGMOD Record, Vol 33(1), pp. 50-57.

[8]. L. Sweeney (2002). k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, Vol 10 (5), pp. 557-570. 42

[9]. A. Machanavajjhala, J. Gehrke ,D. Kifer, M. Venkitasubramaniam (2007). ℓDiversity: Privacy Beyond k-Anonymity. ACM Transactions on Knowledge Discovery from Data, Vol 1(1), Article: 3.

[10]. 10. R. Agrawal, R. Srikant (2000). Privacy preserving data mining. ACM SIGMOD Record, Vol 29(2), pp. 439–450.

[11]. M. R. Z. Mirakabad and A. Jantan (2008). Diversity versus Anonymity for Privacy Preservation. The 3rd International Symposium on Information Technology (ITSim2008), Vol 3, pp. 1-7.

[12]. J.Lin, and M. Wei (2008). An Efficient Clustering Method for k-Anonymization. In: Proceedings of the 2008 international workshop on Privacy and anonymity in information society, Vol. 331, pp. 46-50.

[13]. UCI Repository of machine learning databases, University of California, Irvine. http:/archive.ics.uci.edu/ml/.

[14]. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten (2009). The WEKA Data Mining Software: An Update. SIGKDD Explorations, Vol 11(1).