Anonymization of high-dimensional data by dimensionality reduction using feature selection

U.SARANYA
Post Graduate Student - CSE
Dr.Mahalingam College of Engineering and Technology,
Pollachi.
Email: usaranya.kumar@gmail.com

MR.K.THIRUKUMAR M.E
Assistant Professor (SG) - CSE
Dr.Mahalingam College of Engineering and Technology,
Pollachi.
Email:thirukumar@drmcet.ac.in

Abstract— **Privacy preservation is an important factor for preserving useful information in the published data. Transactional data which involves hundreds or thousands of dimensions requires protection of individuals in many applications. The anonymization approach is used to preserve privacy and the gray encoding based sorting handles the sparsity of data. Correlation-aware anonymization of high-dimensional data (CAHD) results in anonymized group formation that ensures the privacy. Preserving privacy becomes inefficient due to the curse of dimensionality. Feature selection using genetic algorithm is used on transactional database to reduce the attributes. Anonymization is performed on the attributes with maximum privacy strength and maximum informativeness. An anonymized group formation of high-dimensional data is expected to reduce the execution time in comparison with the existing system.**

.

I. INTRODUCTION

A database with sensitive information can be anonymized in order to eliminate privacy threats. Anonymization focuses on retaining the usefulness of data in the development of further applications in various fields.

Anonymization enables to transfer the information about an individual to reduce the risk of disclosure. Database anonymization is the key to secure the databases and unable to gain access the sensitive personal information. To eliminate the privacy risk, the data should be anonymized.

De-identification of sensitive attributes from their corresponding identifiers should be performed on the database. Anonymization ensures privacy that allows researchers to be provided with useful data. There are different types of anonymization techniques that depends on cost, complexity, and robustness. K-anonymity and l-diversity are the anonymization techniques used in various situations.

K-anonymity states that every record in the table is indistinguishable from at least k -1 other records with respect to every set of quasi-identifier attributes. K-anonymity decreases the utility with increase in size of the data.

K-anonymity suffers from severe privacy problems due to homogeneity attack and background knowledge attack. L-diversity contains atleast 'l' well represented values for the sensitive attributes. L-diversity does not provide protection against attribute disclosure.

There are several anonymization operations such as anatomization, permutation and perturbation that are carried out to reduce the risk of re-identification of sensitive attributes. Anatomization approach preserves privacy and correlation by releasing all the quasi-identifier and sensitive values into two separate tables.

Permutation is based on partitioning a set of data records into groups and shuffling their sensitive values within each group is performed. Perturbation is used to replace the original value with some synthetic data values.

Anonymization on high dimensional data is difficult to preserve privacy. Data reorganization method should be used to create anonymized group formation. The privacy requirements should satisfy both the high dimensionality and sparsity characteristics in transactional database.

II. RELATED WORK

Many anonymization techniques are used to preserve privacy. There are number of techniques for modifying or transforming the information to preserve privacy.

Sumana, Hareesh (2010) explains the issues in k-anonymity and l-diversity techniques[1]. The k-anonymity focused on generalization and suppression which results in loss of information. The l-diversity deals with generalization and permutation based approach which does not preserve correlation between quasi-identifier and sensitive attributes.

Different anonymization methods have been used for privatizing the data. K-anonymity does not prevent attribute disclosure [8]. The notion of l-diversity have been proposed to address the attribute disclosure.

Aggarwal and Yu describes the concept of privacy preservation in data mining. The author discussed the topic dimensionality curse and the difficulties to preserve privacy on high dimensional data [2].

The author examined different privacy preservation methods but found to be either infeasible or ineffective on curse of dimensionality [10].

Tiwari, Singh (2010) explains the concepts of feature selection using genetic algorithm to reduce the dimensionality [3]. It focuses on integration of data sources to build a data warehouse for a large number of related data.

The preprocessed data should be stored in data warehouse for data mining process. Feature selection is one of the important and frequently used techniques in data preprocessing for data mining [9]. It reduces the number of features, removes irrelevant, redundant, or noisy data, and used in many applications. Genetic algorithm is an optimization technique among the available techniques for attribute selection.

Reddy and Kumari (2011) explains the computation of privacy strength and informativeness for all possible subsets [4]. Dimensionality reduction and attribute selection aims at choosing a subset of attributes sufficient to describe the data set. Reduction from dimension d to k (k<d) reduces complexity and provides privacy.

To preserve privacy, the privacy strength and informativeness is defined for each of the data sources in an organization. Privacy strength (P) of a given node is the number of leaves in the sub tree with this node as the root. Informativeness (I) is defined as the ratio between the path length from root to the current node to sum of the path length from root to the current node and length of the tree with current node as the root. The best possible subset that produces maximum privacy and maximum informativeness should be selected [11].

Canahuate et al (2006) suggest that evaluating columns using gray code sorting improves the execution time [5]. The gray code ordering to reorganize the data is optimal in terms of memory requirements. It provides the bit-level similarity between consecutive data elements. Since consecutive numbers differ only at one bit, Gray code numbers have maximum bit-level similarity between consecutive numbers. The performance of gray code ordering is found to be better [12].

Rajalakshmi et al(2011) explains the performance and quality of various privacy preserving algorithms [6]. The performance of Correlation-aware Anonymization of High dimensional Data (CAHD) is discussed and found to be less efficient for large amount of data.

III. EXISTING WORK

Privacy preservation focused on anonymizing transactional data with hundreds or thousands of dimensions that may contain the purchased details of some sensitive attributes like pregnancy test and Viagra. Hence anonymization should be performed effectively and data must be published without revealing sensitive information.

The data in the transactional dataset is represented in binary format. The binary representation takes longer time for execution and results in computational overhead. In the existing system, the author suggests that gray encoding based sorting is best technique to perform data transformation.

A. DATA REORGANIZATION

Gray code sorting technique is a data reorganization technique used to preserve correlation and anonymizing transactional data. Algorithm for Gray sort is given in Fig1.

GraySort(T)
Input: transaction set T
1. **foreach** $t \in T$
2. $t'[Q] = compute\_\mathcal{G}^{-1}(t[Q])$
3. sort T increasingly on key $t'[Q]$
4. output T

compute $\mathcal{G}(t')$
Input: transaction t' in natural binary code
5. $t[|Q|] = t'[|Q|]$
6. **for** $i = |Q| - 1$ **downto** 1 **do** $t[i] = t'[i+1] \oplus t'[i]$
7. **return** t

compute $\mathcal{G}^{-1}(t)$
Input: transaction t in Gray binary code
8. $t'[|Q|] = t[|Q|]$
9. **for** $i = |Q| - 1$ **downto** 1 **do** $t'[i] = t'[i+1] \oplus t[i]$
10. **return** t'

Fig 1: Graysort Algorithm

The main objective is to minimize the hamming distance between the consecutive data elements in the transactional data. Due to higher dimensionality of attributes in the dataset, the sparsity of the data should be handled.

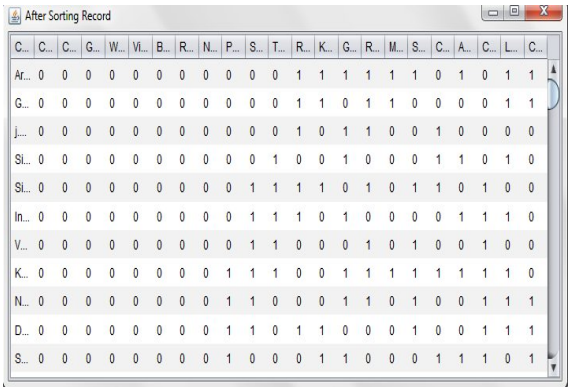

Fig 2: Data Reorganization Snapshot

To provide higher correlation, the identical transactions are placed in consecutive order which facilitates the transactions with similar quasi-identifier in the same anonymized group.

B. CORRELATION-AWARE ANONYMIZATION OF HIGH-DIMENSIONAL DATA (CAHD)

The next step after data reorganization is to create anonymized groups of transactions. Correlation-aware Anonymization of High- dimensional Data (CAHD) technique groups the transactions together that are nearer to each other in the data sequence.

CAHD technique scans each row linearly in the transaction set T to find the first sensitive transaction in the data sequence and form an anonymized group formation. Likewise, each sensitive transaction forms a group consisting of non-conflicting transactions such that degree (G)>=P [7].

The disadvantage of existing system is preserving privacy in high-dimensional dataset is difficult. Due to the sparsity of elements in the transactional data, preserving correlation on anonymized high dimensional data is inefficient.

IV. PROPOSED SYSTEM

In the proposed system feature selection using genetic algorithm is used to handle the high-dimensional data efficiently. Preprocessing is performed on the high-dimensional data set in order to handle the missing values. Generally all the feature in a dataset will not be supportive. Hence Feature selection using genetic algorithm is employed to identify the best set of features.

A.SIMPLE GENETIC ALGORITHM PROCEDURE

Initial Population

A population is a multiset of genotypes. Genetic algorithms are generally stated with an initial population that is randomly generated.

Fitness-Based Selection

In this kind of parent selection, each chromosome has a chance of selection that is directly proportional to its fitness.

Reproduction

The steady-state method selects two chromosomes and performs crossover on them to obtain one or two children, perhaps applies mutation as well, and installs the result back into that population.

Crossover Operator

It merges information from two parent genotypes into one or two offspring genotypes.

Mutation

Mutation has the effect of ensuring that all possible chromosomes are reachable. The mutation operator randomly selects any bit position in a string and make changes on it.

Steps involved in dimensionality reduction:

1. The preprocessed data contains a variety of chromosomes.
2. Privacy strength and informativeness are calculated for each individual chromosome.
3. Perform crossover and mutation operation using the combination of chromosomes.
4. Evaluate the newly generated chromosome.
5. Repeat the process until a chromosome with maximum privacy strength and maximum informativeness is obtained.
6. The reduced set of attributes are produced after feature selection process.

The reduced set of attributes perform Gray-encoding based sorting. The sorted transactions forms an anonymized group using correlation-aware anonymization of high-dimensional data.

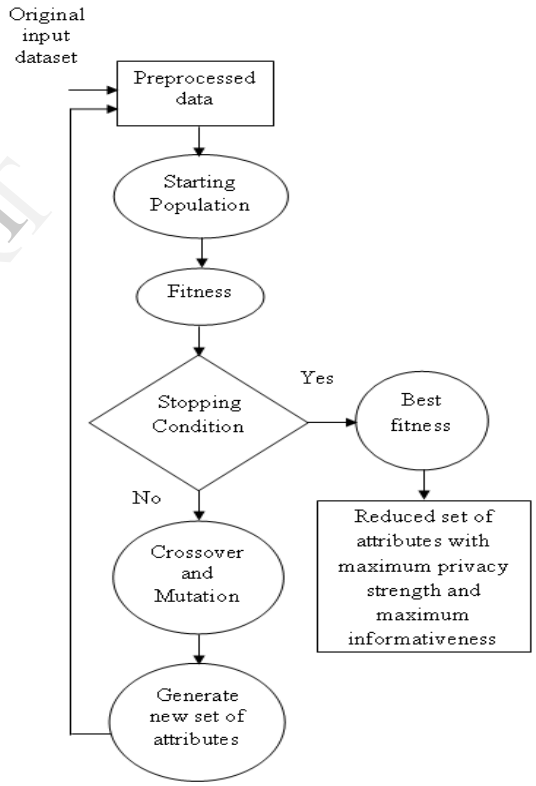

Fig 3: Block Diagram of Proposed System

The performance of proposed method is compared in terms of privacy strength, informativeness and execution time. The dataset after dimensionality reduction is achieved with maximum privacy strength and maximum informativeness.

V. PERFORMANCE EVALUATION

The transactional dataset after dimensionality reduction will be used for performance evaluation. Informativeness and privacy strength are the measures that are evaluated. The privacy strength is defined as the number of sensitive attributes used for anonymization and provides the anonymized group formation with higher privacy.

Co...	C...	C...	G...	W...	Via...	B...	N...	Pre...	S...	TAXOL	R...	G...	RI...	MI...	C...	LA...	CINNAMON
Dh...	1	1	1	1	1925	1	1	1910	1	1949	1	1	0	1	0	1927	0
Ra...	1	1	1	1	1925	1	1	1910	1	1949	1	1	0	1	1	1927	0
Di...	1	1	1	1	1925	1	1	1910	1	1949	1	1	1	1	0	1927	1
Ma...	1	1	1	1	1925	1	1	1910	1	1949	1	1	1	0	0	1927	1
C....	1	1	1	1	1925	1	1	1910	1	1949	1	1	0	1	1	1927	0
Di...	1	1	1	1	1925	1	1	1910	1	1949	1	1	0	0	1	1927	1
Sa...	1	1	1	1	1925	1	1	1910	1	1949	1	0	0	1	1	1927	1
Pa...	1	1	1	1	1925	1	1	1910	1	1949	1	0	0	0	0	1927	0
jeb...	1	1	1	1	1925	1	1	1910	1	1949	0	1	1	1	1	1927	0
Pr...	1	1	1	1	1925	1	1	1910	1	1949	0	1	1	1	1	1927	1
Kri...	1	1	1	1	1925	1	1	1910	1	1949	0	0	0	0	0	1927	1
J.T...	1	1	1	1	1925	1	1	1910	1	1949	0	1	1	1	1	1927	1
Pr...	1	1	1	1	1925	1	1	1910	1	1949	0	1	0	0	1	1927	0
Dh...	1	1	1	1	1925	1	1	1910	1	1949	0	1	0	0	1	1927	1

Fig 4: Anonymized Group Formation Snapshot

The informativeness is defined as the amount of information published after anonymizing the data. The optimal solution is defined as the subset of attributes which satisfy both the maximum privacy strength and maximum informativeness.

VI. CONCLUSION

Anonymization on high-dimensional data is generally difficult due to curse of dimensionality. The data reorganization is performed on the binary representation of data. The feature selection approach based on genetic algorithm is used to obtain the relevant attributes from the dataset. The performance measures are evaluated after dimensionality reduction. The implementation of anonymization on the reduced attributes enhances privacy and reduces execution time.

VII. ACKNOWLEDGEMENT

I would like to express my gratitude to Mr.K.Thirukumar M.E., Assistant Professor (SG), Department of CSE, Dr.Mahalingam College of Engineering and Technology for his useful comments, remarks and engagement through the learning process of this project. Furthermore I would like to thank Ms.G.Anupriya M.E., Assistant Professor (SG), Department of CSE, Dr.Mahalingam College of Engineering and Technology for introducing me to the topic as well as for the support. Also, I would like to thank my family members, who have supported me throughout entire process, both by keeping me harmonious and helping me. I will be grateful forever for their love.

REFERENCES

[1] Sumana M, Dr Hareesh K S,(2010), "Anonymity: An Assessment and Perspective in Privacy Preserving Data Mining", International Journal of Computer Applications (0975 – 8887), Volume 6– No.10, pp. 1-5.

[2] Charu C. Aggarwal, Philip S. Yu(2000), "Privacy-Preserving Data Mining Models and Algorithms", Springer, pp. 433-460.

[3] Rajdev Tiwari, Manu Pratap Singh, (2010), "Correlation-based Attribute Selection Using Genetic Algorithm", International Journal of Computer Applications(0975-8887), Volume 4-No.8, pp. 28-34.

[4] Ram Prasad Reddy S, KVSVN Raju, Valli Kumari V, (2011), "A Dynamic Programming Approach for Privacy Preserving Collaborative Data Publishing", International Journal of Computer Applications(0975-8887), Volume 22-No.4, pp. 18-23.

[5] Guadalupe Canahuate, Hakan Ferhatosmanoglu, Ali Pinar, (2006), "Improving Bitmap Index Compression by Data Reorganization", IEEE Transactions on Knowledge and Data Engineering, pp.1-35.

[6] Rajalakshmi V, Anandha Mala G S, Balasubramanian R, (2011), "An Estimation of Privacy in Incremental Data Mining", ICCCECS624.

[7] Gabriel Ghinita, Member, IEEE, Panos Kalnis, and Yufei Tao, (2011), "Anonymous Publication of Sensitive Transactional Data", IEEE transactions on knowledge and data engineering, vol. 23, no. 2, pp.161-174.

[8] Li Liu, Murat Kantarcioglu and Bhavani Thuraisingham, (2009), "Privacy Preserving Decision Tree Mining from Perturbed Data", Proceedings of the 42nd Hawaii International Conference on System Sciences, pp. 1-10.

[9] Huan Liu and Lei Yu., (2005), "Toward Integrating Feature Selection Algorithms for Classification and Clustering", IEEE Transactions on Knowledge and Data Engineering, Volume 17, Issue 4, pp: 491 – 502.

[10] Liu K,Kargupta H, and Ryan J, (2006), "Random projection-based multiplicative data perturbation for privacy preserving distributed data mining", IEEE Trans. Knowl. Data Eng., vol. 18, no. 1, pp. 92–106.

[11] Oliveira, Stanley R M., Zaiane Osmar R,(2004), "Privacy preservation when sharing data for clustering", In: Proc. Workshop on Secure Data Management in a Connected World, pp.67-82.

[12] Wang K, Fung B C M., and Dong G, (2005), "Integrating private databases for data analysis", In IEEE ISI, pp.171-182.