Hybrid techniques for Privacy preserving in Data Mining

J.Paranthaman[1], Dr. T Aruldoss Albert Victoire [2]

[1.] Assistant Professor in Computer Science and Engineering, University College of Engineering, Pattukottai Campus, Rajamadam- 614 701.
[2.] Associate Professor, Department of Electrical and Electronics Engineering, Regional Centre-Coimbatore, Anna University, Coimbatore - 641 047.

Abstract: With the technology advancements, most of the corporations maintain their huge amount of electronic data in the databases and these databases are accessible using internet. These data are used by data miners to extract useful information. There is a threat to the privacy of the data while performing data mining tasks. Anonymization is one of the methods that transform actual data using generalization or suppression techniques, so that private information of individuals is masked. K-Anonymity transforms data into a set of equivalence classes and each class has a set of K- records indistinguishable from each other.In this proposed work, k-anonymity is used for privacy preserving while applying data mining algorithms. Hybrid technique simulated annealing with a genetic algorithm is used to optimize the feature selection. For evaluation, the mushroom data set anonymized to different levels for preserving privacy and hybrid technique for optimization is used. Experimental results demonstrate that the proposed method achieve satisfactory results.

[J.Paranthaman, Dr. T Aruldoss Albert Victoire. **Hybrid techniques for Privacy preserving in Data Mining**. *Life Sci J* 2013; 10(7s): 471-475] (ISSN:1097-8135). http://www.lifesciencesite.com.

Keywords: Privacy preserving data mining; k-anonymity; simulated annealing; Genetic algorithm

1. Introduction

In recent years, using recent technology advancements, most of the corporations maintain their huge size of electronic data in the databases and these databases are accessible using internet. Data mining methods are used to find useful information from the raw data available in these databases. Therefore there is a threat to the privacy of the data while performing data mining steps. Some of the data available in these databases are sensitive data and should not be given to users of the information gained from the data mining results. Large number of methods is available for preserving the privacy.

Most of the methods make reduction in the granularity of the representation of the data. This makes information loss but improves privacy. So always there is a trade-off between loss of information and the privacy [1]. Effective techniques are needed by no compromising with security mechanisms. Some of the techniques used for privacy preserving are Randomization method, k-anonymity model and l-diversity and Distributed privacy preservation.

In the randomization method, small amount of noise is added to the existing data so that the actual behavior of the individual records is changed. But the data miner can reconstruct the collective behavior of the data distribution by removing the noise by subtracting the noise from the data. After reconstruction, the distribution is more sufficient for all the data mining tasks [2].For randomization, there are two perturbation tasks are used. 1) Additive Perturbation: Randomized noise is added to the actual data records. 2) Multiplicative Perturbation: For the perturbation of the actual data records random rotation or projection techniques are used.

The k-anonymity model is based on a quasi-identifier. It is a collection of attributes in a database that makes identifier of the entire data. All the data in the database is assumed as a set of tables, and each tuple is information of an individual customer. The quasi-identifier is denoted as Q that contains identical values for Q. some of the tuples in the dataset form an equivalence class with respect to the selected set of attributes. If there is more number of tuples in the equivalence class then the identification of the individual is very difficult. All the tuples in the data set D is k-anonymous with respect to Q when every equivalence class has the size with respect to Q is k or more. So every tuple in the released table will be linked to an individual and privacy of individuals is preserved [3].

General k-anonymity models provide efficiency in preserving privacy if the micro data is not provided. Among the different versions of anonymity model complete (alpha, k) anonymity model sets different frequency constraint alpha on the sensitive data and uses clustering algorithms and distance between equivalence classes and generalization trees. So it provides less data distortion [4].

The ℓ -diversity method uses a principle: An equivalence class is having ℓ-diversity if the class has at least ℓ distinct values for the attributes that are sensitive. A table is having ℓ-diversity if all the equivalence classes of the table has ℓ-diversity [5] and the equivalence class entropy (E)is defined as,

$$\text{Entropy}(E) = -\Sigma s \epsilon S p\,(E,s) \log p(E,s)$$

where S is the domain of the sensitive attribute and p(E, s) is a fraction of records in E with sensitive value s.

A table has entropy if for all the equivalence classes E, Entropy (E) ≥ log ℓ. If entropy is too low then there is a less conservation on diversity. This method does not prevent attribute disclosure. This method suffers from two attacks. Skewness Attack: ℓ-diversity cannot prevent attribute disclosure when the overall distribution is skewed. Similarity Attack: When the values of sensitive attributes and values in an equivalence class are diverse but are based on similar semantics, important information can be learnt.

In some datasets, individual data entities derive aggregate outcomes from entire data sets that can be partitioned across these entities. There are two types of partitioning: horizontal and vertical. In horizontal partitioning, the records are distributed among multiple entities. In vertical partitioning, the attributes are distributed among multiple entities. When the individual entities are unwilling to share complete data sets but are willing to share limited information under a variety of protocols. The result of partitioning and aggregation maintains privacy for each entity in the dataset.

Usually statistics-based and the crypto-based algorithms are used for Privacy reserving Data Mining. In the statistics-based approach, the data owners clean the data through perturbation or generalization before explosion. Knowledge models such as decision trees are used on the data cleaning. Statistics-based approach efficiently handles a large volume of datasets [6]. In the crypto-based methods, data owners implement specially designed data mining algorithms [7]. All these algorithms achieve verifiable privacy protection and better data mining performance, but it suffers from performance and scalability issues [8].

2. Related Work

Sumana and Hareesh analyzed various anonymization techniques in Privacy Preserving Data Mining used for preserving privacy of the data [11]. The main goal of anonymization was to secure access to the sensitive information and at the same time providing aggregate information to the public. Confidential information revealing could be linked to the individuals, so the task was a challenging one. Aggregate data should give minimum loss in the accuracy of data mining results. Many types of anonymization methods were compared. K-anonymity method protected against identity disclosure, it did not provide sufficient protection against attribute disclosure. ℓ-diversity method solved disclosure problem by requiring that each equivalence class has at least ℓ well-represented values for each sensitive attribute A complete (α,k) anonymity model satisfied sensitive values individuation secure requirement by setting frequency constraint for each sensitive value. Complete (α,k) used clustering algorithm. Experimental results showed that the complete (α,k)-anonymity model could preserve privacy effectively with less data distortion.

Tzung-Pei et al presented Evolutionary privacy preserving in data mining [12]. Collection of data, dissemination and mining from large datasets introduced threats to the privacy of the data. Some sensitive or private information about the individuals and businesses or organizations had to be masked before it is disclosed to users of data mining. An evolutionary privacy-preserving data mining method was proposed to find about what transactions were to be hidden from a database. Based on the preference and sensitivity of the individual's data in the database different weights were assigned to the attributes of the individuals. The concept of prelarge item sets was used to minimize the cost of rescanning the entire database and speed up the evaluation process of chromosomes. The proposed approach was used to make a good trade-off between privacy preserving and running time of the data mining algorithms.

Han and Keong Ng presented Privacy-Preserving Genetic Algorithms for Rule Discovery [13]. Entire data set was partitioned between two parties, and genetic algorithm was used to find the best set of rules without publishing their actual private data. Two parties jointly developed fitness function to evaluate the results using each party's private data but not compromising the privacy of the data by Secure Fitness Evaluation Protocol. To meet the privacy related challenges, results generated by genetic algorithm were not compromising privacy of those two parities having partitioned data. Creation of initial population and ranking the individuals for reproduction were done jointly by both parties.

Pei–Ling Chiu [14] presented A Simulated Annealing Algorithm for General Threshold Visual Cryptography Schemes. For improving the quality of still image, simulated annealing method had been used. The problem had been formulated as a mathematical optimization model in order to maximize the contrast of recovered images that are subject to density-balance and blackness constraints. The experimental results showed that the SA optimization-based approach significantly improved the performance of previous methods.

Elhaddad presented Combined Simulated Annealing and Genetic Algorithm to Solve Optimization Problems [15]. Many types of evolutionary algorithms were used for optimization of results. To improve different methods and to get high quality results and in less time hybrid techniques could be used. Genetic Algorithm (GA) and Simulated Annealing (SA) had been combined to solve optimization problems. Both GA and SA methods searched a solution space in iterative manner till the convergence. These two algorithms were significantly different. The GA mechanism was parallel on a set of solutions and exchanged information using the crossover operation. SA works on a single solution at a time. SA and GA were combined in order to minimize the disadvantages of both algorithms.

3. Materials and Methods

K-anonymity techniques are based on the reduction of granularity of representation of data using pseudo-identifiers. Major techniques used for granularity reduction was generalization and suppression. In generalization, the attribute values are converted into a range that reduces the granularity and reduces the risk of identifying individual values. In suppression method, actual value of the attribute is removed completely. But these two methods introduce loss of some detail which may affect the accuracy.

Finding optimal k-anonymous datasets using generalization or suppression has been proved as a NP-hard problem [16, 17]. So some standard heuristic search techniques such as genetic algorithms, particle swam optimization and simulated annealing can be used to find optimal datasets. The disadvantage of GA is that it may get struck in local minima and consumption of time. Simulated annealing is an effective technique used to generate k-anonymous representations of the data. But it works on one solution at a time. But GA searches different solutions in parallel.

Many scientific and practical applications need hybrid optimized solutions. The demerits of one optimization method may be suppressed by the other optimization method. Simulated annealing and genetic algorithm are combined to overcome the disadvantages of both algorithms.

Genetic Algorithms use heuristic search techniques and optimization techniques that mimic the process of natural evolution. At every iteration, it selects best attributes by discarding the remaining attributes. GA proceeds to initialize a population of solutions and then to improve it through repetitive operation of the mutation, crossover, inversion and selection operators [9]. So every time solution is improved. Fitness function is used to evaluate the obtained solution. Simple genetic algorithm is given as,

```
simplegenealg()
{
    Initialize the population;
    Calculate fitness function;
    While (fitness value! = termination condition)
    {
        Perform Selection;
        Perform Crossover;
        Perform Mutation;
        Calculate fitness function;
    }
}
```

Simulated annealing (SA) is a random-search technique based on annealing process. The annealing process consists cooling of metal and freezing into a minimum energy crystalline structure. It forms the basis of an optimisation technique for combinatorial and other problems and search for a minimum in a more general system. Simulated annealing methods can be used for global optimization. Global optimization method searches the global optimal solution in the search space. Since general search techniques are not capable of focusing the search towards some ROI in which contains global optimized solution [9].

In this proposed work, the granularity reduction technique by k-anonymity is used for privacy preserving while performing data mining. Simulated annealing with a genetic algorithm is used to optimize the feature selection.

4. Results and Discussion

Mushroom data set is used, and the proposed algorithm is tested. Results show that the algorithm is able to find an optimal solution or near optimal solution for varying Levels of k in anonymity model. The performance metrics used are average accuracy, precision and recall. Figures 1 to 3 depict the experimental results of the same.

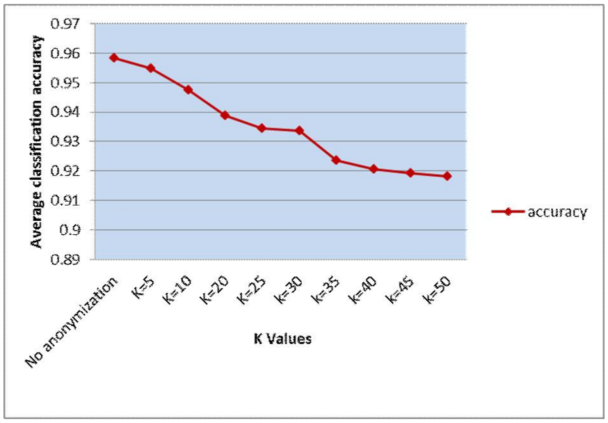

Figure1: Average accuracy

It is apparent from Figure 1 that the performance of the classifier decreases with the increase in the level of anonymity. The accuracy reduces by 4.19% for k=50 when compared with non-anonymized data.

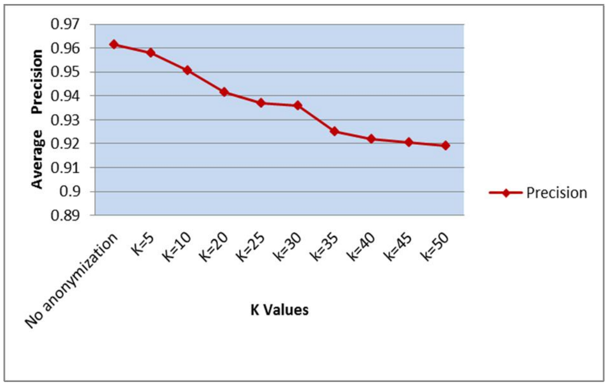

Figure 2: Average precision

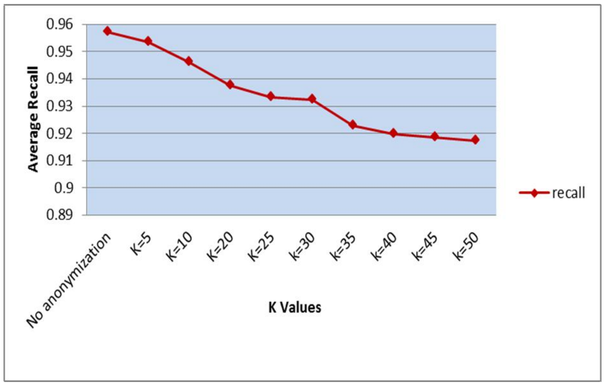

Figure 3: Average recall

From Figure 2 and 3 it is seen that the precision and recall decrease with the increase in k levels.

5. Conclusion

This work proposed a hybrid algorithm for privacy preserving in data mining. K-anonymity method with varying k-levels is used. When mining large data set, evolutionary algorithms such as genetic algorithm and simulated annealing are used to find optimal data set. Mushroom data set has been used for evaluation, and the performance parameters like accuracy, precision and recall were represented graphically with varying k-levels for granularity reduction using k-anonymity. Experimental results demonstrate that with the increase in the k-anonymity levels, the performance of the classifier decreases but within an acceptable level.

References

1. Charu C. Aggarwal And Philip S. Yu, "Privacy-Preserving Data Mining: Models And Algorithms", Springer publications, 2007.
2. Charu C. Aggarwal And Philip S. Yu, " A Survey of Randomization Methods for Privacy-Preserving Data Mining " , Privacy preserving Data Mining Advances in Database Systems , volume 34, 2008.
3. Raymond ChiWing Wong, Jiuyong Li, Ada WaiChee Fuand Ke Wang "(alpha, k) Anonymity: An Enhanced k- Anonymity Model for Privacy Preserving Data Publishing ", ACM, 2006.
4. Han Jian-min, Yu Hui-qun, Yu Juan, Cen Ting-ting "A Complete (α, k)-Anonymity Model for Sensitive Values Individuation Preservation", 2008 IEEE DOI 10.1109/ISECS.2008.92.
5. Ninghui Li, Tiancheng Li and Suresh Venkatasubramanian, "t-Closeness: Privacy Beyond *k*-Anonymity and *ℓ*-Diversity, Citiseer ", 2007.
6. Malin, B., Benitez, K., & Masys, D. (2011). Never too old for anonymity: a statistical standard for demographic data sharing via the HIPAA Privacy Rule. Journal of the American Medical Informatics Association, 18(1), 3-10.
7. Singh, M. D., Krishna, P. R., & Saxena, A. (2010, January). A cryptography based privacy preserving solution to mine cloud data. In Proceedings of the Third Annual ACM Bangalore Conference (p. 14).ACM.
8. Patrick Sharkey, HongweiTian, Weining Zhang, and Shouhuai Xu, 2008, Privacy-Preserving Data Mining through Knowledge Model Sharing, Springer-Verlag Berlin Heidelberg, pp. 97–115, 2008.
9. Mehrdad Dianati, Insop Song, and Mark Treiber "An Introduction to Genetic Algorithms and Evolution Strategies" , 2006.
10. Franco Busetti, "Simulated annealing overview" ,2000.
11. Sumana M, Dr Hareesh K S , " Anonymity: An Assessment and Perspective in Privacy

Preserving Data Mining " International Journal of Computer Applications (0975 – 8887) Volume 6– No.10, September 2010.

12. T zung-Pei, Hong Kuo-Tung Yang, Chun-Wei Lin and Shyue-Liang Wang, "Evolutionary privacy preserving in data mining ", IEEE World Automation Congress conference , 2010.
13. Shuguo Han Wee Keong Ng, "Privacy-Preserving Genetic Algorithms for Rule Discovery", 2007.
14. Pei –Ling Chiu, "A Simulated Annealing Algorithm for General Threshold Visual Cryptography Schemes", IEEE transactions, 2011.
15. Younis R. Elhaddad, "Combined Simulated Annealing and Genetic Algorithm to Solve Optimization Problems ", World Academy of Science, Engineering and Technology 68 2012
16. Winkler W.: Using simulated annealing for k-anonymity. Technical Report 7, US Census Bureau.
17. Atallah, M., Elmagarmid, A., Ibrahim, M., Bertino, E., Verykios, V.: Disclosure limitation of sensitive rules, Workshop on Knowledge and Data Engineering Exchange, 1999.

1/8/2013