Maximum Entropy Oriented Anonymization Algorithm for Privacy Preserving Data Mining

Stergios G. Tsiafoulis[1], Vasilios C. Zorkadis[2], and Elias Pimenidis[3]

[1] Hellenic Open University
Ministry of Public Administrative Reform and e-Government
Vasilissis Sofias Av. 15, 10674, Athens, Greece
stetsiafoulis@gmail.com
[2] Hellenic Open University
Hellenic Data Protection Authority Athens
Kifissias Av.1-3, 115 23, Athens, Greece
zorkadis@dpa.gr
[3] University of East London
e.pimenidis@uel.ac.uk

Abstract. This work introduces a new concept that addresses the problem of preserving privacy when anonymising and publishing personal data collections. In particular, a maximum entropy oriented algorithm to protect sensitive data is proposed. As opposed to *k*-anonymity, ℓ-diversity and *t*-closeness, the proposed algorithm builds equivalence classes with possibly uniformly distributed sensitive attribute values, probably by means of noise, and having as a lower limit the entropy of the distribution of the initial data collection, so that background information cannot be exploited to successfully attack the privacy of data subjects data refer to. Furthermore, existing privacy and information loss related metrics are presented, as well as the algorithm implementing the maximum entropy anonymity concept. From a privacy protection perspective, the achieved results are very promising, while the suffered information loss is limited.

Keywords: Privacy preservation, maximum entropy anonymity, *k*-anonymity, ℓ-diversity, *t*-closeness, maximum entropy, (SOMs), neural-network clustering.

1 Introduction

Data contained in databases may be personal, i.e. information referring to an individual directly or indirectly identifiable and therefore its processing should be restricted to lawful purposes. However, exploiting such personal data collections may offer many benefits to the community and support the policy and action plan development process and even contribute to prognosis and treatment of diseases [1]. To address these at first sight contradicting requirements, privacy preserving data mining techniques have been proposed [2-10].

A few years ago, the most common and simplest method to protect from privacy breaches was to remove the identifiers from the database. But an attacker can associate published databases from different sources and extract personal information of

H. Jahankhani et al. (Eds.): ICGS3/e-Democracy 2011, LNICST 99, pp. 9–16, 2012.

an individual. An attack of this kind is called "linking attack". A study held in 2000 linked a Massachusetts voter list with an anonymized database that contained medical records demonstrating that 87% of the population of the United States can be uniquely identified.

Existing privacy-preserving data mining algorithms can be classified into two categories: algorithms that protect the sensitive data itself in the mining process, and those that protect the sensitive data mining results [1]. The most popular concepts in the privacy preserving data mining research literature are k-anonymity, ℓ-diversity and t-closeness. All these concepts belong to the first category and apply generalization and suppression methods to the original datasets in order to preserve the anonymity of individuals or entities data refer to [15].

In this paper, the authors propose a new concept called maximum entropy anonymity concept. It is based on the idea of creating equivalence classes with maximum entropy with respect to the sensitive attribute values.

The paper is structured as follows. Section 2, provides an introduction to the proposed maximum entropy concept, while in section 3, anonymity and information loss metrics are briefly presented. In section 4, the algorithm that implements the proposed concept is presented and in section 5 the experimental studies and results are discussed.

2 Maximum Entropy Oriented Anonymity Concept

The concept of k-anonymity does not take into account the distribution of the sensitive attribute values in each equivalence class, thus leaving space for successful privacy related attacks, while the concept of ℓ-diversity reduces this risk by requiring at least ℓ different sensitive attribute values in each equivalence class. Finally, t-closeness aims at having sensitive attribute values, in each equivalence class, that follow the related distribution of the initial data table being anonymised in order to cope with background knowledge based attacks.

From the privacy protection perspective, the maximum entropy oriented anonymity concept sets a much more ambitious goal, namely that of building equivalence classes with possibly uniformly distributed sensitive attribute values, i.e., showing maximum entropy with regard to sensitive attribute values and thus maximizing the uncertainty of an aspiring attacker exploiting background knowledge. Background knowledge related attacks are radically encountered, regardless of the information an attacker may possess. However, maximum entropy may not be achieved in all equivalence classes, and depending on the initial distribution it may be restricted to only a few. For such a reason, the original goal may be reduced and the new requirement could be set for each equivalence class to have maximum entropy or at least equal entropy to that of the initial data collection. Noise must be constructed and introduced into those equivalence classes for which the defined goal cannot be achieved otherwise, while keeping the information loss to possibly negligible levels.

The algorithmic implementation of the maximum entropy oriented anonymity concept is attained by dividing the initial sensitive attribute distribution into possibly equivalence class uniform distributions, while minimizing the required noise and information loss.

3 Performance Evaluation

The anonymization process has two objectives, that of preserving privacy, in other words to achieve a high degree of anonymity, and, that of minimizing the resulting information loss. Therefore, any performance evaluation criteria should take into account the above two objectives [15].

Information theoretic anonymity metrics have been proposed mainly based on the entropy concept [15, 20]. The entropy *H(X)* refers to an attacker's a priori knowledge regarding for instance possible senders of a message or a number of messages,

$$H(X) = \sum_{i \in X} p(x_i) \log_2 p(x_i) \tag{1}$$

while *H(X/C)* is the conditional information quantity for an attacker after having received the anonymized table (published table), while exploiting available background The higher the entropies, the better the anonymity, i.e. the more uncertain the attacker is about data subject identities [15, 20].

From the information loss perspective, several criteria have been proposed so far in the literature [12, 13, 16, 22]. In most previous work that proposed group based ano-nymization, the relevant evaluation metrics used are: *Discernibility metric* [12, 13, 22], *Classification metric* [12, 16] and *Normalized Certainty penalty* (NCP)[13].

$$H(X / C) = \sum_{i \in X, j \in C} p(x_i, c_j) \log_2 p(x_i / c_j) \tag{2}$$

Discernibility metric assigns a penalty to each tuple based on how many tuples in the transformed dataset are indistinguishable from it. This can be mathematically stated as follows:

$$C_{DM}(g,k) = \sum_{\forall Es.t. |E| \geq k} |E|^2 + \sum_{\forall Es.t. |E| < k} |D||E| \tag{3}$$

where |*D*| the size of the input dataset, E refers to the equivalence classes of tuples in *D* induced by the anonymization *g*.

The Normalized Certainty penalty calculates the information loss introduced by the generalization depth incurred in every attribute value of each tuple, considering also the importance of each attribute by assigning them with a proper weight. If the attribute is a numerical one then the information loss is measured as follows

$$NCP_A(t) = w_i \cdot \frac{z_i - y_i}{|A_i|} \tag{4}$$

where z_i-y_i is the range of the generalization to the tuple *t* on the values of the attribute *Ai* and |*Ai*| is the range of all tuples on this.

4 The Maximum Entropy Anonymization Algorithm

In the beginning of the proposed algorithm, equivalence classes with distinct sensitive attribute values are being created. The algorithm that is presented in [15] with the proper modifications is being used to create those equivalence classes.

```
Input: A database T in a table format
Output: An anonymized table T*
Variables: E←{}, the set of the equivalence classes EQ
QIC←{}, set of equivalence classes with similar quasi identifier set QI
SAP←{}, set of tuples with same SA values

01. Begin
02.         d=number of distinct values of the SA
03.       While d>2 do {
04.             Cluster (T)
05.    For every QIC set created from the clustering procedure do {
06.                         Bucketize the tuples according the SA value to SAP sets
07.                                   For every SAP set do {
08.                                             While |SAP|>=d do {
09.                                             Create equivalence classes (SAP)
10.                                             E=E U EQ
11.                                             Return E
                                                }
                                    }
                }
12.               d=d-1
          }
13.    Entropy_procedure (E)
14. End
```

Fig. 1. The Main Algorithm

```
Input: Table T
Output: QIC← {}, set of tables with tuples with similar QI sets

01. Begin
02.         Insert T to the neural network
03.         QIC= {QIC1, QIC2, ..., QICm}
04.         Return QIC
05. End
```

Fig. 2. The Clustering Procedure

```
Input: SAP
Output: EQ

01. Begin
02.     Randomly selection of tuple t_m from the smallest group
03.     EQ={t_m}
04.         For p←1 until p=d-1 do {
05.             Select a tuple t_p that minimizes the NCP
06.             EQ=EQ U t_p
07.             Remove t_p from T
08.             Remove t_m from T
                }
09.     Generalize the values in the EQ
10.     Return EQ
11. End
```

Fig. 3. The "*Create Equivalence Classes*" Procedure

```
Input: Table E
Output: set of EQ with entropy more than a threshold ε

01. Begin
02.     Until there is no EQ with entropy lower than ε do {
03.         Calculate the entropy to each EQ
04.         Create a table LE with the tuples that belong to
            EQ classes with entropy lower than ε
05.         Delete those tuples from E
06.         Incorporate the tuples from LE to E
            }
07. End
```

Fig. 4. The "*Incorporation*" Procedure

After the initial creation of equivalence classes, the *incorporate procedure* is undertaken. In this procedure, firstly the entropy of the created equivalence classes is calculated. Secondly, the tuples that belong to equivalence classes with entropy lower than the set threshold ε are removed from the temporary anonymized set and are incorporated to the *Low Entropy* table. Thirdly, for each tuple of this table, the equivalence classes with the most common quasi identifier set from the temporary anonymized table are searched. Finally, the tuple is incorporated to the class that does not contain the same value to the sensitive attribute field in order to achieve larger entropy value. The entropies of the created equivalence classes are calculated once more and the same procedure of the incorporating step is being repeated until there are no more classes with entropy lower than the threshold ε. The proposed algorithm is shown in the above Figures 1 – 4.

5 Experimental Data Set Up and Results

The Adult database from machine learning repositories offered by the California University has been used for the implementation of the suggested algorithm of this paper. This database includes 30162 tuples with 14 attributes. Eight out of those (age, work class, education, marital-status, occupation, race, sex, native-country) were chosen for the experimental part of this work. The attributes were represented in numerical form according to their distributions and their domain generalization hierarchy as stated in [16] and [23], respectively and extends by setting the restriction of the valid generalization [12]. For the categorical attributes "work class" and "marital status" the same taxonomy trees as those stated in [12] were used. To the categorical attributes "race" and "sex" a simple two level taxonomy tree was applied.

The mapping to numeric values from the categorical attributes was applied according to the valid generalization notion [12]. Age, education, occupation and native-country were considered as numerical attributes. The generalizations for the attribute age were defined through rounding to median while that to the former ones through total generalization. For the evaluation of the algorithm, total weight certainty penalty *NCP(T)* and the discernibility metric *CDM* that were discussed in section 3, are computed. The experiments were conducted under the Windows XP professional operating system on a PC with a 2.4 GHz AMD Athlon processor and 2 GB RAM.

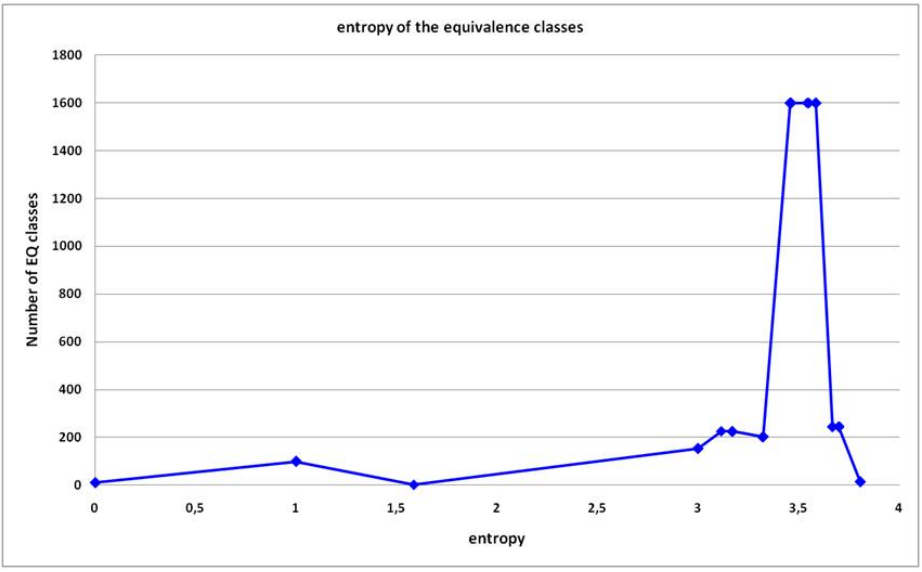

Fig. 5. Entropy of the created equivalence classes considering occupation as sensitive attribute

For the first experiment in this work, the set {*age, education, marital-status, occupation, race, sex, native-country*} was chosen as quasi indentifying set, while *work class* was considered as sensitive attribute. For the second experiment we choose the set {*age, work class, education, marital-status, race, sex, native-country*} as quasi

indentifier set and *occupation* as the sensitive attribute. The distribution of this attribute is much closer to uniform than the previous one. Fig.5 shows the entropy values of the created equivalence. While the entropy of the initial dataset is 3.4, in the resulting intermediate equivalence classes, most of them consisting of 22831 tuples have entropy higher than 3.4. Noise is to be added into the rest of the equivalence classes in order to satisfy the above mentioned entropy threshold.

6 Conclusions

The maximum entropy anonymity concept was introduced and an algorithm that implements it was designed. The performance of our algorithm was measured with respect to privacy by the entropy of the sensitive attribute in each equivalence class and with respect to information loss by means of the *NCP* and discernibility metrics.

We conclude from our work that keeping the distribution of the sensitive attribute values in each equivalence class possibly uniform, or at least the same as the distribution of the initial table, leads to better privacy preservation. We intent to further explore the impact of introducing noise into the equivalence classes, in order to achieve almost perfect privacy preservation, while the resulting information loss is kept to a minimum.

References

1. Wickramasinghe Nilmini, B.R.K., Chris, G.M., Jonathan, S.: Realizing the Knowledge Spiral in Healthcare: the role of Data Mining and Knowledge Management. The International Council on Medical & Care Compunetics, 147–162 (2008)
2. Dalenius, T.: Finding a Needle In a Haystack or Identifying Anonymous Census Records. Journal of Official Statistics 2(3), 329–336 (1986)
3. Sweeney, L.: k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10(5), 557–570 (2002)
4. Sweeney, L., Samarati, P.: Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. In: IEEE Symposium on Research in Security and Privacy (1998)
5. Meyerson, A., Williams, R.: General k-Anonymization is Hard. In: PODS 2004 (2003)
6. Ashwin Machanavajjhala, D.K., Gehrke, J., Venkitasubramaniam, M.: L-Diversity: Privacy Beyond k-Anonymity. ACM Transactions on Knowledge Discovery from Data 1(1), 52, article 3 (2007)
7. Li, N., Li, T., Venkatasubramanian, S.: t-Closeness: Privacy Beyond k-Anonymity and ℓ-Diversity. In: 23rd International Conference on Data Engineering, ICDE 2007, pp. 106–115 (2007)
8. Ye, Y., Deng, Q., Wang, C., Lv, D., Liu, Y., Feng, J.-H.: *BSGI*: An Effective Algorithm towards Stronger *l*-Diversity. In: Bhowmick, S.S., Küng, J., Wagner, R. (eds.) DEXA 2008. LNCS, vol. 5181, pp. 19–32. Springer, Heidelberg (2008)
9. Xiao, X., Tao, Y.: Anatomy: Simple and effective privacy preservation. In: 32nd International Conference on Very large Data Bases, VLDB 2006, pp. 139–150 (2006)

10. LeFevre, K.R., Dewitt, D.J., Ramakrishnan, R.: Incognito: efficient full-domain K-anonymity. In: International Conference on Management of Data ACM SIGMOD 2005, Baltimore, Maryland (2005)
11. LeFevre, K., Dewitt, D.J., Ramakrishnan, R.: Mondrian Multidimensional K-Anonymity. In: ICDE 2006 (2006)
12. Iyengar, V.S.: Transforming Data to Satisfy Privacy Constrains. In: KDD 2002 (2002)
13. Xu, J., Wang, W., Pei, J., Wang, X., Shi, B., Fu, A.W.-C.: Utility-Based Anonymization Using Local Recoding. In: KDD 2006(2006)
14. UCI. Irvin Machine Learning Repository, http://archive.ics.uci.edu/ml/
15. Tsiafoulis, S.G., Zorkadis, V.C.: A Neural Network Clustering Based Algorithm for Privacy Preserving Data Mining. In: 2010 International Conference on Computational Intelligence and Security, Nanning, Guangxi Zhuang Autonomous Region, China (2010)
16. Bayardo, R.J., Agrawal, R.: Data privacy through optimal k-anonymization. In: 21th ICDE 2005 (2005)
17. Webb, G.I.: Opus: An Effcient Admissible Algorithm for Unordered Search. Journal of Artificial intelligence Research 3, 431–465 (1995)
18. Rymon, R.: Search Through Systematic Set Enumeration (1992)
19. Whitley, D.: The Genitor Algorithm and Selective Pressure: Why rank-based allocation of reproductive trials is best. In: Proceedings of Third International Conference on Genetic Algorithms, 1989, pp. 116–121.
20. Kelly, D.J., Raines, R.A., Grimaila, M.R., Baldwin, R.O., Mullins, B.E.: A Survey of State-of-the Art ion Anonymity Metrics. In: NDA 2008. ACM, Fairfax (2008)
21. Dakshi Agrawal, C.C.A.: On the Design and Quantification of Privacy Preserving Data Mining Algorithms. In: 20th Symposium on Principles of Database Systems Santa Barbara California, USA (May 2001)
22. Evfimievski, A.V., Srikant, R., Gehrke, J.: Limiting privacy breaches in privacy preserving data mining. In: Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems table of Contents, San Diego, California, pp. 211–222 (2003)
23. Sweeney, L.: Achieving k-anonymity privacy protection using generalization and suppression. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10(5), 571–588 (2002)