Information Driven Evaluation of Data Hiding Algorithms

Elisa Bertino[1] and Igor Nai Fovino[2]

[1] CERIAS and CS Department, Purdue University
bertino@cerias.purdue.edu
[2] Institute for the Protection and the Security of the Citizen,
Joint Research Centre
igor.nai@jrc.it

Abstract. Privacy is one of the most important properties an information system must satisfy. A relatively new trend shows that classical access control techniques are not sufficient to guarantee privacy when datamining techniques are used. Privacy Preserving Data Mining (PPDM) algorithms have been recently introduced with the aim of modifying the database in such a way to prevent the discovery of sensible information. Due to the large amount of possible techniques that can be used to achieve this goal, it is necessary to provide some standard evaluation metrics to determine the best algorithms for a specific application or context. Currently, however, there is no common set of parameters that can be used for this purpose. This paper explores the problem of PPDM algorithm evaluation, starting from the key goal of preserving of data quality. To achieve such goal, we propose a formal definition of data quality specifically tailored for use in the context of PPDM algorithms, a set of evaluation parameters and an evaluation algorithm. The resulting evaluation core process is then presented as a part of a more general three step evaluation framework, taking also into account other aspects of the algorithm evaluation such as efficiency, scalability and level of privacy.

1 Introduction

Intense work in the area of data mining technology and in its applications to several domains has resulted in the development of a large variety of techniques and tools able to automatically and intelligently transform large amounts of data in knowledge relevant to users. The use of these tools is today the only effective way to extract useful knowledge from the increasing number of very large databases, that, because of their sizes, cannot be manually analyzed. However, as with other kinds of useful technologies, the knowledge discovery process can be misused. It can be used for example by malicious subjects to reconstruct sensitive information for which they do not have access authorization. Because of the nature of datamining techniques that use non sensitive data in order to infer hidden information, the usual security techniques are actually not able to prevent illegal accesses carried out through the use of data mining techniques. For this reason, many research efforts have been recently devoted to addressing

A Min Tjoa and J. Trujillo (Eds.): DaWaK 2005, LNCS 3589, pp. 418–427, 2005.

the problem of privacy preserving in data mining. As a result, different sanitization techniques have been proposed for hiding sensitive items or patterns by removing some of the data to be released or inserting noise into data. Because, however, the sets of parameters adopted for assessing the various algorithms are very heterogeneous, it is difficult to compare these algorithms in order to determine which is the most suitable one for a given context or application. Moreover all parameters adopted by the various metrics do not take into account an important issue of the privacy-preserving data mining (PPDM) process, that is, the quality of the data obtained as result from the sanitization process. In this paper, we explore the DQ properties that are relevant for the evaluation of PPDM algorithms. We propose a model to describe aggregated information, their constraints and their relevance in order to evaluate the various DQ parameters and we provide a three-step framework to evaluate PPDM algorithms using DQ as final discriminant.

2 Related Work

The first approach dealing with the problem of DQ for perturbed data has been developed by Agrawal and Srikant [6]. More in detail, they propose an approach to estimate the privacy level introduced by their PPDM algorithm. Such an approach evaluates the accuracy with which the original values of a modified attribute can be determined. A similar approach has been developed by Rivzi and Haritsa [8]. They propose a distortion method to pre-process the data before executing the mining process. The privacy measure they propose deals with the probability with which the distorted entries can be reconstructed.

These initial approaches have then been extended by taking into account other evaluation parameters and by considering specific data mining techniques. In particular, Agrawal and Aggarwal [5] analyze the proposed PPDM algorithm with respect to two parameters, that is, privacy and information loss. In the context of clustering techniques, Oliveira and Zaiane in [3][2] define some interesting parameters to evaluate the proposed PPDM algorithms, each concerning different aspects of these algorithms. In particular, in the context of Data quality, interesting parameters they introduced are the *misclassification error*, used to estimate how many legitimate data points are incorrectly classified in the distorted database and the *Artifactual Pattern* estimating the artifactual patterns introduced by the sanitization process that are not present in the original database. A more comprehensive evaluation framework for PPDM algorithms has been recently developed by Bertino et al. [16]. Such an evaluation framework consists of five parameters: the *Efficiency*, the *Scalability*, the *Data quality*, the *Hiding failure* and the *Level of privacy* As it can noticed from the above overview, only the approaches by Bertino et al. [16] and Olivera and Zaiane [2,3] have addressed the problem of DQ in the context of the PPDM process. However, these approaches have some major differences with respect to the work presented in this paper. As we discussed before, in the evaluation framework developed by Bertino et al., DQ is directly measured by the dissimilarity parameter. By contrast, Olivera and Zaiane only measure DQ through the estimation of the artifactual patterns and the misclassification error. However, none of these approaches completely captures the concept of DQ. In particular, there are some aspects related to DQ evaluation that are heavily related not only with the PPDM algorithm, but also with the structure of the database, and with the meaning and relevance of the information stored in the database with respect to a well defined context. Parameters such as dissimilarity, artifactual information and misclassification error are not able to capture these aspects. The goal of this paper is to explore the role and relevance of DQ in the PPDM process by developing a more appropriate set of instruments to assess the quality of the data obtained by the sanitization process.

3 Data Quality in the Context of PPDM

Traditionally DQ is a measure of the consistency between the data views presented by an information system and the same data in the real-world [10]. This definition is strongly related with the classical definition of information system as a “model of a finite subset of the real world” [12]. More in detail Levitin and Redman [13] claim that DQ is the instrument by which it is possible to evaluate if data models are well defined and data values accurate. The main problem with DQ is that its evaluation is relative [9], in that it usually depends from the context in which data are used. In the scientific literature DQ is considered a multi-dimensional concept that in some environments involves both objective and subjective parameters [11,14]. In the context of PPDM, we are interested in assessing whether, given a target database, the sanitization phase will compromise the quality of the mining results that can be obtained from the sanitized database. The parameters we consider relevant in the context of PPPDM are the following: the *Accuracy*, measuring the proximity of a sanitized value a^I to the original value a; the *Completeness*, evaluating the percentage of data from the original database that are missing from the sanitized database and finally the *Consistency* that is related to the semantic constraints holding on the data and it measures how many of these constraints are still satisfied after the sanitization. We now present the formal definitions of those parameters for use in the remainder of the discussion. Let OD be the original database and SD be the sanitized database resulting from the application of the PPDM algorithm. Without loosing generality and in order to make simpler the following definitions, we assume that OD (and consequently SD) be composed by a single relation. We also adopt the positional notation to denote attributes in relations. Thus, let od_i (sd_i) be the i-th tuple in OD (SD), then od_{ik} (sd_{ik}) denotes the k^{th} attribute of od_i (sd_i). Moreover, let n be the total number of the attributes of interest, we assume that attributes of positions $1, \ldots, m$ ($m \leq n$) are the primary key attributes of the relation.

Definition 1: Let sd_j be a tuple of SD. We say that sd_j is **Accurate** if $\neg\exists od_i \in OD$ such that $((od_{ik} = sd_{jk})\forall k = 1..m \wedge \exists(od_{if} \neq sd_{jf}), (sd_{jf} \neq NULL), f = m+1, .., n)$. □

Definition 2: A sd_j is **Complete** if $(\exists od_i \in OD$ such that $(od_{ik} = sd_{jk})\forall k = 1..m) \wedge (\neg\exists(sd_{jf} = NULL), f = m+1,..,n)$. □

Let C the set of the constraints defined on database OD, in what follows we denote with c_{ij} the j^{th} constraint on attribute i. We assume here constraints on a single attribute, but, as we show in Section 4 it is easily possible to extend the measure to complex constraints.

Definition 3: An instance sd_k is **Consistent** if $\neg\exists c_{ij} \in C$ such that $c_{ij}(sd_{ki}) = false, i = 1..n$ □

4 Information Driven Data Quality Schema

Current approaches to PPDM algorithms do not take into account two important aspects:

- **Relevance of data:** not all the information stored in the database has the same level of relevance and not all the information can be dealt at the same way.
- **Structure of the database:** information stored in a database is strongly influenced by the relationships between the different data items. These relationships are not always explicit.

We believe that in a context in which a database administrator needs to choose which is the most suitable PPDM algorithm for a target real database, it is necessary to also take into account the above aspects. To achieve this goal we propose to use Data Quality in order to assess how and if these aspects are preserved after a data hiding sanitization.

4.1 The Information Quality Model

In order evaluate DQ it is necessary to provide a formal description that allow us to magnify the aggregate information of interest for a target database and the relevance of DQ properties for each aggregate information (AI) and for each attribute involved in the AI. The Information Quality Model (IQM) proposed here addresses this requirement. In the following, we give a formal definition for an Attribute Class (AC), a Data Model Graph (DMG) (used to represent the attributes involved in an aggregate information and their constraints) and an Aggregation Information Schema (AIS).Before giving the definition of DMG, AIS and ASSET we introduce some preliminary concepts.

Definition 4: An Attribute Class is defined as the tuple $AT_C =< name, AW, AV, CW, CV, CSV, Slink >$ where:

- *Name* is the attribute id
- *AW* is the accuracy weigh for the target attribute
- *AV* is the accuracy value

- CW is the completeness weigh for the target attribute
- CV is the completeness value
- CSV is the consistency value
- *Slink* is list of simple constraints. □

Definition 5: A Simple Constraint Class is defined as the tuple $SC_C =< name, Constr, CW, Clink, CSV >$ where:

- *Name* is the constraint id
- *Constraint* describes the constraint using some logic expression
- CW is the weigh of the constraint. It represents the relevance of this constraint in the AIS
- CSV is the number of violations to the constraint
- *Clink* it is the list of complex constraints defined on SC_C. □

Definition 6: A Complex Constraint Class is defined as the tuple $CC_C =< name, Operator, CW, CSV, SC_C\_link >$ where:

- *Name* is the Complex Constraint id
- *Operator* is the "Merging" operator by which the simple constraints are used to build the complex one.
- CW is the weigh of the complex constraint
- CSV is the number of violations
- SC_Clink is the list of all the SC_C that are related to the CC_C. □

Let D a database, we are able now to define the DMG, AIS and ASSET on D.

Definition 7: A **DMG** (Data Model Graph) is an oriented graph with the following features:

- A set of nodes N_A where each node is an Attribute Class
- A set of nodes SC_C where each node describes a Simple Constraint Class
- A set of nodes CC_C where each node describes a Complex Constraint Class
- A set of direct edges $L_{Nj,Nk} : L_{Nj,Nk} \in ((N_A X SC_C) \cup (SC_C X CC_C) \cup (SC_C X N_A) \cup (CC_C X N_A))$. □

Definition 8: An **AIS** ϕ is defined as a tuple $< \gamma, \xi, \lambda, \vartheta, \varpi, W_A IS >$ where: γ is a name, ξ is a DMG, λ is the accuracy of AIS, ϑ is the completeness of AIS, ϖ is the consistency of AIS and W_{AIS} represent the relevance of AIS in the database. □

We are now able to identify as **ASSET** (Aggregate information Schema Set) as the collection of all the relevant AIS of the database.

The DMG completely describes the relations between the different data items of a given AIS and the relevance of each of these data respect to the data quality parameter. It is the "road map" that is used to evaluate the quality of a sanitized AIS.

4.2 Data Quality Evaluation of AIS

By adopting the IQM scheme, now we are able to evaluate the data quality at the attribute level. By recalling *Definition (1,2)*, we define the *Accuracy lack* of an attribute k for an AIS A as the proportion of non accurate items in a database SD. Ate the same way, the *Completeness lack* of an attribute k is defined as the proportion of non complete items in SD. The accuracy lack index for an AIS can the be evaluated as follows:

$$ACL = \sum_{i=0}^{i=n} DMG.N_i.AV * DMG.N_i.AW \tag{1}$$

where $DMG.N_i.AW$ is the accuracy weight associated with the attribute identified by the node N_i. Similarly the completeness lack of an AIS can be measured as follows:

$$CML = \sum_{i=0}^{i=n} DMG.N_i.CV * DMG.N_i.CW \tag{2}$$

Finally the consistency lack index associated with an AIS is given by number of constraint violations occurred in all the sanitized transaction multiplied by the weight associated with every constraints (simple or complex).

$$CSL = \sum_{i=0}^{i=n} DMG.SC_i.csv * DMG.SC_i.cw + \sum_{j=0}^{j=m} DMG.CC_j.csv * DMG.CC_j.cw \tag{3}$$

4.3 The Evaluation Algorithm

In this section we present the methodology we have developed to evaluate the data quality of the AIS. This methodology is organized in two main phases:

- **Search**: in this phase all the tuples modified in the sanitized database are identified. The primary keys of all these transaction (we assume that the sanitization process does not change the primary key), are stored in a set named *evalset*. This set is the input of the Evaluation phase.
- **Evaluation**: in this phase the accuracy, the consistency and the completeness associated with the DMG and the AIS are evaluated using information on the accuracy and completeness weight associated with the DMG and related to the transactions in Evalset.

The algorithms for these phases are reported in Figures 1 and 2. Once the evaluation process is completed, a set of values is associated with each AIS that gives the balanced level of accuracy, completeness and consistency. However, this set may not be enough. A simple average of the different AIS's values could not be significant, because even in this case not all the AIS's in the ASSET have the same relevance. For this reason, a weight is associated with each AIS that represents the importance of the high level information represented by the AIS in the target context. The accuracy, the completeness and the consistency of the ASSET for each PPDM algorithm candidate are then evaluated as follows:

$$Accuracy_{\text{Asset}} = \frac{\sum_{i=0}^{i=|Asset|} AIS_i.accuracy * AIS_i.W}{|Asset|} \tag{4}$$

```
INPUT:   Original database OD, Sanitized database SD
OUTPUT:  a set Evalset of primary keys
Begin
  Foreach $t_i \in OD$ do
    {j=0;
      While $(s_{jk} \neq t_{ik})$and$(j < |SD|)$do $j++$;
      l=0;
      While $(s_{jl} = til)$and$(l < n)$ do l++;
      If$(l < n)$Then $Evalset = Evalset \cup t_{ik}$
    }
End
```

Fig. 1. Search algorithm

```
INPUT:   the original database OD, the sanitized database SD,Evalset, IQM.
OUTPUT:  the IQM containing a data quality evaluation
Begin
  Foreach IES in IQM do
  {$DMG = IQM.IES.link$;
    $avet = 0$;$cvet = 0$;
    Foreach $(t_{ik} \in Evalset)$do
      For$(j = 0; j < n; j++)$ do
        If $(t_{ij} \neq s_{ij})$ Then
          {
            If $s_{ij} = NULL$ Then $cvet[j]++$;
            Else $avet[j]++$;
            validate_constr(IES,DMG,j)
          }
    For$(m = 0; m < n; m++)$do
      { $DMG.N_m.AV = \frac{avet[m]}{|SD|}$; $DMG.N_m.CV = \frac{cvet[m]}{|SD|}$;}
    $IQM.IES.AV = \sum_{i=0}^{i=n}(DMG.N_i.AV * DMG.N_i.AW)$;
    $IQM.IES.CV = \sum_{i=0}^{i=n}(DMG.N_i.CV * DMG.N_i.CW)$;
    $IQM.IES.CSV = \sum_{i=0}^{i=n} DMG.SC_i.CSV * DMG.SC_i.CW +$
                   $\sum_{j=0}^{j=m} DMG.CC_j.CSV * DMG.CC_j.CW$;}
End
Procedure validate_constr(IES,DMg,j) Begin
  $N_A$=AIS.DMG.j
  For k=1; $k < |N_A.slink|$; k++
  { $N_C = N_A.Slink[k]$
    if$N_C.Clink == NULL$ then{if !($N_C$.constr($s_{ij}$)) then $N_C.CSV++$;}
    else{ $N_O$=$N_C.Clink$; globalconstr=composeconstr($N_O$, $s_{ij}$)
      if !(globalconstr) then $N_O.CSV++$}}
End
```

Fig. 2. Evaluation Algorithm

$$Completeness_{\text{Asset}} = \frac{\sum_{i=0}^{i=|Asset|} AIS_i.completeness * AIS_i.W}{|Asset|} \tag{5}$$

$$Consistency_{\text{Asset}} = \frac{\sum_{i=0}^{i=|Asset|} AIS_i.consistency * AIS_i.W}{|Asset|} \tag{6}$$

where $AIS_i.W$ represents the weight (relevance) associated with the i-th AIS.

5 Evaluation Framework

As shown by the approaches reported in Section 2, and especially by the approach by Bertino et al. [16], in many real world applications, it is necessary to take into account even other parameters that are not directly related to DQ. On the other hand we believe that DQ should represent the invariant of a PPDM evaluation and should be used to identify the best algorithm within a set of previously selected "Best Algorithms". To preselect this "best set", we suggest to use some parameters as discriminant to select the algorithms that have an acceptable behavior under some aspects generally considered relevant especially in "production environments" (efficiency, scalability, hiding failure and level of privacy). In order to understand if these four parameters are sufficient to identify an acceptable set of candidates, we performed an evaluation test. We identified a starting set of PPDM algorithms for Association Rules Hiding (the algorithms presented in [4] and a new set of three algorithms based on data fuzzification [16]). Then, by using the IBM synthetic data generator[3]we generated a categoric database representing an hypothetical Health Database storing the different therapies associated with the patients. We also built the associated DMG. On this database, we applied the different algorithms and then we measured the previous parameters. Once we built the "Best Set" we discovered that some algorithms that performed less changes to the database, which in some way indicates a better quality, are not in this set. A reason is for example a low efficiency. For this reason we believe that even in the preselection phase a "coarse" DQ parameter must be introduced. In our opinion, the *Coarse DQ Measure* depends on the specific class of PPDM algorithms. If the algorithms adopt a perturbation or a blocking technique, the coarse DQ can be measured by the dissimilarity between the original dataset D and the sanitized one D' by measuring, for example, in the case of transactional datasets, the difference between the item frequencies of the two datasets before and after the sanitization. Such dissimilarity can be estimated by the following expression:

$$Diss(D, D') = \frac{\sum_{i=1}^{n} |f_D(i) - f_{D'}(i)|}{\sum_{i=1}^{n} f_D(i)} \tag{7}$$

where i is a data item in the original database D, and $f_D(i)$ is its frequency within the database, whereas i' is the given data item after the application of a privacy preservation technique and $f_{D'}(i)$ is its new frequency within the transformed database D'. The same method can be used, extending the previous formula, also in the case of blocking techniques. If the data modification consists of aggregating some data values, the coarse DQ is given by the loss of detail in the data. As in the case of the $k-$Anonymity algorithm [15], given a database DB with N_A attributes and N transactions, if we identify as generalization scheme a domain generalization hierarchy GT with a depth h, it is possible to measure the coarse quality of a sanitized database SDB as:

$$Quality(SDB) = 1 - \frac{\sum_{i=1}^{i=N_A} \sum_{j=1}^{i=N} \frac{h}{|GT_{Ai}|}}{|DB| * |N_A|} \tag{8}$$

where $\frac{h}{|GT_{Ai}|}$ represent the detail loss for each cell sanitized.

Once we have identified the *Best Set* we are able to apply our DQ-driven evaluation.

[3] http://www.almaden.ibm.com/software/quest/Resources/datasets/syndata.html

We now present a three steps Evaluation Framework based on the previous concepts.

1. A set of "Interesting" PPDM's is selected. These algorithms are tested on a generic database and evaluated according the general parameters (Efficiency, Scalability, Hiding failure, Coarse Data Quality, Level of privacy). The result of this step is a restricted set of *Candidate algorithms*
2. A test database with the same characteristics of the target database is generated. An IQM schema with the AIS and the related DMG is the result of this step.
3. The Information Driven DQ Evaluation Algorithm is applied in order to identify the algorithm that finally will be applied.

As it is probably obvious to the readers, the most "time consuming" step in terms of required user interactions is step 2. The design a good IQM is the core of our evaluation framework. We believe that a top down approach is, in this cases, the most appropriate. More in detail, the first task should be the identification of the high level information that is relevant and for which we are interested in measuring the impact of PPDM algorithms. It could also can be useful to involve in this task some authorized users (e.g. in case of Health DBA's, doctors, etc.) in order to understand all the possible uses of the database and the relevance of the retrieved information. The use of datamining tools could be useful to identify non-evident aggregate information.

A second task would then, given the high level information, determine the different constraints (both simple and complex) and evaluate their relevance. Also in this case, discussions with authorized users and DB designers, and the use of DM tools (e.g. discover association rules) could help to build a good IQM. Finally it is necessary, by taking into account all the previous information, to rate the relevance of the attributes involved. This top down analysis is useful not only for the specific case of PPDM evaluation, but, if well developed, is a powerful tool to understand the real information contents, its value and the relation between the information stored in a given database. In the context of an "Information Society" this is a non negligible added value.

6 Conclusion

The Data Quality of Privacy Preserving Data Mining Algorithms is an open problem. In this paper we have proposed an approach to represent three important aspects of the data quality, an algorithm to magnify the impact of a PPDM algorithm on the data quality of a given database and a framework to select the most suitable algorithm for such database. We have also carried out extensive tests for the case of association rules PPDM algorithms. We plan to extend our work with some tests on other types of PPDM algorithms. Another direction that we plan to explore is the use and then the test of our framework over non-homogeneous algorithms. Because our principal focus is to provide tools supporting the database administrator in identifying the most suitable algorithm for their database, we also plan to develop a tool that will be integrated with GADAET [16] in order to allow an intuitive Asset, AIS and DMG design and an automated algorithms test.

References

1. V. S. Verykios, E. Bertino, I. Nai Fovino, L. Parasiliti, Y. Saygin, Y. Theodoridis. *State-of-the-art in Privacy Preserving Data Mining.* SIGMOD Record, 33(1):50-57 (2004).
2. S. R. M. Oliveira and O. R. Zaiane. *Privacy Preserving Frequent Itemset Mining.* In Proceedings of the 28th International Conference on Very Large Databases (2003).
3. S. R. M. Oliveira and O. R. Zaiane. *Privacy Preserving Clustering by Data Transformation.* In Proceedings of the 18th Brazilian Symposium on Databases, Manaus, Amazonas, Brazil, October 2003, pp.304-318.
4. University of Milan - Computer Technology Institute - Sabanci University. *CODMINE.* IST project. 2002-2003.
5. D. Agrawal and C. C. Aggarwal. *On the Design and Quantification of Privacy Preserving Data Mining Algorithms.* In Proceedings of the 20th ACM Symposium on Principle of Database System (2001), pp. 247-255.
6. R. Agrawal and R. Srikant. *Privacy Preserving Data Mining.* In Proceedings of the ACM SIGMOD Conference of Management of Data (2000), pp. 439-450.
7. A. Evfimievski. *Randomization in Privacy Preserving Data Mining.* SIGKDD Explor. Newsl. (2002), Vol.4, n. 2, pp. 43-48, ACM
8. S. J. Rizvi and J. R. Haritsa. *Maintaining Data Privacy in Association Rule Mining.* In Proceedings of the 28th International Conference on Very Large Databases (2003).
9. G. Kumar Tayi, D. P. Ballou. *Examining Data Quality.* Comm. of the ACM, Vol. 41, Issue 2, pp. 54-58, Feb. 1998.
10. K. Orr. *Data Quality and System Theory.* Comm. of the ACM, Vol. 41, Issue 2, pp. 66-71, Feb. 1998.
11. Y. Wand and R. Y. Wang. *Anchoring Data Quality Dimensions in Ontological Foundations.* Comm. of the ACM, Vol. 39, Issue 11, pp. 86-95 Nov. 1996.
12. W. Kent. *Data and reality.* North Holland, New York, 1978.
13. A.V. Levitin and T.C. Redman. *Data as resource: properties, implications and prescriptions.* Sloan Management review, Cambridge, Fall 1998, Vol. 40, Issue 1, pp. 89-101.
14. R. Y. Wang, D. M. Strong. *Beyond Accuracy: what Data Quality Means to Data Consumers.* Journal of Manahement Information Systems Vol. 12, Issue 4, pp. 5-34, (1996).
15. L. Sweeney. *Achieving* k*-Anonymity Privacy Protection using Generalization and Suppression.* Journal on Uncertainty, Fuzzyness and Knowledge-based Sys., Vol. 10, Issue 5, pp. 571-588,(2002).
16. E. Bertino, I. Nai Fovino, L. Parasiliti Provenza. *A Framework for Evaluating Privacy Preserving Data Mining Algorithms.* Data Mining and Knowledge Discovery Journal, 2005.