International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Bottom-Up Generalization: A Data Mining Solution to Privacy Protection

Vishakha B. Dalvi[1], Ranjit R. Keole[2]

[1]Department of Computer Science & Information Technology, ME First Year, HVPM's COET, SGBA University, Amravati, India

[2]Department of Computer Science & Information Technology, HVPM's COET, SGBA University, Amravati, India

Abstract: ***In recent years, privacy-preserving data mining has been studied extensively, because of the wide proliferation of sensitive information on the internet. This paper investigates data mining as a technique for masking data; therefore, termed data mining based privacy protection. This approach incorporates partially the requirement of a targeted data mining task into the process of masking data so that essential structure is preserved in the masked data. The following privacy problem is considered in this paper: a data holder wants to release a version of data for building classification models, but wants to protect against linking the released data to an external source for inferring sensitive information. An iterative bottom-up generalization is adapted from data mining to generalize the data. The generalized data remains useful to classification but becomes difficult to link to other sources. The generalization space is specified by a hierarchical structure of generalizations. A key is identifying the best generalization to climb up the hierarchy at each iteration.***

Keywords: Generalization, k-anonymity, privacy-preserving data mining, randomization, re-identification.

1. Introduction

Information becomes sensitive when they are specific to a small number of individuals. Data mining, on the other hand, typically makes use of information shared by some minimum number of individuals to ensure a required statistical significance of patterns. As such, sensitive information is to be discarded for reliable data mining. This observation motivates to apply the requirement of an intended data mining task to identify useful information to be released, therefore, sensitive information to be masked. This approach is called *data mining based privacy protection*. A well-studied technique for masking sensitive information, primarily studied in statistics, is *randomizing* sensitive attributes by adding random error to values. In these works, privacy was quantified by how closely the original values of a randomized attribute can be estimated. This is very different from the *K*-anonymity that quantifies how likely an individual can be linked to an external source. The *privacy-preserving data mining* in [1] extends traditional data mining techniques to handle randomized data. Data mining itself is investigated as a technique for masking data. The masked data does not require modification of data mining techniques in subsequent data analysis. Instead of randomizing data, *generalizing* data makes information less precise. Grouping continuous values and suppressing values are examples of this approach. Compared to randomization, generalization has several advantages. First, it preserves the "truthfulness" of information, making the released data meaningful at the record level. This feature is desirable in exploratory and visual data mining where decisions often are made based on examining records. In contrast, randomized data are useful only at the aggregated level such as average and frequency. Second, preferences can be incorporated through the taxonomical hierarchies and the data recipient can be told what was done to the data so that the result can be properly interpreted.

The increasing ability to accumulate, store, retrieve, cross-reference, mine and link vast number of electronic records brings substantial benefits to millions of people. An example given in [4] is that a *sensitive* medical record was uniquely linked to a *named* voter record in a publicly available voter list through the shared attributes of Zip, Birth date, Sex. Indeed, since "the whole is greater than the sum of the parts", protection of individual sources does not guarantee protection when sources are cross-examined. Consider the following *anonymity problem* [5]. A data holder wants to release a person-specific data *R*, but wants to prevent from linking the released data to an external source *E* through shared attributes $R \cap E$, called the *virtual identifier*. One approach is to generalize specific values into less specific but semantically consistent values to create *K-anonymity*: if one record r in *R* is linked to some external information, at least $K - 1$ other records are similarly linked by having the same virtual identifier value as r. The idea is to make the inference ambiguous by creating extraneous linkages. An example is generalizing "birth date" to "birth year" so that everybody born in the same year are linked to a medical record with that birth year, but most of these linkages are non-existing in the real life.

2. k- Anonymity: A Model for Protecting Privacy

The attributes are generalized until each row is identical with at least k-1 other rows. At this point the database is said to be k- anonymous. k-anonymity [7],[8],[10] is a property that captures the protection of released data against possible re-identification of the respondents to whom the released data refer. Consider a private table PT, where data have been de-identified by removing explicit identifiers (e.g., SSN and Name). However, values of other released attributes, such as ZIP, Date of birth, Marital status, and Sex can also appear in some external tables jointly with the individual respondents' identities. If some combinations of values for these attributes are such that their occurrence is unique or rare, then parties

observing the data can determine the identity of the respondent to which the data refer or reduce the uncertainty over a limited set of respondents.

k-anonymity demands that every tuple in the private table being released be indistinguishably related to no fewer than k respondents. Since it seems impossible, or highly impractical and limiting, to make assumptions on which data are known to a potential attacker and can be used to (re-)identify respondents, k-anonymity takes a safe approach requiring that, in the released table itself, the respondents be indistinguishable (within a given set of individuals) with respect to the set of attributes, called quasi-identifier, that can be exploited for linking. In other words, k-anonymity requires that if a combination of values of quasi-identifying attributes appears in the table, then it appears with at least k occurrences.

To illustrate, consider a private table reporting, among other attributes, the marital status, the sex, the working hours of individuals, and whether they suffer from hypertension. Assume attributes Marital status, Sex, and Hours are the attributes jointly constituting the quasi-identifier. Figure 1 is a simplified representation of the projection of the private table over the quasi-identifier. The representation has been simplified by collapsing tuples with the same quasi-identifying values into a single tuple.

Marital_status	Sex	Hours	# tuple (Hyp. Values)
divorced	M	35	2 (0Y, 2N)
divorced	M	40	17 (16Y, 1N)
divorced	F	35	2 (0Y, 2N)
married	M	35	10 (8Y, 2N)
married	F	40	9 (2Y, 7N)
single	M	50	26 (6Y, 20N)

Figure 1: Simplified representation of a private table

The numbers at the right hand side of the table report, for each tuple, the number of actual occurrences, also specifying how many of these occurrences have values Y and N, respectively, for attribute Hypertension. For simplicity, in the following we use such a simplified table as our table PT. The private table PT in Figure 1 guarantees k-anonymity only for $k \leq 2$. In fact, the table has only two occurrences of divorced (fe)males working 35 hours. If such a situation is satisfied in a particular correlated external table as well, the uncertainty of the identity of such respondents can be reduced to two specific individuals. In other words, a data recipient can infer that any information appearing in the table for such divorced (fe)males working 35 hours, actually pertains to one of two specific individuals.

3. Bottom – Up Generalization

Wang et al. [1] present an effective bottom-up generalization approach to achieve k-anonymity. They employed the sub-tree generalization scheme. A generalization g : child(v) → v, replaces all instances of every child value c in child(v) with the parent value v. Although this method is designed for achieving k-anonymity, it can be easily modified to adopt the LKC-privacy model in order to accommodate the high-dimensional data.

3.1 The Anonymization Algorithm

Algorithm 3.1.1 presents the general idea of bottom-up generalization method. It begins the generalization from the raw data table T. At each iteration, the algorithm greedily selects the Best generalization g that minimizes the information loss and maximizes the privacy gain. This intuition is captured by the information metric ILPG(g) = IL(g)/PG(g). Then, the algorithm performs the generalization child(Best) → Best on the table T , and repeats the iteration until the table T satisfies the given k-anonymity requirement.

Algorithm 3.1.1 Bottom-Up Generalization
1: while T does not satisfy a given k-anonymity requirement do
2: for all generalization g do
3: compute ILPG(g);
4: end for
5: find the Best generalization;
6: generalize T by Best;
7: end while
8: output T;

Let A(QID) and Ag(QID) be the minimum anonymity counts in T before and after the generalization g. Given a data table T, there are many possible generalizations that can be performed. Yet, most generalizations g in fact does not affect the minimum anonymity count. In other words, A(QID) = Ag(QID). Thus, to facilitate efficiently choosing a generalization g, there is no need to consider all generalizations. Indeed, we can focus only on the "critical generalizations."

DEFINITION 3.1: A generalization g is critical if Ag(QID) > A(QID).

Wang et al. [1] made several observations to optimize the efficiency of Algorithm 3.1.1: A critical generalization g has a positive PG(g) and a finite ILPG(g), whereas a non-critical generalization g has PG(g) = 0 and infinite ILPG(g). Therefore, if at least one generalization is critical, all non-critical generalizations will be ignored by the ILPG(g) information metric. If all generalizations are non-critical, the ILPG(g) metric will select the one with minimum IL(g). In both cases, Ag(QID) is not needed for a non-critical generalization g. Based on this observation, Lines 2-3 in Algorithm 3.1.1 can be optimized as illustrated in Algorithm 3.1.2.

Algorithm 3.1.2 Bottom-Up Generalization
1: while T does not satisfy a given k-anonymity requirement do
2: for all critical generalization g do
3: compute Ag(QID);
4: end for
5: find the Best generalization;
6: generalize T by Best;
7: end while
8: output T;

3.2 Data Structure

To further improve the efficiency of the generalization operation, Wang et al. [1] propose a data structure, called Taxonomy Encoded Anonymity (TEA) index for QID = D1, . . . , Dm. TEA is a tree of m levels. The ith level represents the current value for Dj . Each root-to-leaf path represents a qid value in the current data table, with a(qid) stored at the leaf node. In addition, the TEA index links up the qids according to the generalizations that generalize them. When a generalization g is applied, the TEA index is updated by adjusting the qids linked to the generalization of g. The purpose of this index is to prune the number of candidate generalizations to no more than |QID| at each iteration, where |QID| is the number of attributes in QID. For a generalization g : child(v) → v, a segment of g is a maximal set of sibling nodes, {s1, . . . , st}, such that {s1, . . . , st} & child(v), where t is the size of the segment. All segments of g are linked up. A qid is generalized by a segment if the qid contains a value in the segment.

A segment of g represents a set of sibling nodes in the TEA index that will be merged by applying g. To apply generalization g, we follow the link of the segments of g and merge the nodes in each segment of g. The merging of sibling nodes implies inserting the new node into a proper segment and recursively merging the child nodes having the same value if their parents are merged. The merging of leaf nodes requires adding up a(qid) stored at such leaf nodes. The cost is proportional to the number of qids generalized by g.

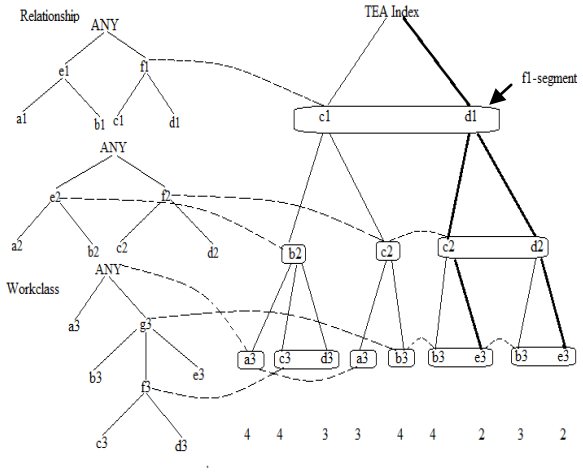

Figure 2: The TEA structure for QID = {Relationship, Race, Workclass}

Example 3.2.1
Figure 2 depicts three taxonomy trees for QID attributes {Relationship, Race, Workclass} and the TEA index for qids:
<c1, b2, a3>
<c1, b2, c3>
<c1, b2, d3>
<c1, c2, a3>
<c1, c2, b3>
<d1, c2, b3>
<d1, c2, e3>
<d1, d2, b3>
<d1, d2, e3>

A rectangle represents a segment, and a dashed line links up the segments of the same generalization. For example, the left-most path represents the qid = <c1, b2, a3>, and a(<c1, b2, a3>) = 4. {c1, d1} at level 1 is a segment of f1 because it forms a maximal set of siblings that will be merged by f1. {c1c2} and {d1c2, d1d2} at level 2 are two segments of f2. {c1b2c3, c1b2d3} at level 3 is a segment of f3. <d1, d2, e3> and <d1, c2, e3>, in bold face, are the anonymity qids.

Consider applying {c2, d2} → f2. The first segment of f2 contains only one sibling node {c1c2}, we simply re-label the sibling by f2. This creates new qids <c1, f2, a3> and <c1, f2, b3>. The second segment of f2 contains two sibling nodes {d1c2, d1d2}. We merge them into a new node labeled by f2, and merge their child nodes having the same label. This creates new qids <d1, f2, b3> and <d1, f2, e3>, with a(<d1, f2, b3>) = 7 and a(<d1, f2, e3>) = 4.

4. Conclusion

The paper investigated data mining as a technique for masking data, called *data mining based privacy protection*. The idea is to explore the data generalization concept from data mining as a way to hide detailed information, rather than discover trends and patterns. Once the data is masked, standard data mining techniques can be applied without modification. The paper demonstrated another positive use of the data mining technology: not only can it discover useful patterns, but also mask private information.

In particular, the paper presented a bottom-up generalization for transforming specific data to less specific but semantically consistent data for privacy protection.

5. Acknowledgement

This paper is benefited from conversations with many different people– far more than can be acknowledged here. Still we would like to particularly thank, HOD of CS&IT department for his guidance and support.

References

[1] K. Wang, P. S. Yu, and S. Chakraborty. Bottom-up generalization: A data mining solution to privacy protection. In Proc. of the 4th IEEE International Conference on Data Mining (ICDM), November 2004.
[2] R. Agrawal and R. Srikant. Privacy preserving data mining. In *SIGMOD*, 2000.
[3] Benjamin C. M. Fung, Ke Wang, Ada Wai-Chee Fu, and Philip S. Yu. Introduction to Privacy-Preserving Data Publishing Concepts and Techniques.
[4] L. Sweeney. k-anonymity: a model for protecting privacy. *International Journal on Uncertainty, Fuzziness and Knownledge-based Systems*, 10(5):557–570, 2002.
[5] L. Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal on Uncertainty, Fuzziness and Knownledge-based Systems*, 10(5):571–588, 2002.
[6] R. C. W. Wong, A. W. C. Fu, K. Wang, and J. Pei. Minimality attack in privacy preserving data publishing.

In Proc. of the 33rd International Conference on Very Large Data Bases (VLDB), Vienna, Austria, 2007.

[7] Valentina Ciriani, Sabrina De Capitani di Vimercati, Sara Foresti, and Pierangela Samarati. K anonymity. In T. Yu and S. Jajodia, editors, Security in Decentralized Data Management. Springer, Berlin Heidelberg, 2007.

[8] P. Samarati. Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering (TKDE), 2001.

[9] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical report, SRI International, March 1998.

[10] P. Samarati and L. Sweeney. Generalizing data to provide anonymity when disclosing information. In Proc. of the 17th ACM SIGACTSIGMOD- SIGART Symposium on Principles of Database Systems (PODS), Seattle, WA, June 1998.

[11] T. M. Truta and V. Bindu. Privacy protection: p-sensitive k-anonymity property. In Proc. of the Workshop on Privacy Data Management (PDM), April 2006.