IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 20, Issue 4, Ver. II (Jul - Aug 2018), PP 42-47
www.iosrjournals.org

Data Privacy through Optimal Classification in Large Datasets

[#1]arshiya Tabassum, [#2]p.Pranitha, [#3]dr.M.Sujatha

M.Tech Student, Associate Professor, Associate Professor, Department Of Cse, Jyothishmathi Institute Of Technological Sciences, Karimnagar,T.S.India.
Corresponding Author: Arshiya Tabassum

Abstract: *The sharing of information has been proven to be beneficial for business partnerships in many application areas such as business planning or marketing. Today, association rule mining imposes threats to data sharing, since it may disclose patterns and various kinds of sensitive knowledge that are difficult to find. Such information must be protected against unauthorized access. The challenge is to protect actionable knowledge for strategic decisions, but at the same time not to lose the great benefit of association rule mining. To address this challenge, a sanitizing process transforms the source database into a released database in which the counterpart cannot extract sensitive rules from it. Unlike existing works that focused on hiding sensitive association rules at a single concept level, this paper emphasizes on building a sanitizing algorithm for hiding association rules at multiple concept levels. Employing multi-level association rule mining may lead to the discovery of more specific and concrete knowledge from datasets. The proposed system uses learning algorithm as a biogeography-based optimization strategy for modifying multi-level items in database in order to minimize sanitization's side effects such as non-sensitive rules falsely hidden and fake rules falsely generated. The new approach is empirically tested and compared with other sanitizing algorithms depicting considerable improvement in completely hiding any given multi-level rule that in turn can fully support security of database and keeping the utility and certainty of mined multi-level rules at highest level using greedy algorithm.*

Keywords: *Database sanitization, learning algorithm, privacy preserving data mining, multi-level association rule hiding.*

Date of Submission: 23-07-2018 Date of acceptance: 07-08-2018

I. Introduction

Vast quantities of personal data are now collected in a wide variety of domains, including personal health records, emails, court documents, and the Web. It is anticipated that such data can enable significant improvements in the quality of services provided to individuals and facilitate new discoveries for society. At the same time, the data collected is often sensitive, and regulations, such as the Privacy Rule of the Health Insurance Portability and Accountability Act of 1996 (when disclosing medical records), Federal Rules of Civil Procedure (when disclosing court records), and the European Data Protection Directive often recommend the removal of identifying information. To accomplish such goals, the past several decades have brought forth the development of numerous data protection models. These models invoke various principles, such as hiding individuals in a crowd (e.g., k-anonymity) or perturbing values to ensure that little can be inferred about an individual even with arbitrary side information (e.g., ε-differential privacy). All of these approaches are predicated on the assumption that the publisher of the data knows where the identifiers are from the outset. More specifically, they assume the data has an explicit representation, such as a relational form, where the data has at most a small set of values per feature. However, it is increasingly the case that the data we generate lacks a formal relational or explicitly structured representation. A clear example of this phenomenon is the substantial quantity of natural language text which is created in the clinical notes in medical records. To protect such data, there has been a significant amount of research into natural language processing (NLP) techniques to detect and subsequently redact or substitute identifiers.

As demonstrated through systematic reviews and various competitions, the most scalable versions of such techniques are rooted in, or rely heavily upon, machine learning methods, in which the publisher of the data annotates instances of personal identifiers in the text, such as patients and doctor name, Social Security Number, and a date of birth, and the machine attempts to learn a classifier (e.g., a grammar) to predict where such identifiers reside in a much larger corpus. Unfortunately, generating a perfectly annotated corpus for training purposes can be extremely costly. This, combined with the natural imperfection of even the best classification learning methods implies that some sensitive information will invariably leak through to the data recipient. This is clearly a problem if, for instance, the information leaked corresponds to direct identifiers (e.g., personal name) or quasi-identifiers (e.g., ZIP codes or dates of birth) which may be exploited in re identification attacks, such as the re-identification of Thelma Arnold in the search logs disclosed by AOL or the Social

Security Numbers in Jeb Bush's emails. Rather than attempt to detect and redact every sensitive piece of information, our goal is to guarantee that even if identifiers remain in the published data, the adversary cannot easily find them. Fundamental to our approach is the acceptance of non-zero privacy risk, which we view as unavoidable.

This is consistent with most privacy regulation, such as HIPAA, which allows expert determination that privacy "risk is very small", and the EU Data Protection Directive, which "does not require anonymization to be completely risk free". Our starting point is a threat model within which an attacker uses published data to first train a classifier to predict sensitive entities based on a labelled subset of the data, prioritizes inspection based on the predicted positives, and inspects and verifies the true sensitivity status of B of these in a prioritized order. Here, B is the budget available to inspect (or read) instances and true sensitive entities are those which have been correctly labelled as sensitive (for example, true sensitive entities could include identifiers such as a name, Social Security Number, and address). We use this threat model to construct a game between a publisher, who 1) applies a collection of classifiers to an original data set, 2) prunes all the positives predicted by any classifier, and 3) publishes the remainder, and an adversary acting according to our threat model. The data publisher's ultimate goal is to release as much data as possible while at the same time redacting sensitive information to the point where re identification risk is sufficiently low. In support of the second goal, we show that any locally optimal publishing strategy exhibits the following two properties when the loss associated with exploited personal identifiers is high: a) an adversary cannot learn a classifier with a high true positive count, and b) an adversary with a large inspection budget cannot do much better than manually inspecting and confirming instances chosen uniformly at random (i.e., the classifier adds little value).

II. Related Work

Wu X, et al. [1] presents a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modelling, and security and privacy considerations.

[2]The Privacy Rule provides the first comprehensive federal protection for the privacy of health information. All segments of the health care industry have expressed their support for the objective of enhanced patient privacy in the health care system. At the same time, HHS and most parties agree that privacy protections must not interfere with a patient's access to or the quality of health care delivery.

[3]The Committee on Rules of Practice and Procedure and the Advisory Committee on the Federal Rules of Civil Procedure, Judicial Conference of the United States, prepared notes explaining the purpose and intent of the amendments to the rules.

[4]The directive regulates the processing of personal data regardless of whether such processing is automated or not.

Personal data are defined as "any information relating to an identified or identifiable natural person ("data subject"); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity;" The data protection rules are applicable not only when the controller is established within the EU, but whenever the controller uses equipment situated within the EU in order to process data. Controllers from outside the EU, processing data in the EU, will have to follow data protection regulation. In principle, any online business trading with EU residents would process some personal data and would be using equipment in the EU to process the data (i.e. the customer's computer). As a consequence, the website operator would have to comply with the European data protection rules. The directive was written before the breakthrough of the Internet, and to date there is little jurisprudence on this subject.

Fung B, et al. [5] stated to establish a nationwide system of electronic medical records that encourages sharing of medical knowledge through computer-assisted clinical decision support. In the data publishing phase, the data publisher releases the collected data to a data miner or to the public, called the data recipient, who will then conduct data mining on the published data. The data publisher is not required to have the knowledge to perform data mining on behalf of the data recipient. Any data mining activities have to be performed by the data recipient after receiving the data from the data publisher.

[6] Sweeney L examines re-identification attacks that can be realized on releases that adhere to k-anonymity unless accompanying policies are respected. The k-anonymity protection model is important because it forms the basis on which the real-world systems known as Data fly, µ-Argus and k-Similar provide guarantees of privacy protection.

[7] The problem is to release statistical information without compromising the privacy of the individual respondents. There are two settings: in the non-interactive setting the curator computes and publishes some statistics, and the data are not used further. Privacy concerns may affect the precise answers released by the curator, or even the set of statistics released.

[8] Sweeney L presents a formal presentation of achieving k-anonymity using generalization and suppression. Generalization involves replacing (or recoding) a value with a less specific but semantically consistent value. Suppression involves not releasing a value at all. While there are numerous techniques available combining these two offers several advantages.

A. Approaches for Anonymizing Structured Data

There has been a substantial amount of research conducted in the field of privacy-preserving data publishing (PPDP) over the past several decades. Much of this work is dedicated to methods that transform well-structured (e.g., relational) data to adhere to a certain criterion or a set of criteria, such as k-anonymization, l-diversity, m-invariance, and ε differential privacy, among a multitude of others. These criteria attempt to offer guarantees about the ability of an attacker to either distinguish between different records in the data or make inferences tied to a specific individual. There is now an extensive literature aiming to operationalize such PPDP criteria in practice through the application of techniques such as generalization, suppression (or removal), and randomization. All of these techniques, however, rely on a priori knowledge of which features in the data are either themselves sensitive or can be linked to sensitive attributes. This is a key distinction from our work: we aim to automatically discover which entities in unstructured data are sensitive, as well as formally ensure that whatever sensitive data remains cannot be easily unearthed by an adversary.

B. Traditional Methods for Sanitizing Unstructured Data

In the context of privacy preservation for unstructured data, such as text, various approaches have been proposed for the automatic discovery of sensitive entities, such as identifiers. The simplest of these rely on a large collection of rules, dictionaries, and regular expressions. An automated data sanitization algorithm aimed at removing sensitive identifiers while inducing the least distortion to the contents of documents. However, this algorithm assumes that sensitive entities, as well as any possible related entities, have already been labelled. Similarly, have developed the t-plausibility algorithm to replace the known (labelled) sensitive identifiers within the documents and guarantee that the sanitized document is associated with least t documents.

C. Machine Learning Methods for Sanitizing Unstructured Data

A key challenge in unstructured data that makes it qualitatively distinct from structured is that even identifying (labelling) which entities are sensitive is non-trivial. For example, while a structured portion of electronic medical records would generally have known sensitive categories, such as a patient’s name, physician’s notes do not have such labels, even though they may well refer to a patient’s name, date of birth, and other potentially identifying information. While rule-based approaches, such as regular expressions, can automatically identify some of the sensitive entities, they have to be manually tuned to specific classes of data, and do not generalize well. A natural idea, which has received considerable traction in prior literature, is to use machine learning algorithms, trained on a small portion of labeled data, to automatically identify sensitive entities. Numerous classification algorithms have been proposed for this purpose, including decision stumps, support vector machines (SVM), conditional random fields (CRFs), hybrid strategies that rely on rules and statistical learning models ensemble methods. Unfortunately, such PPDP algorithms fail to formally consider the adversarial model, which is crucial for the decision making of the data publisher. A recent work by Carrel considers enhancing such redaction methods by replacing removed identifiers with fake identifiers which appear real to a human reader. Our approach builds on this literature, but is quite distinct from it in several ways. First, we propose a novel explicit threat model for this problem, allowing us to make formal guarantees about the vulnerability of the published data to adversarial re-identification attempts. Our model bears some relationship to a recent work by Li who also consider an adversary using machine learning to re-identify residual identifiers. However, our model combines this with a budget-limited attacker who can manually inspect instances; in addition, our publisher model involves the choice of a redaction policy, whereas Li et al. focus on the publisher’s decision about the size of the training data, and use a traditional learning-based redaction approach. Second, we introduce a natural approach for sanitizing data that uses machine learning in an iterative framework. Notably, this approach performs significantly better than a standard application of CRFs, which is the leading approach for text sanitization to date, but can actually make use of arbitrary machine learning algorithms.

D. Game Theory in Security and Privacy

Our work can be seen within the broader context of game theoretic modelling of security and privacy, including a number of efforts that use game theory to make machine learning algorithms robust in adversarial environments. In both of these genres of work, a central element is an explicit formal threat (i.e., attacker) model, with the game theoretic analysis generally focused on computing defensive privacy preserving strategies.

None of this work to date, however, addresses the problem of PPDP of unstructured data with sensitive entities not known a priori.

In existing system there has been a substantial amount of research conducted in the field of privacy-preserving data publishing (PPDP) over the past several decades.

These criteria attempt to offer guarantees about the ability of an attacker to either distinguish between different records..

- .Linked to sensitive attributes.
- Crucial for the decision making of the data publisher.

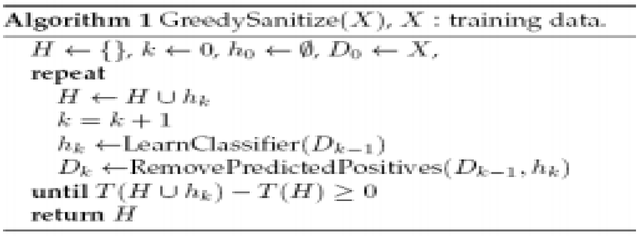
III. Proposed Model

Before delving into the technical details, we offer a brief high-level intuition behind the main idea in this paper. Suppose that a publisher uses a machine learning algorithm to identify sensitive instances in a corpus, these instances are then redacted, and the residual data is shared with an attacker. The latter, aspiring to uncover residual sensitive instances (e.g., identifiers) can, similarly, train a learning algorithm to do so (using, for example, a subset of published data that is manually labelled). At the high level, consider two possibilities: first, the learning algorithm (learning algorithm is a method used to process data to extract patterns appropriate for application in a new situation. In particular, the goal is to adapt a system to a specific input-output transformation task) enables the attacker to uncover a non-trivial amount of sensitive information, and second, the learning algorithm is relatively unhelpful in doing so. In the latter case, the publisher can perhaps breath freely: few sensitive entities can be identified by this attacker, and the risk of published data is low. The former case is, of course, the problem. However, notice that, in principle, the publisher can try out this attack in advance of publishing the data, to see whether it can in fact succeed in this fashion. Moreover, if the attacker is projected to be sufficiently successful, the publisher has a great deal to gain by redacting the sensitive entities an attacker would have found. Of course, there is no need to stop at this point: the publisher can keep simulating attacks on the published data, and redacting data labelled as sensitive, until these simulations suggest that the risk is sufficiently low. This, indeed, is the main idea. However, many details are clearly missing: for example, what does an attacker do after training the learning algorithm, when, precisely, should the publisher stop, and what can we say about the privacy risk if data is published in this manner, under this threat model? Next, we formalize this idea, and offer precise answers to these and other relevant questions.

- Supporting the goal of maximizing the quantity of released data.
- The number of required algorithm iterations.
- The number of residual true positives is always quite small.
- Choosing entities to manually inspect.

IV. A Greedy Algorithm For Automated Data Sanitization

We can now present our iterative algorithm for automated data sanitization, which we term Greedy Sanitize. Our algorithm (shown as Algorithm 1) is simple to implement and involves iterating over the following steps: 1) compute a classifier on training data, 2) remove all predicted positives from the training data, and 3) add this classifier to the collection. The algorithm continues until a specified stopping condition is satisfied, at which point we publish only the predicted negatives, as above. While the primary focus of the discussion so far, as well as the stopping criterion, have been to reduce privacy risk, the nature of Greedy Sanitize is to also preserve as much utility as feasible: this is the consequence of stopping as soon as the re identification risk is minimal. It is important to emphasize that Greedy Sanitize is qualitatively different from typical ensemble learning schemes in several ways. First, a classifier is retrained each iteration on data that includes only predicted negatives from all prior iterations. To the best of our knowledge this is unlike the mechanics of any ensemble learning algorithm.1 Second, our algorithm removes the union of all predicted positives, whereas ensemble learning typically applies a weighted voting scheme to predict positives; our algorithm, therefore, is fundamentally more conservative when it comes to sensitive entities in the data. Third, the stopping condition is uniquely tailored to the algorithm, which is critical in enabling provable guarantees about privacy-related performance.



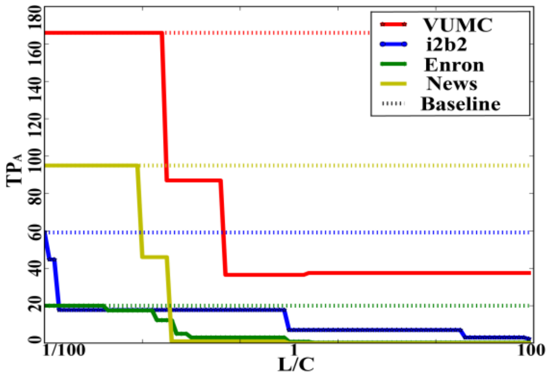
Algorithm 1 GreedySanitize(X), X : training data.

```
$H \leftarrow \{\}$, $k \leftarrow 0$, $h_0 \leftarrow \emptyset$, $D_0 \leftarrow X$,
repeat
    $H \leftarrow H \cup h_k$
    $k = k + 1$
    $h_k \leftarrow$LearnClassifier($D_{k-1}$)
    $D_k \leftarrow$RemovePredictedPositives($D_{k-1}, h_k$)
until $T(H \cup h_k) - T(H) \geq 0$
return $H$
```

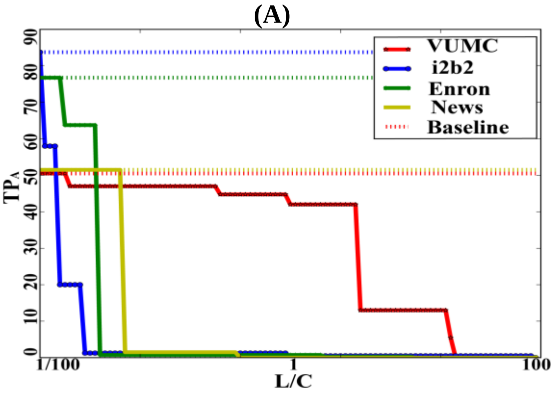
V. Result Analysis

Privacy Risk:

When the budget of the attacker is small, the theoretical results provide an upper bound on the expected number of identified instances. While this upper bound suggests that risk becomes arbitrarily small when the associated loss is large, it is not tight. In Figure 3 demonstrate that the number of identified instances (which is equivalent to the number of true positives for the attacker's classifier) typically becomes negligible even when L is quite small relative to C.



(A)



(B)

The number of residual true positive instances TPA (equivalently, identified instances for an attacker with a small budget) after running GreedySanitize for the i2b2, VUMC, Enron, and Newsgroup datasets. (a) GreedySanitize using CRF (dashed lines, or baseline, correspond to standard application of CRF). (b) GreedySanitize using the best classifier from fCRF, SVM, Ensembleg (dashed lines correspond to the baseline application of the best classifier from this collection).

VI. Conclusion & Future Work

Our ability to take full advantage of large amounts of unstructured data collected across a broad array of domains is limited by the sensitive information contained therein. This paper introduced a novel framework for sanitization of such data that relies upon 1) a principled threat model, 2) a very general class of publishing strategies, and 3) a greedy, yet effective, data publishing algorithm. The experimental evaluation shows that our algorithm is: a) substantially better than existing approaches for suppressing sensitive data, and b) retains most of the value of the data, suppressing less than 10% of information on all four data sets we considered in evaluation. In contrast, cost-sensitive variants of standard learning methods yield virtually no residual utility, suppressing most, if not all, of the data, when the loss associated with privacy risk is even moderately high. Since our adversarial model is deliberately extremely strong- far stronger, indeed, than is plausible - our results suggest feasibility for data sanitization at scale.

References

[1]. X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97–107, 2014.

[2]. U.S. Dept. of Health and Human Services, "Standards for privacy and individually identifiable health information; final rule," Federal Register, vol. 65, no. 250, pp. 82 462–82 829, 2000.

[3]. Committe on the Judiciary House of Representatives, "Federal Rules of Civil Procedure," 2014.

[4]. European Parliament and Council of the European Union, “Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data,” Official Journal of the EC, vol. 281, pp. 0031–0050, 1995.

[5]. B. Fung, K. Wang, R. Chen, and P. S. Yu, “Privacy preservingdata publishing:Asurvey of recent developments,” ACM Computing Surveys, vol. 42, no. 4, p. 14, 2010.

[6]. L. Sweeney, “k-anonymity: A model for protecting privacy,” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 05, pp. 557–570, 2002.

[7]. C. D work, “Differential privacy: A survey of results,” in International Conference on Theory and Applications of Models of Computation, 2008, pp. 1–19.

[8]. L. Sweeney, “Achieving k-anonymity privacy protection using generalization and suppression,” International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 05, pp. 571–588, 2002.

[9]. Y. He and J. F. Naughton, “Anonymization of set-valued data via top-down, local generalization,” VLDB Endowment, vol. 2, no. 1,pp. 934–945, 2009.

[10]. G. Poulis, A. Gkoulalas-Divanis, G. Loukides, S. Skiadopoulos, and C. Tryfonopoulos, “SECRETA: A system for evaluating and comparing relational and transaction anonymization algorithms,” in International Conference on Extending Database Technology, 2014, pp. 620–623.

[11]. G. Poulis, G. Loukides, A. Gkoulalas-Divanis, and S. Skiadopoulos, “Anonymizing data with relational and transaction attributes,” in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2013, pp. 353–369.

[12]. M. Terrovitis, N. Mamoulis, and P. Kalnis, “Privacypreserving anonymization of set-valued data,” VLDB Endowment, pp. 115–125, 2008.

arshiya Tabassum "Data Privacy Through Optimal Classification In Large Datasets "IOSR Journal of Computer Engineering (IOSR-JCE) 20.4 (2018): 42-47.