Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 37 (2014) 511 – 516

Procedia Computer Science

International Workshop on Privacy and Security in HealthCare (PSCare14)

Using Feature Selection to Improve the Utility of Differentially Private Data Publishing

Yasser Jafer*[1], Stan Matwin[1,2,3], Marina Sokolova[1,2,4]

[1] *School of Electrical Engineering and Computer Science, University of Ottawa, Canada*
[2] *Institute for Big Data Analytics, Dalhousie University, Canada*
[3] *Institute for Computer Science, Polish Academy of Sciences*
[4] *Faculty of Medicine, University of Ottawa, Canada*

Abstract

Protection of patient's privacy is an obligation enforced by laws and regulations in the US, Canada, and other jurisdictions. With exponential growth of exchange of personal health information (PHI) brought about by e-health, there is a need for smart algorithms that help the data publisher to protect PHI. Within exiting privacy models, differential privacy is considered one of the strongest privacy protection techniques that does not make any assumption about the attacker's background knowledge. One way to achieve differential privacy in the non-interactive mode is to derive a contingency table of the raw data over the database domain, to add noise to each count, and to publish the resulting noisy table of counts. This approach, however, is not suitable for high-dimensional data with large domains as the added noise substantially destroys the utility of the data. In this work, we show that when the K-anonymity is preceded by feature selection, it is possible to obtain a contingency table with higher counts. As a result, when noise is added to satisfy differential privacy, its distorting effect is minimized and high utility of the data is preserved. We propose the *TOP_Diff* algorithm which offers a trade-off between anonymization level K and the privacy budget ε, and enables us to publish privacy preserving datasets with high utility. Our approach is capable of handling both numerical and categorical features.

Peer-review under responsibility of the Program Chairs of EUSPN-2014 and ICTH 2014.

Keywords: Privacy, Feature Selection, K-anonymity, Differential Privacy, Classification

* Corresponding author.
E-mail address: yjafe089@uottawa.ca

Peer-review under responsibility of the Program Chairs of EUSPN-2014 and ICTH 2014.
doi:10.1016/j.procs.2014.08.076

1. Introduction

Protection of patient's privacy is an obligation enforced by laws and regulations such as HIPAA(Health Information Portability and Accountability Act)[9] in US, PHIPA(Personal Health Information Protection Act)[10] in Ontario, etc. Health care organizations are major Data Holders of patient's personal health information (PHI) and as such, are obliged to implement the best practices of the PHI protection. For example, raw data needs to be modified before release, and the modification is done via a number of anonymization operations[3].

In general, attributes in a dataset can be categorized into (i) *explicit identifier*, (ii) *quasi identifiers*, (iii) *sensitive* and (iv) *non-sensitive* attributes. *Explicit identifiers* refer to a set of attributes that explicitly identify individuals. *Quasi Identifiers (QI)* refer to a set of attributes that could be linked to external datasets and potentially breach the privacy. *Sensitive* attributes correspond to person-specific private information. Finally, *non-sensitive* attributes consist of attributes that do not fall into any of the above categories. While the explicit identifiers are removed from the table, the QI set is transformed into a less specific form (QI') by applying anonymization operations. For example, a table is considered K-anonymous if the QI values of each tuple are indistinguishable from "at least" K-1 other tuples. K-anonymity belongs to syntactic anonymity approaches which are known to be susceptible to various attacks[6]. There are also common limitations associated with these approaches such as information loss, ad hoc assumption on auxiliary information, and sub-optimality[4]. In order to respond to the needs for a firm foundation for privacy preserving data publishing, differential privacy was proposed by Dwork[5]. Differential privacy ensures that adding or removing a single dataset item does not substantially influence the outcome of any analysis. Differential privacy supports a rigorous notion of privacy. However, a study of its utility is still in its infancy[6]. A fruitful research direction is to combine the benefits associated with syntactic anonymity approaches and differential privacy[7,8] in order to enhance utility while guaranteeing differential privacy. A main approach to guarantee differential privacy of the data is through *non-interactive* means. The current *non-interactive* strategy is to publish a noisy contingency table (i.e. table of counts) [11]. This is achieved by deriving a frequency matrix of the original data over the database domain. After obtaining the counts, noise is added to each count in order to satisfy differential privacy. However, the issue with publishing noisy contingency tables is that such approach is not suitable for high-dimensional data that represent large domains. In latter setting, the added noise becomes very large compared with the counts and therefore, the utility of the data is substantially degraded to the level that it makes the data useless.

Privacy preserving data publishing focuses on anonymizing and releasing datasets which are used for data mining and other analytics purposes. Usually, in this scenario, the purpose of data analysis is not known before hand. However, if the data publishing techniques are customized according to a particular type of analysis, better results can be obtained[1]. In this work, we follow this assumption and consider a scenario which includes a Data Holder (DH) that holds the original data (e.g. hospital) and a Data Recipient (DR) who wants the data in order to apply certain data mining task[2] (e.g. a research center).

In this work, we propose a novel technique for privacy preserving data publishing satisfying differential privacy and use feature selection in order to minimize the negative impact of injecting noise into the contingency table. We show that when feature selection is applied on the dataset prior to K-anonymization, we obtain contingency tables with high counts. Consequently, when noise is added to each count to satisfy differential privacy, the amount of noise is well compensated by the higher counts resulting from incorporating feature selection into K-anonymity. Our technique enables us to trade-off the level of anonymization and the amount of noise and to obtain a dataset that satisfies both the privacy and the utility requirements. Since the data publishing approach presented here is designed so as to precede data use for, e.g., model building or other kinds of data analytics, we view this as an instance of the Privacy by Design paradigm applied in a data analytics context.

2. Preliminaries

2.1 Feature Selection

Feature selection aims at removing irrelevant and/or redundant attributes in order to improve the quality of data. It is also considered an effective dimensionality reduction method[12]. There are two broad categories of feature selection techniques, namely, *filters* and *wrappers*. Filter approach attempts to assess the merits of features from the data without considering the induction algorithm. The wrapper model, on the other hand, uses a target learning algorithm in order to estimate the worth of attribute subsets. Previous works have shown that the wrapper feature