Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Privacy and utility preserving data clustering for data anonymization and distribution on Hadoop

J. Jesu Vedha Nayahi[a,*], V. Kavitha[b]

[a] Department of Computer Science and Engineering, Anna University Regional Campus - Tirunelveli, Tirunelveli - 627007, Tamil Nadu, India

[b] Department of Computer Science and Engineering, University College of Engineering, Kancheepuram Campus, Kancheepuram - 631552, India

HIGHLIGHTS

- We have proposed a clustering algorithm to achieve anonymization overcoming similarity attack, probabilistic inference attack and the other common attacks feasible with anonymized data set.
- A privacy-preserving distributed framework is proposed to anonymize the data set and distribute the anonymized data set on a distributed environment without threat to data privacy.
- The anonymized data set is distributed on Hadoop Distributed File System
- The better trade-off between privacy and data utility is achievable and the data utility is shown in terms of traditional metrics.
- Moreover different classifiers are applied on the privacy-preserved data set to estimate the data utility in terms of FMeasure and Percentage of Correctly Classified Instances.

ARTICLE INFO

Article history:
Received 21 June 2016
Received in revised form
20 October 2016
Accepted 23 October 2016
Available online xxxx

Keywords:
Data privacy
Clustering
Data anonymization
Hadoop
Classification
Distributed data

ABSTRACT

Data privacy is a stringent need when sharing and processing data on a distributed environment or in Internet of Things. Collaborative privacy-preserving data mining based on secured multiparty computation incur high communication and computational cost. Data anonymization is a promising technique in the field of privacy-preserving data mining used to protect the data against identity disclosure. Information loss and common attacks possible on the anonymized data are serious challenges of anonymization. Recently, data anonymization using data mining techniques has showed significant improvement in data utility. Still the existing techniques lack in effective handling of attacks. Hence in this paper, an anonymization algorithm based on clustering and resilient to similarity attack and probabilistic inference attack is proposed. The anonymized data is distributed on Hadoop Distributed File System. The method achieves a better trade-off between privacy and utility. In our work the data utility is measured in terms of accuracy and FMeasure with respect to different classifiers. Experiments show that the accuracy, FMeasure and the execution time of the classification algorithms on the privacy-preserved data sets formed by the proposed clustering algorithms are better than the existing algorithms.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Today data is inherently distributed on various geographically distributed sites. Researchers are interested in the knowledge extracted from the integrated collection of data available at these sites. Consider the case of a sudden epidemic disease spreading in the society and the researchers are looking for the causes, symptoms and treatments for that particular disease. The data mining performed on data of patients available from hospitals worldwide/nationwide would be more efficient than what done on the data from a single hospital.

Internet of Things (IoT) is an emerging field in which data and other things are exchanged through a network of objects. IoT has its applications in various fields such as patient monitoring system, traffic control system, inventory management system, etc. In all these applications the user's identity and sensitive information related to their interaction and mobility should be protected. This leads to guaranteeing the data privacy of individuals in IoT. Moreover the health records of patients maintained at any hospital

[*] Corresponding author.
E-mail addresses: vedhaj2000@gmail.com, vedhaj2000@yahoo.com (J.J.V. Nayahi), kavinayav@gmail.com (V. Kavitha).

http://dx.doi.org/10.1016/j.future.2016.10.022
0167-739X/© 2016 Elsevier B.V. All rights reserved.

Please cite this article in press as: J.J.V. Nayahi, V. Kavitha, Privacy and utility preserving data clustering for data anonymization and distribution on Hadoop, Future Generation Computer Systems (2016), http://dx.doi.org/10.1016/j.future.2016.10.022

should be kept confidential according to the privacy law [1]. This demands effective mechanism to maintain the privacy of individuals while sharing the data on distributed environment [2], IoT [3–5] etc. The data owners need privacy-preserving data publishing [6,7] or privacy-preserving data mining [8] techniques to publish their microdata or share knowledge to third party respectively.

Data anonymization [9–14], data randomization [15–18] and cryptography [19,20] are some of the major techniques used in the field of privacy-preserving data mining or data publishing. k-anonymization is the process of anonymizing the records such that k individuals become indistinguishable from each other. This mechanism protects the record from identity disclosure or linking attack. Linking attack or identity disclosure is the possibility that an attacker discloses the sensitive attribute value of a person with the known values of Quasi Identifier *(QID)* attributes. Anonymization is achieved either by generalization or suppression [14] of the *QID* attributes. Generalized equivalence classes are formed by replacing the specific *QID* values using more generalized values. Suppression completely removes the *QID* values by replacing the original value with a '*'. Generalization is an NP-hard problem and suppression leads to tremendous loss of information. k-anonymization is the commonly used technique to preserve the health care data against identity disclosure [21].

Han et al. [22] proposed a two steps clustering based k-anonymization algorithm suitable for both numerical and categorical data. This method combines both the common methods for anonymization namely generalization and microaggregation. Mortazavi and Jalili [23] proposed fast data-oriented microaggregation algorithm for anonymization. The partitions are formed minimizing the information loss and satisfying the anonymization parameter k. Gkoulalas-Divanis et al. [24] proposed privacy and utility preserving anonymization algorithm and proved that anonymization is the best known mechanism to overcome identity disclosure. Wimmera et al. [25] proposed multiagent privacy-preserving medical data decision making using k-anonymization. The data integrated from multiple sources are anonymized to an acceptable level of accuracy. k-anonymization is prone to some of the common attacks such as homogeneity attack, similarity attack, background knowledge attack and probabilistic inference attack.

Machanavajjhala et al. [26] proposed l-diversity principle to bring l well represented sensitive values in each equivalence class group. Li et al. [27] proposed t-closeness principle to overcome the limitations of l-diversity principle. The distinct l-diversity is not sufficient to protect against probabilistic inference attack and it would be infeasible to achieve entropy diversity when some sensitive values are more frequent than the others. t-closeness principle makes the distribution of sensitive values in each equivalence class formed to be as close to the distribution of the sensitive values in the original data set. But it is practically infeasible to determine t-close equivalence classes. Domingo-Ferrer and Soria-Comas [28] proposed a stochastic t-closeness principle combining k-anonymity and ϵ-differential privacy. The authors overcome the shortcomings of the deterministic t-closeness principle. Soria-comas et al. [29] proposed three different microaggregation algorithms to achieve k-anonymous t-close equivalence classes. Among which the clustering algorithm used to achieve t-close equivalence classes first show minimum loss of information. The algorithms are suitable only for numerical data.

The traditional methods are prone to common attacks and also fail to achieve a better trade-off between privacy and utility. The use of data mining techniques to achieve anonymization would improve the accuracy of knowledge extracted from these data [30,31]. Wen-Yang et al. [32] proposed a MS (k, θ^2) privacy model for Adverse Drug Reaction data using greedy clustering algorithm. The algorithm forms anonymized data satisfying anonymization parameter k and preserves the data utility. Similarity attack is a very serious problem with data anonymization. Existing methods on k-anonymization and l-diversity are not commonly employed for preserving data privacy in a distributed environment. According to Chen and Liu [16] data anonymization would be a comparatively better technique to achieve privacy in a fully distributed setting. Hence in this paper we intend to employ k-anonymization to preserve the data privacy in a distributed environment. We proposed a clustering algorithm to achieve k-anonymization and l-diversity resilient to similarity attack and probabilistic inference attack. Later the privacy-preserved anonymized data sets are distributed on Hadoop [33,34].

2. Related work

In this section we present the techniques developed to handle similarity attack and other attacks feasible with anonymized equivalence classes. The merits and demerits of privacy-preserving data mining algorithms on distributed data using secured multiparty computation is studied. We have also studied some of the recent approaches used to preserve data privacy based on anonymization and hybrid approaches.

2.1. Techniques handling common attacks

(α, k) anonymity achieves k-anonymity property and α-deassociation property to achieve privacy-preserving data publishing. The relative frequency of the most frequently occurring sensitive value in each equivalence class should be less than or equal to the user defined threshold α. The authors show that achieving (α, k) anonymization is also NP-hard [35]. Traian et al. [36] proposed (p, α) and $(p+, \alpha)$ sensitive k-anonymity for data publishing. (p, α) sensitive property is proposed to achieve p distinct sensitive values in each equivalence class with a total weight of at least α. Even this property may not be sufficient to protect data against similarity attack. Hence [37] categorize the sensitive attribute values into four categories ranging from Top Secret to Non Secret and then form equivalence classes each with diverse sensitive values from each category. $(p+, \alpha)$ sensitive property is proposed to achieve at least p distinct categories of sensitive attribute values with total weight at least α. Sun et al. [38] proposed (L, α) diversity privacy models based on l-diversity and categorization of sensitive attributes values into different levels of confidentiality. Tian and Zhang [39] enhances l-diversity using functional (τ, l) diversity. The sensitive values are also generalized along with the *QID* attributes until the (τ, l) diversity is satisfied. All the parameters mentioned in the above mentioned algorithms are user defined and are used to set appropriate threshold on the size of the equivalence class and the number of distinct values of the sensitive attribute value. When the input data set contains only a few possible values for the sensitive attribute or if some sensitive values occur more frequently than the others, overcoming similarity attack with the existing techniques is infeasible.

Sattar et al. [40] proposed anonymization based on generalization with sampling to disguise the adversary's confidence on the sensitive information of an individual. Amiri et al. [41] proposed a k-anonymous β-likeness model for protection of data against both identity disclosure and attribute disclosure. The authors have developed two clustering algorithms to create k-anonymous β-likeness privacy model and overcome background knowledge attack. Komishani et al. [42] proposed personalized privacy model based on anonymization for trajectory data. The sensitive attribute is generalized whereas the trajectory data is suppressed. The method overcomes linking attack and similarity attack. Zhang et al. [43] proposed a two phase clustering algorithm to preserve data privacy on cloud using anonymization. The semantic diversity of sensitive attribute value is considered to form equivalence classes with diverse sensitive values.

Download English Version:

https://daneshyari.com/en/article/4950389

Download Persian Version:

https://daneshyari.com/article/4950389

Daneshyari.com