Information Sciences 232 (2013) 437–448

Contents lists available at SciVerse ScienceDirect

Information Sciences

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

Evaluation of a perturbation-based technique for privacy preservation in a multi-party clustering scenario

R. Vidya Banu [a,*], N. Nagaveni [b]

[a] Sri Krishna College of Engineering and Technology, Coimbatore, TamilNadu, India
[b] Coimbatore Institute of Technology, Coimbatore, TamilNadu, India

ARTICLE INFO

Article history:
Available online 7 March 2012

Keywords:
Data transformation
Multi-party clustering
K-means clustering
Self-organising map
Geometrical data transformation method
Principal component analysis

ABSTRACT

Data processing techniques and the growth of the internet have resulted in a data explosion. The data that are now available may contain sensitive information that could, if misused, jeopardise the privacy of individuals. In today's web world, the privacy of personal and personal business information is a growing concern for individuals, corporate entities and governments. Preserving personal and sensitive information is critical to the success of today's data mining techniques. Preserving the privacy of data is even more crucial in critical sectors such as defence, health care and finance. Privacy Preserving Data Mining (PPDM) addresses such issues by balancing the preservation of privacy and the utilisation of data.

Traditionally, Geometrical Data Transformation Methods (GDTMs) have been widely used for privacy preserving clustering. The drawback of these methods is that geometric transformation functions are invertible, which results in a lower level of privacy protection. In this work, a Principal Component Analysis (PCA)-based technique that preserves the privacy of sensitive information in a multi-party clustering scenario is proposed. The performance of this technique is evaluated further by applying a classical K-means clustering algorithm, as well as a machine learning-based clustering method on synthetic and real world datasets. The accuracy of clustering is computed before and after privacy-preserving transformation. The proposed PCA-based transformation method resulted in superior privacy protection and better performance when compared to the traditional GDTMs.

1. Introduction

Advancements in data compression, the advent of lower storage costs and the increasing usage of the World Wide Web have played a large role in capturing and storing large volumes of data; a suitable goal for these data is their use to gain insight into hidden patterns by using data mining tools and techniques. With these advancements, there are growing concerns about the privacy of personal and sensitive information. People hesitate to share their personal data, which can result in skewing the outcome of the data mining because the data collected may then contain incorrect or incomplete information. An overview of the risks that are associated with data mining, along with the strategies that have been proposed over the years to mitigate these risks, is given in [22]. Clifton and Marks [6] have analysed how data mining can bring threats to databases and discussed many solutions to protect privacy during data mining. Most of the traditional PPDM algorithms preserve the privacy of data by transforming the original data in such a way that the utility of the data is not lost. The ability to analyse private data without violating the privacy of the individuals has contributed to the popularity of PPDM.

* Corresponding author.
E-mail address: vidhyabanu@yahoo.com (R. Vidya Banu).

0020-0255/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2012.02.045

The mere removal of unique identifiers in the data will not ensure the preservation of privacy because the actual values can still be guessed by associating other attributes with publicly available knowledge [16]. Data privacy is viewed as a legal, social and cultural concept. Now, with the emergence of the web, privacy has become an issue of serious concern. Hence, any data mining algorithm must be properly scrutinised, to eliminate privacy breaches. There are multiple factors that contribute to a violation of privacy in data mining, and data can be misused in a number of ways. One source of data privacy violation is the use of data magnets; these tools are used for collecting private data. They include techniques such as collecting information through on-line registration, identifying users through IP addresses and indirectly collecting information for secondary usage. In most of the cases, the users will be totally or partially unaware of the fact that their personal information is being collected.

The idea behind privacy-preserving clustering (PPC) is to protect the sensitive attribute values of objects that are subjected to clustering analysis. Cano et al. [2] have suggested that synthetic data generation using the ISPO family of methods is a way to preserve data privacy in clustering. Oliveira and Zaiane [18] explored different types of scenarios in Privacy Preserving Clustering. These scenarios are the following:

1.1. PPC over horizontally partitioned data

Assume that Alice A and Bob B have two unlabelled samples, DA and DB. Assume that each sample in DA and DB has all of the attributes or that the data sets are horizontally partitioned between A and B. Alice and Bob want to cluster the joint data set DA U DB without revealing the individual items in their data sets.

1.2. PPC over centralised data

In this scenario, two parties A and B are involved, party A owning a dataset D and party B wanting to mine it for clustering. The dataset is assumed to be a data matrix $D_{m\times n}$, where each of the m rows represents an entity or object, and each entity contains values for each of the n attributes. The matrix $D_{m\times n}$ may contain binary, categorical, or numerical attributes. Before sharing the dataset D with party B, party A must transform D to preserve the privacy of the individual data records. However, the transformation applied to D must not jeopardise the similarity between the objects.

1.3. PPC over vertically partitioned data

Consider a scenario where k parties, $k >= 2$, have different attributes for a common set of objects. The goal here is to perform a join over the k parties and then to cluster the common objects. After performing a join over the k parties, the problem of PPC over vertically partitioned data becomes a problem of PPC of centralised data.

Privacy issues are further aggravated because of the rapid growth of internet technologies. Preserving privacy on the web has become an important concern because of the remarkable growth of e-business and e-commerce. Reay et al. [20] have analysed privacy issues on the web and discussed how the economic sector is currently endangered by consumers' well-founded concerns for privacy. The ready availability of personal data through the web has made it easier for malicious users to collect sensitive information, thereby causing serious threats to individual privacy. To handle these privacy issues, a standardisation framework for PPDM should be developed in terms of definitions, principles, policies and requirements.

2. Related work

Most of the algorithms for Privacy Preserving Data Mining (PPDM) that are proposed in the literature are based on perturbation, randomisation and secure multi-party computation. To limit disclosure risk when providing data for mining, Sweeney [24] introduced the k-anonymity privacy requirement, which requires each record in an anonymised table to be indistinguishable from at least $k-1$ other records within the dataset, with respect to a set of quasi-identifiers. Very recently, Matatov et al. [15] suggested an approach for achieving k-anonymity by partitioning the original dataset into several projections such that each adheres to k-anonymity. Top-down specialisation techniques for privacy preservation are discussed in [9]. Samariti [21] proposed the usage of generalisation and suppression techniques to protect the identities of respondents while releasing micro data. Tian and Zhang [25] have proposed an extended l-diversity model called the functional diversity measure, to constrain the frequencies of base sensitive attribute values that are induced by general sensitive attribute values. Yeh and Hsu [29] have proposed algorithms that focus on privacy preserving utility mining for hiding sensitive item sets from adversaries. Yang and Qiao [28] have proposed a method that anonymises data by randomly breaking links among attribute values in records. Chen and Liu [4] has presented the geometric perturbation approach to multi-party privacy-preserving collaborative mining. Oliveira and Zaiane [17] have defined a family of Geometric Data Transformation Methods (GDTMs) for privacy preserving transformation, as follows.

The Translation Data Perturbation Method (TDP) uses additive noise perturbation to perturb the confidential attributes. In the Scaling Data Perturbation Method (SDP), the confidential attributes are perturbed using multiplicative noise while the Rotation Data Perturbation Method (RDP) works by identifying a common rotation angle between any two attributes and applying this transformation to the confidential attributes. The Hybrid Data Perturbation Method (HDP) combines the vigour