A Hybrid Data Transformation Approach for Privacy Preserving Clustering of Categorical Data

Dr.A.M.Natarajan
Principal
Kongu Engineering College
Perundurai
Erode, TamilNadu
India
principal@kongu.ac.in

R.R.Rajalaxmi
Assistant Professor / CSE-PG
Kongu Engineering College
Perundurai
Erode, TamilNadu
India
rrr_kec@yahoo.co.in

N.Uma
Final ME
Kongu Engineering College
Perundurai
Erode, TamilNadu
India
umamanivannan@yahoo.com

G.Kirubhakar
Final ME
Kongu Engineering College
Perundurai
Erode, TamilNadu
India
kirubhakar@rediffmail.com

Abstract - In today's information age there is a large availability of repositories storing various types of information about individuals. Data mining technology has emerged as a means of identifying patterns and trends from large quantities of data. The application of data mining technology to identify interesting patterns from these repositories leads to serious privacy concerns. Despite its benefit in a wide range of applications, data mining techniques also have raised a number of ethical issues. Some such issues are privacy, data security, intellectual property rights and many others. In this paper, we address the privacy problem against unauthorized secondary use of information. We focus primarily on privacy preserving data clustering on categorical data. In the proposed method, the categorical data is converted into binary data and it is transformed using geometric data transformation method. Then, clustering using conventional clustering algorithm is done on the transformed data to ensure privacy.

Index Terms: Clustering, Categorical data, Data Transformation, Binary data, Translation, Rotation, Scaling.

1. INTRODUCTION

The organizations collect data, often concerning individuals, and use them for various purposes, ranging from scientific research, as in the case of medical data, to demographic trend analysis and marketing purposes [1]. Organizations may also give access to the data they own or even release such data to third parties. The number of increased data sets that are thus available poses serious threats against the privacy of individuals and organizations.

Privacy preserving data mining (PPDM) is a novel research direction in data mining, where data mining algorithms are analyzed for side effects they incur in data privacy. The main consideration in privacy preserving data mining is two fold[1]. First, sensitive raw data like identifiers, names, addresses and the like should be modified from the original data base, in order for the recipient of the data not be able to compromise another persons privacy. Second, sensitive knowledge which can be mined from a database by using data mining algorithm should also be excluded, because such knowledge can equally well compromise data privacy. So, the main objective in PPDM is to develop algorithms for modifying the original data in some way, so that the private data and knowledge remain private even after the mining process [6].

In this paper, we focus on privacy preserving data clustering, notably when personal data are shared before clustering analysis. The idea is to partition a dataset into new clusters of similar objects. The goal is to group objects to achieve high similarity between objects within individual clusters and low similarity between objects that belong to different clusters.

To address privacy concerns in clustering analysis, we need to design specific data transformation methods that enforce privacy without loosing the benefit of mining. The key idea is to convert the categorical attribute value into binary value and it is transformed using geometric data transformation method. Then, clustering using conventional clustering algorithm is done on the transformed data to ensure privacy.

2. PRIVACY PRESERVING TECHNIQUES

There are many approaches adopted for privacy preserving data mining [4]. It can be classified based on the following dimensions:

- Data distribution
- Data modification
- Data mining algorithm
- Data or rule hiding
- Privacy preservation

The following sections discuss the approaches involved in each dimension.

T. Sobh (ed.), Innovations and Advanced Techniques in Computer and Information Sciences and Engineering, 403–408.

2.1 Data distribution

Some of the approaches have been developed for centralized data, while others refer to a distributed data scenario. Distributed data scenarios can also be classified as horizontal data distribution and vertical data distribution. Horizontal distribution refers to these cases where different database records reside in different places, while vertical data distribution, refers to the cases where all the values for different attributes reside in different places.

2.2 Data modification

In general, data modification is used in order to modify the original values of a database that needs to be released to the public and in this way to ensure high privacy protection [4]. It is important that a data modification technique should be in concert with the privacy policy adopted by an organization. Methods of modification include:

•*Perturbation***,** which is done by the alteration of an attribute value by a new value (i.e., changing a 1-value to a 0-value, or adding noise)

•*Blocking*, which is the replacement of an existing attribute value with a "?"

• *Aggregation* or merging which is the combination of several values into a coarser category

• *Swapping* that refers to interchanging values of individual records, and

• *Sampling* which refers to releasing data for only a sample of a population.

2.3 Data mining algorithm

This is actually something that is not known beforehand, but it facilitates the analysis and design of the data hiding algorithm. Various data mining algorithms have been considered in isolation of each other. Among them, the most important ideas have been developed for classification data mining algorithms, like decision tree inducers, association rule mining algorithms, clustering algorithms, rough sets and Bayesian networks.

2.4 Data or rule hiding

It refers to whether raw data or aggregated data should be hidden. The complexity for hiding aggregated data in the form of rules is of course higher, and for this reason, mostly heuristics have been developed. The lessening of the amount of public information causes the data miner to produce weaker inference rules that will not allow the inference of confidential values. This process is also known as "rule confusion".

2.5 Privacy preservation

This refers to the privacy preservation technique used for the selective modification of the data [4]. Selective modification is required in order to achieve higher utility for the modified data given that the privacy is not jeopardized. The techniques that have been applied for this reason are:

• Heuristic-based techniques like adaptive modification that modifies only selected values that minimize the utility loss rather than all available values

• Cryptography-based techniques like secure multiparty computation where a computation is secure if at the end of the computation, no party knows anything except its own input and the results, and

• Reconstruction-based techniques where the original distribution of the data is reconstructed from the randomized data. It is important to realize that data modification results in degradation of the database performance. In the proposed system, we use the data modification approach for privacy preservation.

3. Cluster Analysis

Clustering is an important data mining problem. The goal of clustering, in general, is to discover dense and sparse regions in a dataset. Most previous work in clustering focused on numerical data whose inherent geometric properties [2] can be exploited to naturally define distance functions between points. However, many datasets also consist of categorical attributes on which distance functions are not naturally defined. Recently, the problem of clustering categorical data started receiving interest.

3.1 Categorical Variables

Categorical variable (nominal variable) is a variable which can take more than two states and the domain of the categorical attribute is small [2]. For example, marital status is a categorical variable that may have, say three states: single, married, divorcee.

Let the number of states of the variable be M. The states can be denoted by letters or symbols. The dissimilarity between two objects i and j, defined by nominal variables can be computed using the simple matching approach:

$$d(i, j)=(p-m)/p$$

Where m is the number of matches (i.e., the number of variables for which I and j are in the same state), and p is the total number of variables.