Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 54 (2015) 422 – 430

Procedia Computer Science

Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015)

A Privacy Preserved Data Mining Approach Based on k-Partite Graph Theory

T. Pranav Bhat, C. Karthik* and K. Chandrasekaran

Department of Computer Science and Engineering, NITK Surathkal 575 025, Karnataka, India

Abstract

Traditional approaches to data mining may perform well on extraction of information necessary to build a classification rule useful for further categorisation in supervised classification learning problems. However most of the approaches require fail to hide the identity of the subject to whom the data pertains to, and this can cause a big privacy breach. This document addresses this issue by the use of a graph theoretical approach based on k-partitioning of graphs, which paves way to creation of a complex decision tree classifier, organised in a prioritised hierarchy. Experimental results and analytical treatment to justify the correctness of the approach are also included.

Peer-review under responsibility of organizing committee of the Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015)

Keywords: Data mining; Graph theory; K-partite; Privacy; Security.

1. Introduction

Information extraction from a given data-repository to determine the behaviour of a particular system, or to determine the predictive outcome of a particular problem statement for the case of an unknown condition or input forms an application of wide horizon in easing the life of human beings, by its penetration into domains ranging from e-commerce to healthcare. Supervised learning algorithms have been widely employed in prediction problems to forecast the outcome of a tweaked problem from an underlying data reflecting the actual outcomes of similar problems.

A major concern that arises out of the above techniques of data repositories for data mining using supervised learning techniques for building of classification rules is the privacy and confidentiality of the information, especially in guarding the identity of the subjects to whom the information pertains to. Various privacy issues could arise due of the mining of such sensitive personal data, and misuse of the data by breach of privacy can cause legal and ethical issues beyond the domain of data mining

Privacy Preserved Data Mining is a new hype which has entered the market and which claims to take care of this particular issue. The goal of privacy preserving data mining is to develop data mining methods without increasing the risk of misuse of the data used to generate those methods. Literature cites a large number of methods, most of which use some form of transformation on the original data to ensure privacy preservation, called key interchange mapping methods, but these methods are quite complex and compute and memory intensive, thus leading to limited

*Corresponding author. Tel.: +91 9035219859.
E-mail address: karthikiyer2000@gmail.com

Peer-review under responsibility of organizing committee of the Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015)
doi:10.1016/j.procs.2015.06.049

usage of these methods. This document suggests an alternative approach to privacy preservation. This method leads to identification of two categories of attributes – *key – attributes*, which directly reveal the identity of an individual with minimal or single operations on them and *quasi – identifier – attributes*, which identify an individual through certain data mining operations, primarily due to the existence of attribute dependencies. The proposed technique harnesses these vulnerabilities in privacy-unpreserved dataset and strives to eliminate these, by using simple principles from the theory of *k-partite graphs*, and builds classifiers from these privacy preserved data-subsets, which then will be grouped based on decision tree root priority approaches, to form a privacy preserving complex classifier or classification rule for test sets.

We organise the paper through 7 sections in the following fashion. Section 2 discusses related work and earlier contributions cited in literature in the domain of privacy preserved data mining. Section 3 details the methodology proposed with the required analytical justifications wherever necessary. Section 4 illustrates the approach and justifies it by the use of an experimental setup, and also analysis of results, followed by conclusion in Section 6.

2. Literature Survey

A lot of work has gone into tackling the issues of data related security. One of the recent issues is the privacy preservation of users and individuals while mining through data. The work by Agarwal and Srikant[1] on PPDM is one of the initial works to address this issue. In their paper they have built a decision tree using training data whose distribution was scattered and still obtained comparable classification accuracy results. In[2,3] the authors have given a detailed description about the *k* anonymization and randomization techniques of PPDM and also addressed the issues and the areas of application for PPDM. In[4] a detailed study has been given of topics such as attribute relations, use of technology for privacy enhancement. They have done this through a survey of data mining related privacy for two methods-randomization and secure multiparty computation. In[5] the authors have proposed a two tier method by which the medical data can be safely mined with increased privacy. The two tiers are horizontal data separation and vertical data separation. A similar work was done in[6] where the authors decided the level of anonymization of attributes based on their sensitivity. In[7] the authors have proposed an enhancement of the *k* anonymity method for privacy preservation.

3. Methodology

The principle of enabling privacy preservation in a dataset under use mainly concerns with the identification of the vulnerabilities or faults in the existing data-mining methodology, or in the existing set of steps involved in the information retrieval process using data mining techniques based on supervised learning concepts. More specifically privacy breach and information misuse can be avoided by eliminating direct or indirect extraction of information pertaining to the subjects of the particular information, especially when the information is quite sensitive as in the cases of healthcare data. This drives home the idea that identity related attributes need to be eliminated or anonymised for privacy preservation, which means that the key attributes can be removed and the quasi-attributes can be played around with, which is the major principle driving privacy preservation in the proposed approach.

The initial cleansing and formatting process is performed on the target data, which is followed by the graph theoretical privacy preservation proposed to generate privacy preserving sub-classifiers, which are then integrated in a decision tree root identification hierarchy methodology proposed by Quinlan *et al.*[8], followed by the classification. This forms the complex privacy preserved data mining setup.

3.1 Assumptions

1. Applicability to supervised learning approaches by the existence of a labelled training dataset.
2. Problem is a binary classification problem (or also called a *concept*), where the output is either true or false (this condition can be relaxed since the method works for multi-class classifications as well.
3. Data stored as a relational database. Data stored in the form of semi-structured data, in the form of XML sheets or NO-SQL databases may need to be transformed into Relational databases and then this method applied.