Journal of King Saud University – Computer and Information Sciences (2015) 27, 297–304

King Saud University

Journal of King Saud University – Computer and Information Sciences

www.ksu.edu.sa
www.sciencedirect.com

An anonymization technique using intersected decision trees

Sam Fletcher *, Md Zahidul Islam

Center for Research in Complex Systems (CRiCS), School of Computing and Mathematics, Charles Sturt University, Bathurst, NSW 2795, Australia

Received 22 February 2014; revised 25 April 2014; accepted 4 June 2014
Available online 19 June 2015

KEYWORDS

Privacy preserving data mining;
Decision tree;
Anonymization;
Data mining;
Data quality

Abstract Data mining plays an important role in analyzing the massive amount of data collected in today's world. However, due to the public's rising awareness of privacy and lack of trust in organizations, suitable Privacy Preserving Data Mining (PPDM) techniques have become vital. A PPDM technique provides individual privacy while allowing useful data mining. We present a novel noise addition technique called Forest Framework, two novel data quality evaluation techniques called EDUDS and EDUSC, and a security evaluation technique called SERS. Forest Framework builds a decision forest from a dataset and preserves all the patterns (logic rules) of the forest while adding noise to the dataset. We compare Forest Framework to its predecessor, Framework, and another established technique, GADP. Our comparison is done using our three evaluation criteria, as well as Prediction Accuracy. Our experimental results demonstrate the success of our proposed extensions to Framework and the usefulness of our evaluation criteria.

1. Introduction

As technology has advanced, so has the ability to gather massive amounts of data. Data Mining plays an important role in data collection, pre-processing, integration and pattern extraction from the collected data. Due to the wide use of data mining, it is important to consider the ramifications. The most prominent ramification is perhaps the breach of individual privacy. This public awareness of privacy and lack of trust in organizations (Arnott, 2011) may introduce additional complexity to data collection. As a result, organizations may not be able to fully enjoy the benefits of data mining. Privacy Preserving Data Mining (PPDM) techniques have therefore become vital. A PPDM technique provides individual privacy while allowing useful data mining on a dataset. Typical PPDM techniques include noise addition to a dataset, data swapping, aggregation and masking (Brankovic et al., 2007; Dankar and Eman, 2012; Adam and Worthmann, 1989; Farkas and Jajodia, 2002). The two main aims of the PPDM techniques are high security and high data quality/utility.

In this paper, we propose a novel technique called Forest Framework as a modification of an existing technique called Framework (Islam and Brankovic, 2011). Forest Framework

* Corresponding author.
E-mail addresses: safletcher@csu.edu.au (S. Fletcher), zislam@csu.edu.au (M.Z. Islam).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

http://dx.doi.org/10.1016/j.jksuci.2014.06.015

adds noise to a dataset in such a way that many existing patterns of the unperturbed dataset are preserved many more than is possible with Framework. We also propose two novel data quality evaluation techniques called "Evaluation of Data Utility using Domain Similarity" (EDUDS) and "Evaluation of Data Utility using Splitting Criteria" (EDUSC). Additionally, we propose a security analysis technique called "Security Evaluation using Record Similarity" (SERS). We carry out experiments on five natural datasets available from UCI Machine Learning Repository (Bache and Lichman, 2013). Our experimental results indicate the effectiveness of all our techniques. The organization of the paper is as follows: Section 2 provides a relevant background study; Section 3 presents Forest Framework; Section 4 presents EDUDS; Section 5 presents EDUSC; Section 6 presents SERS; and experimental results are presented in Section 7. Section 8 gives concluding remarks and avenues for future research.

We consider a dataset as a two dimensional table where rows represent the records and columns represent attributes. Each attribute can be numerical or categorical. Out of the attributes, one is a class attribute. Fig. 1 shows an example decision forest having two trees T_1 and T_2 obtained from a dataset. Decision tree algorithms iteratively discover which attribute best explains the class attribute for the given segment of records defined by the preceding attribute splits (e.g. $A > 7$ in Fig. 1) (Quinlan, 1993; Quinlan, 1996). In this example, each of the trees has three leaves L_1, L_2, and L_3. The leaves of a tree divide the dataset into mutually exclusive horizontal segments of records.

2. Background study

There are many privacy preserving data mining techniques in the literature, ranging from output privacy (Wang and Liu, 2011) to categorical noise addition (Giggins, 2012) to differential privacy (Friedman and Schuster, 2010), to many others discussed in surveys (Brankovic et al., 2007; Adam and Worthmann, 1989; Farkas and Jajodia, 2002; Wu et al., 2010). Framework – one such privacy preservation technique – was proposed in 2011 (Islam and Brankovic, 2011). It first builds a decision tree from an original dataset in order to learn the existing patterns (logic rules) of the dataset. It then adds noise to all attributes (both numerical and categorical) of a dataset in a way that preserves patterns discovered by the decision tree built from the original dataset. The basic idea of Framework is to add noise to the value of a numerical attribute of a record, but in such a way that the perturbed value falls within the range that satisfies the logic rule of the leaf where the record originally belongs to. That is, if a record in an unperturbed dataset falls in a leaf of the tree obtained from the original dataset, then Framework adds noise in such a way that the record still falls in the same leaf of the tree even after noise addition.

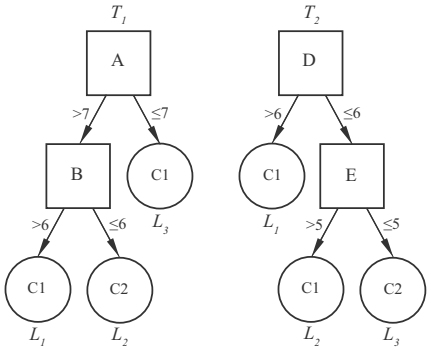

Fig. 1 An example decision forest with two trees.

For a categorical attribute, it first discovers which values are similar. Using a user-defined probability it then changes each value to another value having high similarity with the original value. For a class attribute, it shuffles the class values of the records belonging to the same leaf in such a way so that the distribution of class values among the records remains the same.

Two main aims of noise addition techniques are to perturb a dataset in order to preserve individual privacy and maintain high utility in the perturbed dataset. Measuring data utility is a challenging task (Ntoutsi et al., 2008; Osei-Bryson, 2004). Generally, the quality of a perturbed dataset is measured through the Prediction Accuracy of a decision tree, built from the perturbed dataset, while it is applied on an unperturbed testing dataset (Islam and Brankovic, 2011; Ray et al., 2011). It has also been shown that an assessment of data quality provided by a comparison of Prediction Accuracy may differ from an assessment provided by a comparison of decision tree similarity (Islam, 2007; Lim et al., 2000). The utility of a perturbed dataset is sometimes evaluated through the similarity of decision trees, the accuracy of the trees, and statistical properties such as mean and correlation matrix (Islam and Brankovic, 2011). Finding a suitable technique to compare the similarity of two trees can be a challenge.

General Additive Data Perturbation (GADP) perturbs only those attributes which are deemed confidential by a user, thus allowing data quality to remain as high as possible for the non-confidential attributes (Muralidhar et al., 1999). However, it takes all attributes into account when perturbing confidential attributes, and thus maintains all correlations among the attributes of a dataset. Modifications of GADP such as CGADP and EGADP have been proposed in order to preserve statistical parameters in datasets having non-multivariate normal distribution or small number of records (Sarathy et al., 2002; Muralidhar and Sarathy, 2005).

We will also be using the same benchmark perturbation technique used in the original paper: Random Technique (RT) (Islam and Brankovic, 2011). RT is a random noise addition technique which is used as a means for evaluating the effectiveness of other perturbation and evaluation techniques. It adds uniform noise to all attributes indiscriminately.

3. Our perturbation technique: forest framework

Forest Framework is a modification of an existing technique called Framework (Islam and Brankovic, 2011), with an aim to better preserve the original data quality in a perturbed dataset. Unlike Framework, it first builds a decision forest (Islam and Giggins, 2011) from an unperturbed dataset in order to learn the existing patterns (logic rules) of the dataset. The construction of a forest allows for far more patterns to be discovered, and therefore preserved. In our experiments we will be using the SysFor forest-building algorithm (Islam and Giggins, 2011), which harnesses the popular C4.5 tree building algorithm (Quinlan, 1993; Quinlan, 1996). Forest Framework