World Engineering & Applied Sciences Journal 6 (1): 41-44, 2015
ISSN 2079-2204

DOI: 10.5829/idosi.weasj.2015.6.1.22187

A Cryptography Based Privacy Preserving Association Rule Mining in Academic Analytics

J. Shana

Department of MCA, Coimbatore Institute of Technology, Coimbatore-641014, TamilNadu, India

Abstract: Association rule mining is one the important data mining technique that discovers the association among attributes and reveals hidden patterns in the given data source. The idea of outsourcing data analytics is becoming increasingly popular; the associated security risks still prevent many potential users from deploying it. In particular, the need to give full access to one's data to a third party, the database service provider, remains a major obstacle. There is one area of data mining called privacy preserving association rule mining that attempts to protect either the database or the discovered rules. This paper attempts the latter by using cryptographic algorithm to protect the rules generated in educational data so that it cannot be identified by the consultant. In academic analytics the nature of data is different compared to a business domain. This work highlights a simple model that implements AES encryption algorithm on target attributes before applying association rule mining algorithm. Through experiments conducted on student course data it is shown that the proposed system is an easy and reliable system to preserve privacy of educational data.

Key words: Privacy preserving association rule mining · Data mining · Academic analytics

INTRODUCTION

Academic analytics is the term for business intelligence used in an academic setting. There is an increasing distinction made between academic analytics and traditional BI because of the unique type of information that university administrators require for decision making. The goal here is to identify association rules to study the factors that influence the result of students in the course. For any educational institution the data representing a student is confidential. And when association rules are generated it reveals many useful patterns about admission, student and staff performance, student retentivity and so on [1]. With huge number of institutions across the country competing for student intake it becomes very important to identify factors that would enable the decision makers to take productive measures to benefit the students and the institution. It is also equally necessary to preserve the privacy of the knowledge mined such as associations generated among the target attributes. This discovered knowledge needs to be protected from the competitors. It is not a problem if the data is stored in-house and analytics also performed in-house. But most of the College Management Systems are being managed by consultants (third parties). But it is equally important for the institution to protect the rules from being exposed. Cryptographic techniques have been used for data security since ages. In this paper it is being used in a educational data analysis. This paper uses dataset of undergraduate students of computer technology for a particular course. This paper has the following sections, Section II gives the background information to understand the basics of privacy preserving association rule mining. Section III is the literature survey on existing and related work. Section IV highlights the architectural components of the proposed system. Section V shows the experimental results and Section VI concludes the work.

The Background

Association Rule Mining: Association rules are if-then statements that help uncover relationships between seemingly unrelated data in a relational database or other information repository. An example of an association rule would be "If a student is from English medium school and attendance >80% then his result is a pass"

Association rules are created by analyzing data for frequent if-then patterns and using the criteria *support* and *confidence* to identify the most important relationships [1]. *Support* is an indication of how

Corresponding Author: J. Shana, Department of MCA, Coimbatore Institute of Technology, Coimbatore-641014, TamilNadu, India.

frequently the items appear in the database. *Confidence* indicates the number of times the if-then statements have been found to be true. For an association rule of form X=>Y the support and confidence is specified as given by (1) and (2) [2].

$$Support(X) = \frac{Support\_count(X)}{n} \times 100 \quad (1)$$

$$Confidence(X \rightarrow Y) = \frac{Support(XY)}{Support(X)} \times 100 \quad (2)$$

Privacy Preserving Association Rule Mining: Given a database D, it needs to be transformed into D' by modifying the database by applying any privacy protection methods. Then association rules are mined from D' such that all the rules in D are intact. Based on the privacy protection technologies used privacy preserving association rule mining algorithms can be classified into data perturbation based techniques and cryptography based methods. Data perturbation techniques are statistical methods that seek to protect confidential data by adding random noise to confidential, numerical attributes, thereby protecting the original data [3]. A primary perturbation technique is data swapping-exchanging data values between records in ways that preserve certain statistics but destroy real values. An alternative is randomization-adding noise to data to prevent discovery of the real values. Because the data no longer reflects real world values, it can't be used to violate individual privacy. Another method is the use of cryptographic techniques to preserve privacy of data. This paper concentrates on this kind of technique. The challenge is in obtaining valid data mining results from the perturbed data. The aim of privacy in association rule mining is (a) Inability to extract any meaningful information from the given encrypted dataset and (b) Reliability in generated rules.

Related Work: Literature shows that much work have been done in privacy preserving data mining and association rule mining also have its share. [4]. used data distortion techniques to modify the confidential data values so that the approximate original data distribution could be obtained from the modified version of the database.

The authors in [5] have worked on data distortion algorithms aimed towards hiding association rules with efficiency and side effects as metrics. These algorithms were the first of their kind in hiding association rules but had large number of side effects. [6] Proposes a set of efficient algorithms for hiding sensitive knowledge from data by minimally perturbing their values. The hiding strategies proposed are based on reducing the support and confidence of rules.

A number of recently proposed techniques address the issue of privacy preservation by perturbing the data and reconstructing the distributions at an aggregate level in order to perform the association rule mining. The authors in [7] use the Bayesian algorithm for distribution reconstruction in numerical data. Also for categorical data [4] suggests a uniform randomization approach on reconstruction. The authors of [8] improved the work over the Bayesian based reconstruction procedure by using an EM algorithm for distribution reconstruction.

Another variation of data reconstruction is based upon sanitizing the knowledge base and [9] first proposed a Constraint based Inverse Itemset Lattice Mining Procedure (CIILM) for hiding sensitive frequent itemsets. Their data reconstruction is based on itemset lattice.

A set of algorithms for distributed privacy-preserving data mining is discussed in [10] which propose new data mining primitives for secure multi-party computation over a variety of horizontally and vertically partitioned data sets. [11] Implements a model incorporating cryptographic techniques to minimize the shared information without incurring much overhead in horizontally partitioned distributed data mining. But the cost of mining is much higher [12].

The work [13] utilizes a one-to-n item mapping together with nondeterministic addition of cipher items to protect the identification of individual items.

A Simple Crypto Association Rule Mining: In this paper a simple method to preserve the privacy of association rules has been introduced. This approach is known as the Alice-to-Alice cryptosystem and has been proposed in [14]. The essence of this approach lies in the fact that the proprietor of the database, who in the context of cryptography is called Alice, encrypts the attribute names in the database. The encrypted data is then transferred to the data miner (consultant) who extracts the set of association rules through available data mining techniques without being able to obtain insight into the meaning of either the data, or the rules. Finally, the association rules are returned to Alice who by decrypting them obtains the true meaning of the extracted rules. The two main characteristics that identify and differentiate

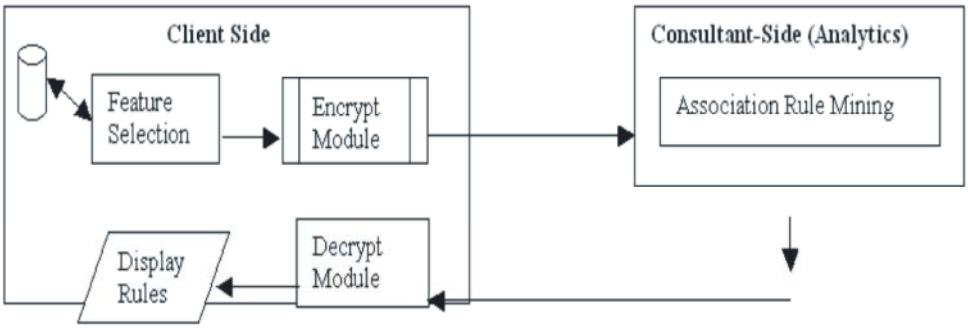

Fig 1: Architecture of the proposed system

one encryption algorithm from another are its ability to secure the protected data against attacks and its speed and efficiency in doing so. Any good encryption scheme can be used. The client side components of the proposed system are shown in Fig.1. For this approach to be reliable, the resulting rules should correspond to the rules that would be obtained had data mining been applied on the real data.

Experimental Results: The prototype of the proposed system is implemented as a part of data analytics in College Management System. The dataset taken consists of 1000 instances of UG student information for a particular course of study along with their personal information gathered through surveys. The emphasis here is to derive association rules to identify the factors that influence the performance of students. The details of the dataset are given in the Table 1.

The certain attributes which do not affect the performance of the students are removed from the target list of attributes. The selected sensitive attribute names are encrypted by the encryption module using the algorithm in Section II. The attribute values are nominal values and are not encrypted. The Blowfish encryption algorithm which has been proved to give considerable performance in [12] is used here. The implementation is in Java and association rules are generated by implementing the well known Apriority algorithm to generate the frequent item sets. The algorithms were executed on Intel core i3 processor machine with 4GB RAM. Table 2 shows the result of the original attributes after encryption.

Table 1: Dataset details

Dataset	# attributes	Instances
Course Data	14	1000

Table 2: Attributes before and after encryption

Attr #	Original	Encrypted
1	Family_ Income	g9X0NGXaKlPI9xSs9EQX6A==
2	Native	iBP5t26rFbfIzAz8thoZcQ==
3	SSLC_Perc	cZn0wUOcc2BtxR2yVym4Ag==
4	Hsc_Perc	17Qx6NqCXTi8BWg83kXb2w==
6	Medium	flGdON0KjnXcgRY8pI0h8w==
7	Prev_Skill	qafklYVORicBdoCQeiOllQ==
8	Motivation	bm3trqPWwiSVMinul2okpw==
9	Stay	g6H51id8yvsfQOT4kF0Npw==
10	Staff_Approach	he/duwvc8ig//j115sJkwg==
11	Sub_Difficulty	DVIrDmxhlXRgy8nq1e79Og==
12	Friends_Circle	xB1NaxGJWwht38LaXLYR1Q==
13	Sub_Interest	/nsppBM1Dcwcc0Xr7I0rHg==
14	Result	aJ5BPPpRnImJsj3/N8LvuQ==

Association rules were generated with a minimum support count of 90 % and confidence as 95%. Table 3 shows the sample rules generated after encrypting the target attributes.

The Table 3 shows the kind of association rules generated after the attributes are encrypted. The cipher text hides the association pattern generated from the dataset. The study shows that the encrypted rules were the same as the ones generated with the original non-encrypted attributes thus ensuring reliability which is one of the important privacy metric as described in Section II. Also there was no time delay in generating the rules after the attributes were encrypted as shown in Table 4.

Table 3: Sample Association Rules Generated

#1	g9X0NGXaKlPI9xSs9EQX6A=high bm3trqPWwiSVMinul2okpw=high g6H51id8yvsfQOT4kF0Npw=1 _ aJ5BPPpRnImJsj3N8LvuQ=1 conf:(0.95)
#2	g9X0NGXaKlPI9xSs9EQX6A=high flGdON0KjnXcgRY8pI0h8w=1 bm3trqPWwiSVMinul2okpw=high g6H51id8yvsfQOT4kF0Npw=1_aJ5BPPpRnImJsj3N8LvuQ=1 conf:(0.95)

Table 4: Rule generation time

	Time for 10 rules (seconds)	
Attributes#	Encrypted rule generation	Normal Rules
13	20:47:51	20:47:51
30	20:59:45	20:59:45
50	21:03:11	21:03:11

CONCLUSION

The experiments done shows a method to protect the association rules generated from a data source. The time taken by the encryption/decryption module is insignificant when the number of attributes is scaled to 50. Such a framework is cost-effective for any educational institution intending to do educational data mining. Future work can concentrate on cryptography based predictive mining and other knowledge extraction techniques in educational data mining. Also further study needs to be done on the effect of overall performance of the system when the number of attributes increases considerably to suit other domains.

REFERENCES

1. Han, J. and M. Kamber, 2009. Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers.
2. Boutsinas, B., G.C. Meletiou and M.N. Vrahatis, 2002. Mining Encrypted Data, In Proc. of the International Conference on Financial Engineering, E-commerce and Supply Chain and Strategies of Development.
3. Wilson, R.L. and Peter A. Rosen, 2003. Protecting Data Through Perturbation Techniques: The Impact on Knowledge Discovery in Database, 14(2): 14-26.
4. Evfimievski, A., R. Srikant, R. Agarwal and J. Gehrke, 2002. Privacy Preserving Mining of Association Rules, SIGKDD.
5. Oliveria, S.R.M. and O.R. Zaiane, 2004. Privacy Preserving Frequent Itemset Mining, In Proc. of IEEE ICDM Wokshop on Privacy, Security and Data Mining, pp: 43-54.
6. Verykios, V.S., A. Elmagarmid, E. Bertino, Y. Saygin and E. Dasseni, 2004. Association Rule Hiding, IEEE Transactions on Knowledge and Data Engineering, 16(4): 434-447.
7. Agrawal, R. and R. Srikant, 2000. Privacy Preserving Data Mining, In Proc. of ACM SIGMOD Conf on Management Data, pp: 439-450.
8. Agrawal, D. and C. Agrawal, 2001. On the Design and Quantification of Privacy Preserving Data Mining Algorithms, Proc. of the 20th Symposium on Principles of Database Systems, USA.
9. Chen, X., M. Orlowska and X. Li, 2004. A New Framework for Privacy Preserving Data sharing, In Proc. of the 4th IEEE ICDM Workshop: Privacy and Security Aspects of Data Mining, IEEE Computer Society, pp: 47-56.
10. Cliffton, C., M. Kanatarcioglou, X. Lin and M.Y. Zhu, 2002. Tools for Privacy Preserving Distributed Data Mining, SIGMOD Explorations, 4(2).
11. Kanatarcioglou, M. and C. Cliffton, 2002. Privacy Preserving Distributed Mining of Association Rules on Horizontally Partitioned Data, ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery(DMKD'02), pp: 24-31.
12. Verma, O.P., R. Agarwal, D. Dafouli and S. Tyagi, 2011. Performance Analysis of Data Encryption Algorithm, International Conference on Electronics Computer Technology (ICECT), pp: 366-372.
13. Agrawal, R. and R. Srikant, 1994. Fast algorithms for mining association rules, in Proc. Int. Conf. Very Large Data Bases, pp: 487-499.
14. Baker, R.S.J.D. and K. Yacef, 2009. The state of educational data mining in 2009: A review and future visions, Journal of Educational Data Mining, pp: 3-17.