DATA MINING AND CRYPTOLOGY

E.C. LASKARI[1,3], G.C. MELETIOU[2,3], D.K. TASOULIS[1,3],
M.N. VRAHATIS[1,3]

[1] *Department of Mathematics, University of Patras, GR-26110 Patras, Greece,
E-mail: {elena, dtas, vrahatis}@math.upatras.gr*

[2] *A.T.E.I. of Epirus, P.O. Box 110, GR-47100 Arta, Greece,
E-mail: gmelet@teiep.gr*

[3] *University of Patras Artificial Intelligence Research Center (UPAIRC),
University of Patras, GR-26110 Patras, Greece*

This paper addresses the issue of mining encrypted data, in order to protect confidential information while permitting knowledge discovery. Common cryptographic algorithms are considered and their robustness against data mining algorithms is evaluated. Having identified robust cryptosystems, data are encrypted and well-known data mining techniques are applied on the encrypted data to produce classification rules which are then compared with those obtained from the initial non-encrypted databases.

1. Introduction

Nowadays business and scientific organizations collect data, orders of magnitude greater than ever before. Considerable attention has been paid in the development of methods that contribute to knowledge discovery in large databases, using data mining techniques, for example see Fayyad et al. in [7].

Data mining can be used either to classify data into predefined classes (classification), or to partition a set of patterns into disjoint and homogeneous groups (clustering), or to identify frequent patterns in the data, in the form of dependencies among concepts-attributes (associations).

Business and scientific databases typically contain confidential information. Clifton and Marks in [3] provide examples in which applying data mining algorithms on a firm's database reveals critical information to business rivals. Clifton in [4] presents a technique to prevent the disclosure of confidential information by releasing only samples of the original data. This

technique is applicable independently of the specific data mining algorithm to be used. In later work, Clifton [5] has proposed ways through which distributed data mining techniques can be applied on the union of databases of business competitors so as to extract association rules, without violating the confidentiality of the data of each firm. This problem is also addressed by Lindell and Pincas [8] for the case when classification rules are to be extracted. Another approach to extract association rules without violating privacy is to artificially decrease the significance of these rules (see [1,6]).

Here, we also consider a scenario in which a company or a scientific organization negotiates a deal with a consultant. We address the privacy problem by encrypting the data. Thus the miner will be unable to extract meaningful information neither from the raw data, nor from the extracted rules. Having applied the data mining algorithms, the consultant provides the organization with the extracted rules. Finally, the organization decrypts those rules so as to restore their true meaning. Two important issues arise in this approach. The first is the need to investigate the robustness of commonly used cryptosystems to potential attacks by data miners. The second key issue is to choose the appropriate cryptosystem that will allow the deduction of correct results, in the sense that the decrypted rules have to be as close as possible to the rules that would be extracted from the original data.

2. A new approach to cryptanalysis

In the context of cryptography, the encryption algorithm is called *cryptosystem* or *cipher*, its input is called *plaintext* and its output is the *ciphertext*. The elementary requirement for a cryptosystem to be considered secure is that it is computationally infeasible for an eavesdropper who obtains the ciphertext to deduce any portion of the plaintext. Cryptanalysis is the study of mathematical techniques to violate cryptographic systems.

Frequency analysis is the first step undertaken by cryptanalysts. The idea of using the underlying frequency distribution of a language to decipher an encrypted message dates back to the early 15th century, and it is attributed to an Arab mathematician named Qalqashandi.

Since, data mining algorithms are able to identify regularities in data in the form of dependencies, we primarily consider the application of data mining techniques for the purposes of cryptanalysis as a generalization of traditional frequency analysis. To this end, we study the robustness of commonly used cryptosystems to alternative data mining algorithms. This

knowledge is critical for the identification of the proper cryptosystem that will be incorporated in the Alice-to-Alice cryptography, described immediately below.

3. Alice to Alice cryptography

This section addresses the key problem of privacy of data mining on encrypted data. This approach is known as the Alice-to-Alice cryptography, and has been recently proposed in [2]. 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 database. The encrypted data is then transferred to the data miner who extracts the set of classification rules through available data mining techniques without being able to obtain insight into the meaning of either the data, or the rules. Finally, the classification rules are returned to Alice who by decrypting them obtains the true meaning of the extracted rules. Clearly, 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.

The main acting agents of the protocol are, "Alice" that represents a business or scientific organization and "Bob" that represents a data mining consultant who handles the data mining process. Alice owns a database with fields and field values that correspond to attributes and attribute values referred by the data mining rules. Attribute values, irrespective of what they represent, have to be encrypted. Since each attribute value has a label like "good customer" or "driver" or "tomato", this label can be transformed to an integer. For instance, a label can be transformed to a string of bits with the help of ASCII code, in turn, each string corresponds to a number (integer). As a result, in both of the above cases each attribute value can be represented as a small integer. The methodology is as follows:

encrypted data mining algorithm

(1) Alice collects the data organized into relational tables
(2) Alice encrypts the relational tables
(3) Alice sends the encrypted tables to the miner.
    Data mining performed.
    The miner returns the obtained rules to Alice
(4) Alice decrypts the rules.

During the first step, Alice selects and preprocesses the appropriate

data and organizes it into relational tables. A relational table is supposed to be two dimensional, however it can be represented as one dimensional considering it in a *row major order* or in a *column major order*.

At the second step, encryption takes place. At the third step, Alice sends the encrypted tables to Bob. Bob applies the proper data mining algorithm to the encrypted tables and a number of data mining rules are extracted. Of course, attribute and attribute values appearing in the rules are encrypted. Then Bob returns these rules to Alice. Finally, Alice decrypts the rules.

The scope of the present study is to extent this line of research by identifying the proper cryptographic methods and data mining techniques so as to meet the objectives of privacy, i.e. the inability of the data miner to extract any meaningful information from the encrypted data he receives; and reliability, in the sense that the encrypted rules once decrypted will correspond as closely as possible to the ones that would be obtained had the data been processed in its original form. To this end, the process will be applied on extensively studied databases.

References

1. M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, V. Verykios, Disclosure Limitation of Sensitive Rules, Proceedings of the *1999 Workshop on Knowledge and Data Engineering Exchange*, Chicago, 45–52, (1999).
2. B. Boutsinas, G.C. Meletiou, M.N. Vrahatis, *Mining Encrypted Data*, Proceedings of the International Conference on *Financial Engineering, E-commerce & Supply Chain, and Strategies of Development, (FEES 2002)*, June 10–12, 2002, Athens, Greece, in press.
3. C. Clifton, D. Marks, Security and Privacy Implication of Data Mining, Proceedings of the *1996 ACM Workshop on Data Mining and Knowledge Discovery*, (1996).
4. C. Clifton, Protecting against Data Mining through Samples, Proceedings of the *13th IFIP Conference on Database Security*, Seattle, Washington, (1999).
5. C. Clifton, *Privacy Preserving Distributed Data Mining*, (2001).
6. E. Dasseni, V. Verykios, A. Elmagarmid, E. Bertino, Hiding Association Rules by Using Confidence and Support, *LNCS 2137*, 369-383, (2001).
7. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, Advances in Knowledge Discovery and Data Mining, *AAAI*, Press/MIT Press (1996).
8. Y. Lindell, B. Pinkas, Privacy Preserving Data Mining, *Advances in Cryptology - CRYPTO '00, LNCS 1880*, 36-53, (2000).