The Use of Fully Homomorphic Encryption in Data Mining with Privacy Preserving

A. Laécio A. Costa¹, B. Ruy J. G. B. de Queiroz²

¹College of Informatics, Instituto Federal do Sertão Pernambucano - IFSertão, Petrolina, Pernambuco, Brazil
²Center of Informatics, Universidade Federal do Pernambuco - UFPE, Recife, Pernambuco, Brazil

Abstract - Data mining is a computational tool widely used today which aims to extract useful information from various databases. Nowadays with the large volume of information that is produced, stored in a remote database (using cloud computing), concerns about confidentiality and privacy of information are arising due to lack of guaranteed security by storage service and the mining algorithm. A new approach of modern cryptography, defined as the Fully Homomorphic Encryption (FHE), allows for the encrypted data to be arbitrarily computed which is a solution that aims to preserve the security, confidentiality and data privacy. This article aims to present a study of the literature to identify research that proposes methods that ensure the confidentiality and privacy in the mining of databases based on fully homomorphic encryption.

Keywords: Data Mining, Privacy Preserving, Fully Homomorphic Encryption

1 Introduction

Many companies need to explore sensitive data derived from data mining of multiple databases in order to obtain useful information. To preserve the integrity and privacy in this exploration, they need to make the ciphertext without disclosing or having knowledge of the content of the data operations. For example, two competing companies need to exchange sensitive information, this information is important for assessing a given market scenario, so each company "X" and "Y" encodes and forwards its information to a platform of homomorphic processing. The processing platform performs operations with the encrypted data of the two companies without knowledge of the original message and returns the result to the two companies "X" and "Y" for analysis. At no time did either company have access to the data of the competitor, but managed to make a more accurate analysis of the market using the output data. Therefore the proposed homomorphic encryption aims to ensure this scenario is safe as well as practical.

The homomorphic encryption is an area of modern cryptography, enabling the completion of computation of arbitrary computations on a ciphertext and still achieves the encrypted result that corresponds to the sequence of operations performed in the original text. For example, one could add two encrypted numbers and then another person could decipher the result, without being able to find the initial value of the computed numbers.

According to Valeria Nikolaenko and Dan Boneh [18], the current mining algorithms data should have access to data in clear text for proper handling. Thus preserving privacy and confidentiality in the current mining process is a vulnerability that needs to be resolved. Thus, several studies are being directed towards the development of schemes and techniques that allow the manipulation and computation of encrypted data, with the prospect that the homomorphic encryption can solve this problem effectively and efficiently [16].

One solution to this problem was first defined in 1978 where Rivest, Adleman and Dertouzos [21] suggested the construction of secret homomorphisms - privacy homomorphisms - as a way of providing a technique that meets this demand. However, the technique presented in their proposal [21] called for unfavorable conditions that would make the practice technique a fully homomorphic cryptosystem.

The homomorphic encryption has two forms, partially homomorphic encryption and fully homomorphic encryption. The partially homomorphic encryption is defined when it has limits on the amount of transactions with encrypted [23] data. Nevertheless, the fully homomorphic encryption is a cryptographic system that allows you to make a set of arbitrary mathematical operations (without limitation) in the ciphertext, which should result in another ciphertext corresponding to the result of the operation in plain text. In this research we focus on fully homomorphic encryption in order to meet the requirements of mining encrypted data.

Fully homomorphic encryption with any circuit can be evaluated homomorphically, allowing the construction of programs that can run with the encodings of its inputs to produce an encryption of its output. Programs such as homomorphic never decode their inputs, they can be used by untrusted third parties, therefore it is impossible to reveal its entry data and internal processes.

The existence of a fully homomorphic and efficient cryptographic system would have great practical implications for the outsourcing of private computing, as in the context of cloud computing. The example of a cloud-based model is shown in Figure 1, which illustrates three different hospitals which provide sensitive data to the cloud. The cloud computing
platform analyzes and extracts useful information (data mining operation) from data entry and delivers it to recipients.

In order to preserve the privacy and confidentiality of patient information and using fully homomorphic encryption, the Platform for Cloud Computing will perform operations using only encoded data and deliver the results to the recipients. Thus, no information can be leaked during calculation or delivery stages of the communication.

In this context, we intend to present a study to identify methods and proposals that use completely homomorphic encryption to maintain the privacy and confidentiality of information during the process of data mining.

This paper is organized as follows: section two presents the fundamentals and concepts related to the fully homomorphic encryption and Privacy in Data Mining. Section three presents the methodology applied in the study, section four the results are presented in a survey and analysis of studies cataloged, and in section five final remarks are made in conclusion.

2 Preliminary

2.1 Fully Homomorphic Encryption

The concept of homomorphic encryption was initially defined by Ronald Rivest, Len Adleman, and Michael Dertouzos [21], after verifying that the RSA cryptographic system developed by Ronald Rivest, Shamir Adl, Len Adleman [22], had multiplicative homomorphism. Thus, Rivest, Adleman and Dertouzos, defined special encryption functions called "privacy homomorphisms", which are a subset of arbitrary encryption schemes.

The homomorphic public key encryption is a cryptographic system that allows the performance of a set of operations on the data when they are encoded, resulting in its data appearing in plain text. The public key homomorphic encryption allows it to perform various calculations on the data without revealing any information of the encoded message, preserving the privacy and confidentiality of data.

The fully homomorphic cryptographic models remained as speculation for more than 30 years, because researchers have failed to develop a fully homomorphic encryption method that could compute arbitrary numbers of operations. But in 2009, the researcher Craig Gentry [12] proposed a system which has a valid encryption scheme with a fully homomorphic public key that can arbitrarily compute on encrypted data, based on ideal lattices.

Ideal lattices is an area that studies mathematical structures, and has applications in cryptography because of the complexity involved in solving difficult problems. It is defined as a set of points in n-dimensional space with a periodic structure. In other words, a lattice vector is a discretized space, using the concept of standard dimension orthogonal linear transformation, among others [20].

As defined by Gentry, the homomorphic encryption schemes completely preserve the operations of addition and multiplication on encrypted blocks, i.e:

Definition: Given that $E(m)$ is the application of the encryption algorithm to a message $m$, a cryptographic scheme is fully homomorphic if:

$$E(m_1 + m_2) = E(m_1) + E(m_2),$$

$$E(m \cdot m_2) = E(m_1) \cdot E(m_2),$$

• For any $m_1$ and $m_2$ block of the message to be encrypted; and
• The same applies to any number of consecutive operations performed on a single block.

The research of Craig Gentry [12] is based on polynomial ideals for obtaining a scheme of restricted homomorphic encryption restricted (SHE - Somewhat homomorphic encryption). This scheme is able to add and multiply encrypted texts in a homomorphic way, but as transactions are conducted noise is added to the ciphertext. According to Gentry the decoding algorithm works provided that such noise does not exceed a certain threshold.

Using the concept that is called bootstrapping, Craig Gentry proposes the construction of a new scheme that can decode and reduce noise homomorphically. However, this adaptation leads directly to increasing the size of the parameters, making it impossible to implement the scheme. From the Gentry scheme other schemes have been proposed in an attempt to be more practical and efficient, an example based on the Learning with Errors (LWE) proposed by Zvika Brakerski and Vinod Vaikuntanathan [4].

2.2 Privacy in Data Mining

Data mining helps in extracting useful knowledge from large data sets, but the process of data collection and data dissemination may, however, result in an inherent risk of threats to confidentiality and data privacy. Some personal information about individuals, companies and organizations must be deleted before it is shared or published, unless such information is encoded. Thus, preserving privacy in data mining has become a very important issue in the last decade.

The problem of learning something without revealing the data itself has not been defined recently. This problem was proposed in 1982 by Andrew C. Yao [24] in his article "Protocols for Secure Computations", where Yao defines a model that can ensure privacy between the parties. To [24], a
protocol between two parties is considered safe to perform operations if participants do not learn anything beyond what is revealed by the output of the circuit.

In this context, the term originated in Privacy Preserving Data Mining (PPDM - Privacy-Preserving Data Mining) that refers to the area of data mining for protecting sensitive information from unsolicited disclosure. Mining techniques over traditional data analyze and model the data set statistically, while the preservation of privacy is primarily concerned with the protection against disclosure of individual data records. This separation of domain points to the technical feasibility of PPDM.

The term Privacy-Preserving in Data Mining - PPDM was introduced by [11] and [17]. These researchers considered two fundamental problems in PPDM: i) preservation of privacy in data collection, and ii) privacy during the mining process of a partitioned data set from several private companies.

The Agrawal and Srikant [11] researchers developed an algorithm of randomness that allows a large number of users to contribute with their private records for centralized data mining, limiting disclosure of their records; Lindell and Pinkas [17] designed a cryptographic protocol for the construction of a decision on a data set horizontally partitioned between two parties tree. These methods were later refined and extended by many researchers.

The goal of preserving privacy in data mining (PPDM) is to extract relevant knowledge from large amounts of data, while protecting sensitive information [9]. Thus, the process to preserve the privacy and confidentiality of data during the data mining requires new technologies and advancements, especially in the area of modern cryptography.

3 Applied Method

Conducting a survey of relevant studies, using the mechanisms of systematic mapping, this mechanism establishes a formal investigation in the literature, in order to give credibility to the ongoing research in the area.

The steps of the systematic literature mapping process were defined in order to explore and find the most relevant primary studies able to respond to the problem: "It is computationally feasible to explore sensitive data in data mining when they are encrypted using fully homomorphic encryption in order to ensure security, privacy and confidentiality efficiently?".

With the implementation of the systematic mapping protocol, studies dealing with Fully Homomorphic Encryption, Data Mining, privacy and confidentiality of data were identified.

Thus, with the implementation of the search strings in the mills: ACM Digital Library, Engineering Village, IEEE Xplore, Springer Link, Scopus and ScienceDirect, a total of 213 trials, distributed according to Figure 2 were collected.

Five rounds for analysis of the criteria for inclusion and exclusion of the collected studies were performed. In each round, the studies were analyzed in order to filter the most relevant articles within the research theme. After performing the first 4 rounds the studies listed came to the total of 25 possible studies for primary selection, as Table I. Thus, preserving privacy in data mining has become a very important issue in the last decade.

![Figure 2: Articles listed in the databases queries. Data generated in January of 2014.](image)

<table>
<thead>
<tr>
<th>TABLE I. DESCRIPTION THE STEPS IN REVIEW</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>STEPS</strong></td>
</tr>
<tr>
<td>Primary Selection</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Secondary Selection</td>
</tr>
</tbody>
</table>

With the results of the primary selection complete, then the secondary selection began with the implementation of the fifth round, which aims to identify studies that show a method or protocol used in mining that aims to preserve the privacy and confidentiality of data in addition to its fully homomorphic encryption base. However, the evaluation of the proposed method was considered in addition to their results. Thus, Table II presents the data of secondary selection with the final results of six cataloged studies after the implementation of the five rounds.
4.1 Statistical analysis using the Linear Regression

Linear regression is a statistical analysis tool widely used in data mining because it allows you to verify the linear relationship between the dependent and independent variables in a dataset. But, in the current approaches, it is necessary to have data in plain text so that we can perform statistical operations. The FHE is a solution to perform statistical analyzes of the encoded data while preserving confidentiality and privacy.

Cataloged in this study “Privacy Preserving linear regression modeling of distributed databases”, the authors Weiwei Fang et al. [10], address the tradeoff between privacy and statistical analysis with a focus on linear regression applied in the field of data mining. The question of this study is defined based on the following survey: how to identify a feasible process to maintain the linear relationship between the dependent and independent variables, without disclosing your personal data in the data mining process.

Thus, the PPRCP (Privacy Preserving Regression Coefficient Protocol) protocol, which allows one to perform a linear regression using the FHE was presented. In evaluating implementation of this protocol, no further information was leaked, so the authors argue that the protocol is computationally indistinguishable. The protocol securely computes the regression coefficient without the leakage of sensitive data.

The security of the proposed protocol is based on the cryptographic scheme, if the fully homomorphic encryption is secure, then the protocol is also secure.

4.2 Recovery blocks of data bits with the protocol PIR

The PIR (Private Information Retrieval) is a family of two-party protocols in which one party has a database, and the other part wants to consult it with some restrictions and guarantees of privacy. This approach was introduced by Chor, Goldreich, Kushilevitz and Sudan [7], and since then has attracted attention of researchers.

The PIR protocol allows a client to retrieve a certain element of your choice in a database without the owner of the database being able to determine which element was selected. To ensure the confidentiality admittedly is a trivial solution, sending the entire database to the client allowing it to see with perfect privacy, but there is a problem regarding the computational cost, because it is detrimental to communication in the occurrence of large databases. Another problem in this proposal is that the user can obtain additional information.

Collected in the study “Single-Database Private Information Retrieval from Fully homomorphic encryption”, was the presentation of the protocol for private information retrieval that allows a user to retrieve the i bit of a database of n bits that are shown, without disclosing to the administrator of the database the value of index i, at a lower cost of communication. Furthermore, propose an extension of the PIR protocol for a recovery protocol of private blocks PBR (Private Block Retrieval) being more efficient, both protocols (PIR and PBR) are based on FHE.

Comparing the proposal with existing protocols PIR and PBR, it is concluded that the presented method is conceptually simpler. Overall, the PBR protocol is practical and more efficient than the existing PBR protocols in terms of total execution time when a high-speed network is available.
In the safety analysis presented, it was demonstrated that the protocol is semantically secure if the fully homomorphic encryption scheme is semantically secure.

4.3 Association Rules of Data with Privacy Preservation

According to Brusso [6], “association rules are descriptive patterns that represent the probability that a set of items will appear in a transaction where another set is present”. An algorithm is widely applied in mining due to the possibility of finding patterns.

Typically, association rules represent patterns in existing stored transactions. For example, from a database that enters their items purchased by customers, a strategy for mining using association rules could generate the following rule: \{belt, purse\} $\rightarrow$ \{shoes\}, which indicates that a customer purchasing a belt and a bag, with a certain degree of certainty, will also buy shoes.

The authors, Mohamend Kaosar, Russell Paulet and Yi You [15] propose a technique ARM (Association Rule Mining) of secure comparison (making checks of interesting correlations in a set of database) based on homomorphic encryption scheme completely promoting greater efficiency due to the reuse of resources.

In order to evaluate the presented method, a prototype software was implemented to test the feasibility of this approach by using a fully homomorphic encryption scheme open source library called Smart-Vercauteren for cryptographic operations. With this library, the authors could measure the time required for the method of comparison of integers.

The authors state that the main contribution of the present study is the use of fully homomorphic encryption to solve the problem of association rules while preserving privacy. The protocol was safe, based on hardness assumption of the cryptographic system. The technique proposed privacy preservation can be used in the collaborative filtering of data between two databases.

4.4 System Search on Encrypted Data Base

The search data is an important tool for mining, where the storing of data in encrypted form is required it will necessary to find a solution that allows for the processing of such data. Thus, the authors present in the study “The Implementation and Application Fully Homomorphic Encryption Scheme” [14], a solution that combines the Attributes Based Encryption (ABE - Attribute Based Encryption) and Fully Homomorphic Encryption that can perform computations with encrypted data. In the proposed solution, the encoded data is computed by servers in the cloud in order to preserve the privacy of the system’s input and output of the circuit, but the FHE schemes based on lattices and ideals based on LWE (Learning With Errors) do not offer solutions that are truly practical for the method presented.

ABE is a collection of cryptographic tools based on attributes and policies assigned to users by an authority. In particular, it allows attaching attributes and policies to the message to be encrypted so that only a receiver that is assigned policies/attributes can decrypt it. The attributes are Boolean variables with arbitrary labels and political calculations are represented as Boolean circuits with the attribute variables (which evaluate to true or false) [3].

The method presented allows one to make searches on encrypted data stored in the cloud, ensuring privacy and confidentiality of information in storage, in the survey conducted, algorithms that run in this process and in transmission between the server and the requesting user. In the proposed system, all the data computed by the cloud servers are encrypted in order to preserve the privacy and confidentiality of data.

But according to reviews, FHE is not yet a practical and efficient scheme that allows its use in making systems as presented. Efforts must be made to improve the efficiency of cryptographic schemes in order to make practical applications that require security, confidentiality and privacy that can provide FHE.

4.5 Set Operations in Distributed Data

The mining algorithms need to unite the various databases so that they can perform operations on these sets. Since the amount of data submitted to the applications of data mining has grown considerably as an indirect result of reductions in the cost of collection, transmission and storage of data there is an increasing amount of sensitive data stored on untrusted remote servers which raises concern for confidentiality.

During the mining process steps are required to perform operations on data sets, such as union and intersection, to identify relationships. These operations are widely used in the mining process and requires confidentiality and privacy, as it will handle third-party data.

Thus, the reported study in [8] has proposed a protocol that preserves the privacy of the data using the disjunctive normal form, called PPDNF (Privacy Preserving Disjunctive Normal Form) in operations with distributed set data, without revealing any data beyond the information inferred from the input operation. The fully homomorphic encryption scheme is used to ensure the safety of the protocol and perform operations on encrypted data. The structure of PPDNF protocol does not depend on the structure of the cryptographic system, however, the efficiency of the proposed protocol depends on the efficiency of the fully homomorphic encryption scheme.

4.6 Preserving Privacy with Protocol SPIR

Data privacy is a natural and fundamental requirement in many contexts, an example occurs in a commercial database that sells information to users, such as stock information, and is collecting the amount of data that the user retrieves. In this example, both the privacy and confidentiality of the user are essential [13]. The Symmetric PIR (SPIR) protocol should prevent the user to learn more than one record from the database during a session. The main method to measure the cost of such systems, like the SPIR protocol, is by its communication complexity.

The SPIR model (Symmetrically Private Information Retrieval) aims to ensure data privacy and user privacy, ie. each invocation of a SPIR scheme, while maintaining the user's privacy, the scheme should also prevent the user (even as a
rogue) from getting any information other than a single physical bit of data.

The evaluation of the SPIR protocol, proposed in [25], is based on fully homomorphic encryption presents an improvement in the communication complexity making it ideal. Furthermore, the step of bootstrapping FHE makes the protocol computationally less efficient.

4.7 Execution of SQL Queries on Encrypted Data Base

In the study "A Secure Database System using homomorphic Encryption Schemes", the authors Youssef Gahi, Mouhcine Guennoun and Khalil El - Khatib [11] present a technique for executing SQL statements about the encrypted data. They developed a system of secure database processing such queries. The parameters of SQL queries are encrypted by the client and sent to the server for proper processing. The server performs the requested operation in an encrypted database and returns an encrypted result to the client. The advantage of this system is that the database server does not get the content nor the position of the records generated by the query. Tests with the fully homomorphic encryption scheme proposed by Craig Gentry were conducted to verify the efficiency and safety of the method.

The proposed system supports a set of SQL operations such as SELECT, UPDATE, DEL and statistical operations such as COUNT and AVG. During the simulation, the researchers claim that the fully homomorphic encryption scheme is not efficient, citing performance issues with the proposed solution.

4.8 Homomorphic Signatures for Polynomial Functions

The authors Dan Boneh and David Mandell Freeman [24], presented the study "Homomorphic Signatures for Polynomial Functions", which recounts the first homomorphic signature scheme capable of evaluating multivariate polynomials on signed data. Thus, the authors' focus is on the functions that perform arithmetic operations on a set of data such as mean, standard deviation, and other data mining algorithms.

With this approach it is possible to store data and their signatures in an untrusted storage server. If a third party needs to make some calculation, as the standard deviation, it can order directly to the untrusted server that will perform the requested calculation and its homomorphic signature from the signatures of each stored data. With homomorphic signature it is possible to validate the integrity of the result.

However, if you publish a pair \((f, \sigma)\), where \(f\) is the result of the operation of the standard deviation and \(\sigma\) is the homomorphic signature bypass operation (derived from calculations homomorphic with the signing of other data), any party who owns the Alice's public key can check the signature and verify the integrity of the results generated by the untrusted server over the computed data.

This method can be applied to a tool known as a decision tree that is widely used in data mining processes. The decision tree is widely used in ranking algorithms, it is a simple representation of knowledge and an efficient way to build classifiers that predict or indicate classes or information based on the values of attributes of a dataset. The decision tree is very useful in the process of extracting previously unknown information from large databases.

This study is presented as the first step towards fully homomorphic signature, since the proposed system uses ideal lattices that is the basis of the proposal of CCH presented by Craig Gentry.

In this context, if the fully homomorphic signature is implemented (based on arbitrary calculations on signed data), an untrusted server can execute more complex data mining algorithms about a particular data set and verify the authenticity of the transactions. For example, given a set of signed data, the server could publish a signed decision tree. As the signatures are private, if publish a signed decision tree, the original dataset will not be exposed and can be permitted to validate the published results.

A fully homomorphic signature scheme would be a parallel system and useful for existing fully homomorphic encryption schemes. Even if a fully homomorphic signature scheme was not developed, it would be very useful to amplify the set of admissible functions proposed in the present study [24].

5 Conclusions

Preservation of privacy and confidentiality for data mining is a problem for data owners and many researchers devote efforts to solve this problem. A fully homomorphic encryption is presented as an amazing possibility for solving this challenge.

Homomorphic encryption provides privacy and security for data processing in untrusted environments, and can be used in protocols that perform the steps of data mining. Thus, it was shown in this study the significance and importance of using a fully homomorphic cryptographic scheme in the application of data mining.

Analyzing the different cataloged studies, it was observed that each author seeks to solve the problem of protecting confidentiality and privacy in data mining, addressing it in isolated parts of the mining process.

All analyzed studies make it clear that the safety of the proposed solution depends on the security of the cryptographic system completely, i.e. if the cryptosystem is secure then the proposed scheme is also secure.

It is quite clear that homomorphic encryption still requires deeper studies in order to become a practical and efficient implementation that will succeed in the preservation of security, integrity, privacy and confidentiality of data in several areas of data mining.

Another noticeable point is that the efficiency of the schemes that have been proposed in the studies catalogued, also depends on the efficiency of the fully homomorphic cryptosystem. At this point the cryptographic system still does not meet the desired requirements, but scholars are seeking to solve this problem with new solutions.
6 Future Works

This section is an important point in the research, it allows for opportunities for future work aimed to continue this research. Aiming to highlight these opportunities there are some suggestions presented for new referrals that were identified during the study in question:

- Conduct a search through the references presented in the studies cataloged in order to identify new studies;
- Evaluate the proposed methods to verify the possibility of integrating them with the goal of keeping all steps of data mining safe and preserving data privacy;
- Assess new practical and efficient fully homomorphic encryption schemes that can be applied in data mining;
- Resolve or improve the computational cost of the cryptographic process used in the fully homomorphic encryption.

As shown above, by using homomorphic encryption in data mining is not feasible due to the cost of the cryptographic system processing. If the fully homomorphic cryptosystems are improved, then the trend is that more protocols to the steps of data mining will be developed, requiring less processing on cryptographic regard and devoting greater computational power to the mining process.

7 References