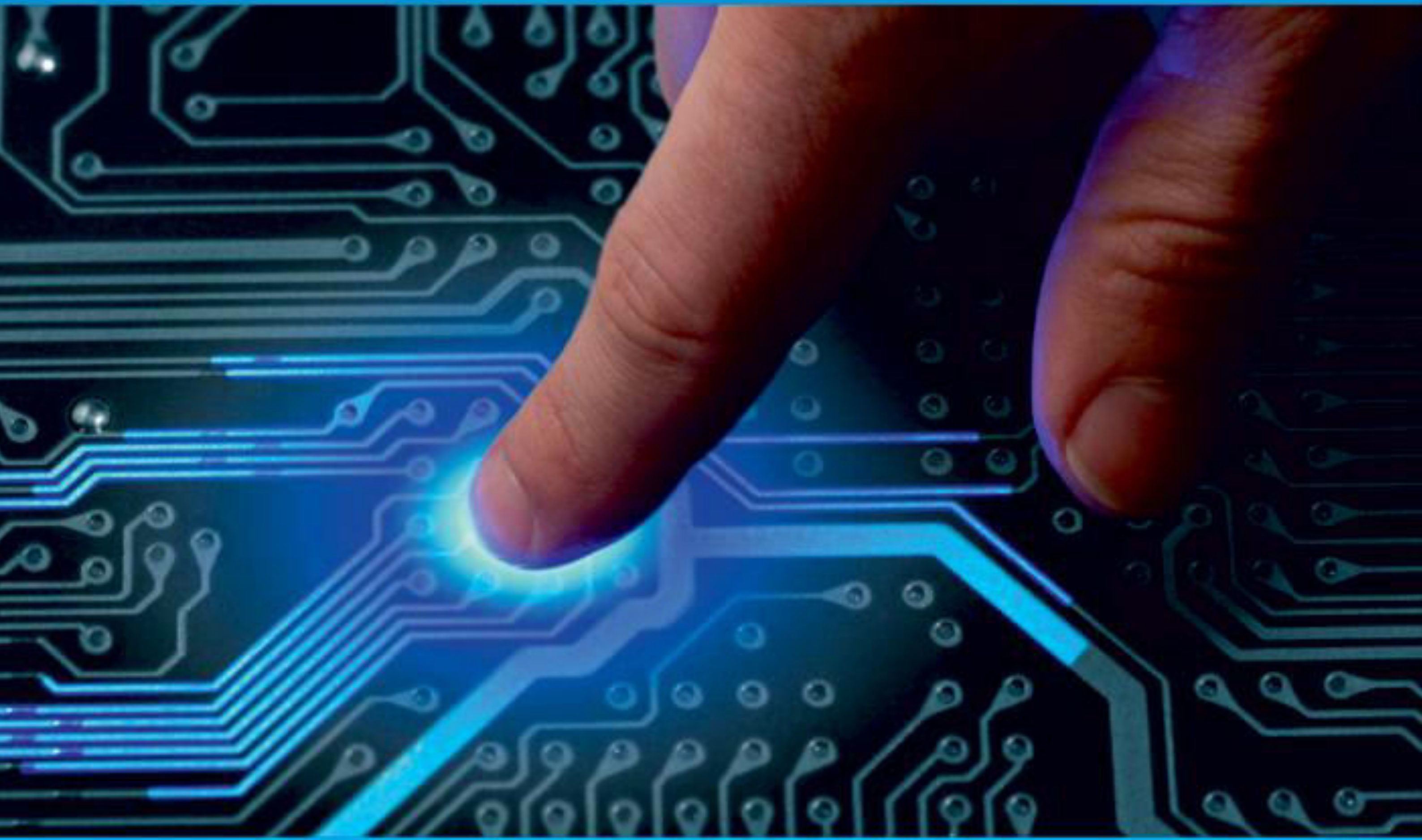
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 7, July 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

| DOI: 10.15680/IJIRCCE.2021.0907211 |

# Enforcing Privacy in Itemset Mining on Encrypted Cloud

**Dr. R. Ramyadevi, Jayavardhini V, Rohini H, Rohini P S,**

Assistant Professor, Dept. of Computer Science and Engineering, Velammal Engineering College, Chennai, India

UG Students, Dept. of Computer Science and Engineering, Velammal Engineering College, Chennai, India

**ABSTRACT:** Data Mining, the process of extracting and discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems, is used to get a particular data from the public cloud services using storage-as-a-service mechanism. In frequent mining usually the interesting associations and correlations between item sets in transactional and relational databases are found. The existing system construct the first homomorphic signature scheme that is capable of evaluating multivariate polynomials on signed data. Using the public key and a signed data set, there is an efficient algorithm to produce a signature on the mean, standard deviation, and other statistics of the signed data. We propose a privacy-preserving framework for secure frequent item set mining on encrypted data, where only one aided server referred to as Evaluator is needed besides the Cloud Service Provider (CSP). The CSP collects encrypted transactions and maintains an encrypted transaction database, and the Evaluator assists the CSP to carry mining on encrypted data. The data miner can submit frequent item set mining queries to the CSP, and we leverage an Evaluator to assist the CSP to send frequent item set mining efficiently on the transaction database. Experimental results show that our protocols are much more efficient than previous work with the same security levels. In this system, the data user register and login to the system, and encrypt the stored data using public key. Then upload the data to the CSP. In order to preserve privacy of frequent item set mining in public cloud services all the transactions are encrypted before contributing to the CSP. The CSP collects encrypted transactions and maintains an encrypted transaction database, and the Evaluator assists the CSP to carry mining query operations. The Evaluator generates key pairs, a Public key and a Secret Key for upload and download process for the user and miner respectively. Finally, the data miner downloads the decrypted data and carry the data mining process for finding frequent item sets.

**KEYWORDS:** Data mining, privacy, Cloud Service Provider, frequent itemset, Evaluator, encrypted, secret key.

## I. INTRODUCTION

In computer science, Data mining is one of the most useful techniques that help entrepreneurs, researchers, and individuals to extract valuable information from huge sets of data. Data mining is also called Knowledge Discovery in Database (KDD). The knowledge discovery process includes Data cleaning, Data integration, Data transformation, Data mining, Pattern evaluation, and Knowledge presentation. It is a process used by organizations to extract specific data from huge databases to solve business problems. It primarily turns raw data into useful information.

Clustering is a division of information into groups of connected objects. Describing the data by a few clusters mainly loses certain confine details, but accomplishes improvement. It models data by its clusters. Data modelling puts clustering from a historical point of view rooted in statistics, mathematics, and numerical analysis. From a machine learning point of view, clusters relate to hidden patterns, the search for clusters is unsupervised learning, and the subsequent framework represents a dataconcept. From a practical point of view, clustering plays an extraordinary job in datamining applications. For example, scientific data exploration, text mining, information retrieval, spatial database applications, CRM, Web analysis, computational biology, medical diagnostics, and much more. In other words, wecan say that Clustering analysis is a datamining technique to identify similar data. This technique helps to recognize the differences and similarities between the data. Clustering is very similar to the classification, but it involves grouping chunks of data together based on their similarities.Storage as a service (SaaS) is a cloud business model in which a company leases or rents its storage infrastructure to another company or individuals to store data.

## II. LITERATURE REVIEW

### A. *Dynamic item set counting andimplication rules for market basket data*

We consider the problem of analysing market-basket data and present several important contributions. First, we present a new algorithm for finding large item sets which uses fewer passes over the data than classic algorithms, and yet uses fewer candidate item sets than methods based on sampling. We investigate the idea of item reordering, which can improve the low-level efficiency of the algorithm. Second, we present a new way of generating ``implication rules," which are normalized based on both the antecedent and the consequent and are truly implications (not simply a measure of co-occurrence), and we show how they produce more intuitive results than other methods. Finally, we show how different characteristics of real data, as opposed to synthetic data, can dramatically affect the performance of the system and the form of the results. Within the area of data mining, the problem of deriving associations from data has recently received a great deal of attention. The problem was first formulated by Agrawal et al, and is often referred to as the ``market-basket" problem. In this problem, we are given a set of items and a large collection of transactions which are subsets (baskets) of these items. The task is to find relationships between the presences of various items within those baskets. There are numerous applications of data mining which fit into this framework. The canonical example from which the problem gets itsname is a supermarket. The items are products and the baskets are customer purchases at the checkout. Determining what products customers are likely to buy together can be very useful for planning and marketing. However, there are many other applications which have varied data characteristics. For example, student enrolment in classes, word occurrence in text documents, users' visits of web pages, and many more. In this paper, we address both performance and functionality issues of market-basket analysis. We improve performance over past methods by introducing a new algorithm for finding large item sets (an important sub problem). We enhance functionality by introducing implication rules as an alternative to association rules (see below). Thus, the algorithm performs as many passes over the data as the maximum number of elements in a candidate item set, checking at pass k the support for each of the candidates in Ck. The two important factors which govern performance are the number of passes made over all the data and the efficiency of those passes.

### B. *Data mining approaches for intrusion detection*

In this paper we discuss our research in developing general and systematic methods for intrusion detection. The key ideas are to use data mining techniques to discover consistent and useful patterns of system features that describe program and user behaviour, and use the set of relevant system features to compute (inductively learned) classifiers that can recognize anomalies and known intrusions. Using experiments on the send mail system call data and the network tcp dump data, we demonstrate that we can construct concise and accurate classifiers to detect anomalies. We provide an overview on two general data mining algorithms that we have implemented: the association rules algorithm and the frequent episodes algorithm. These algorithms can be used to compute the intra- and inter- audit record patterns, which are essential in describing program or user behaviour. The discovered patterns can guide the audit data gathering process and facilitate feature selection. To meet the challenges of both efficient learning (mining) and real-time detection, we propose an agent-based architecture for intrusion detection systems where the learning agents continuously compute and provide the updated (detection) models to the detection agents. An intrusion can be defined [HLMS90] as “any set of actions that attempt to compromise the integrity, confidentiality or availability of a resource”. Intrusion prevention techniques, such as user authentication (e.g., using passwords or biometrics), avoiding programming errors, and information protection (e.g., encryption) have been used to protect computer systems as a first line of defence. Intrusion prevention alone is not sufficient because as systems become ever more complex, there are always exploitable weakness in the systems due to design and programming errors, or various “socially engineered” penetration techniques. For example, after it was first reported many years ago, exploitable “buffer overflow” still exists in some recent system software due to programming errors. The policies that balance convenience versus strict controlof a system and information access also make it impossible for an operational system to be completely secure. Intrusion detection is therefore needed as another wall to protect computer systems. Many researchers have proposed and implemented different models which define different measures of system behaviour, with an ad hoc presumption that normalcy and anomaly (or illegitimacy) will be accurately manifested in the chosen set of system features that are modelled and measured. Intrusion detection techniques can be categorized into misuse detection, which uses patterns of well-known attacks or weak spots of the system to identify intrusions; and anomaly detection, which tries to determine whether deviation from the established normal. The main shortcomings of such systems are: known intrusion patterns have to be hand-coded into the system; they are unable to detect any future (unknown) intrusions that have no matched patterns stored in the system.

Our research aims to eliminate, as much as possible, the manual and ad-hoc elements from the process of building an intrusion detection system. We take a data-centric point of view and consider intrusion detection as a data analysis process. Anomaly detection is about finding the normal usage patterns from the audit data, whereas misuse detection is about encoding and matching the intrusion patterns using the audit data. The central theme of our approach is to apply data mining techniques to intrusion detection. Data mining generally refers to the process of (automatically) extracting models from large stores of data [FPSS96]. The recent rapid development in data mining has made available a wide variety of algorithms, drawn from the fields of statistics, pattern recognition, machine learning, and database. Several types of algorithms are particularly relevant to our research, maps a data item into one of several predefined categories. These algorithms normally output “classifiers”, for example, in the form of decision trees or rules. An ideal application in intrusion detection will be to gather sufficient “normal” and “abnormal” audit data for a user or a program,then apply a classification algorithm to learn a classifier that will determine (future) audit data as belonging to the normal class or the abnormal class; Finding out the correlations in audit data will provide insight for selecting the right set of system features for intrusion detection; Sequence analysis: models sequential patterns.

## III. PROPOSED WORK

### A. *Overview*

We present three practical privacy-preserving frequent item-set mining protocols in the paper. We propose a privacy-preserving framework for secure frequent item set mining on encrypted data, where only one aided server referred to as Evaluator is needed besides the Cloud Service Provider (CSP). The CSP collects encrypted transactions and maintains an encrypted transaction database, and the Evaluator assists the CSP to carry mining on encrypted data. The data miner can submit frequent item set mining queries to the CSP, and we leverage an Evaluator to assist the CSP to send frequent item set mining efficiently on the transaction database. Experimental results show that our protocols are much more efficient than previous work with the same security levels.

### B. *System Architecture*

We provided separate registration for data user and data miner. After successful registration the data user login in system, they can upload the data in public cloud. The stored data is in encrypted form. The data miner will request for a particular data of his interest. The evaluator generates and send public key to the user and secret key to the miner. The user accepts or rejects the request of the data miner. Once the user accepts the request, the data miner can decrypt and view the data using secret key. The process of clustering helps to view the data efficiently.

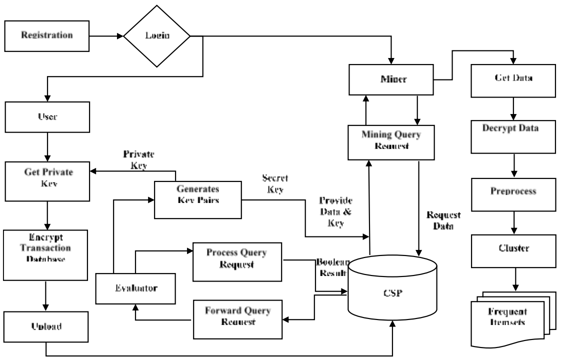


Fig 1: system architecture

### C. *Algorithm*

Boneh-Goh-Nissim (BGN) encryption isa somewhat homomorphic encryption, which was proposed byBoneh et al. [9]. With the homomorphic property, it can computean arbitrary number of additions and also one multiplication onencrypted data. Specifically, we have

- Given $\|m1\|$ and $\|m2\|$, compute $\|m1 + m2\| = \|m1\|\odot\|m2\|$
- Given $\|m1\|$ and $\|m2\|$, compute $\|m1 \times m2\| = \|m1\|\otimes\|m2\|$

where $\|m\|$ represents the BGN ciphertext of a plaintext m, $\odot$ and$\otimes$ denote the corresponding operations in the ciphertext domain.More construction details of BGN can be found in [9].

## IV. CONCLUSION

We propose three practical privacy-preserving frequent item set mining protocols on encrypted cloud data. Our Protocol 1 achieves an extremely higher mining performance while our Protocol 2 provides a stronger privacy guarantee. To-ward more practical efficiency, we optimize the performance of our second protocol by leveraging a minor leakage of privacy to achieve our Protocol 3.

## V. FUTURE WORK

We focus on further improving the efficiency of frequent item set mining on larger scale database. The special properties of streaming data (real-time and continuous) and its extensively practical application, we focus on the study of the efficient privacy-preserving frequent item set mining algorithm on stream data.

## REFERENCES

[1] X. Ge, L. Yan, J. Zhu, and W. Shi, “Privacy-preserving distributed association rule mining based on the secret sharing technique,” in Software Engineering and Data Mining (SEDM), 2010 2nd International Conference on. IEEE, 2010, pp. 345–350.
[2] J. Lai, Y. Li, R. H. Deng, J. Weng, C. Guan, and Q. Yan, “Towards semantically secure outsourcing of association rule mining on categorical data,” Information Sciences, vol. 267, pp. 267–286, 2014.
[3] V. Nikolaenko, U. Weinsberg, S. Ioannidis, M. Joye, D. Boneh, and N. Taft, “Privacy-Preserving Ridge Regression on Hundreds of Millions of Records,” in Proc. Of IEEE S&P’13, 2013.
[4] A. Peter, E. Tews, and S. Katzenbeisser, “Efficiently Outsourcing Multi-party Computation Under Multiple Keys,” IEEE Transactions on Information Forensics and Security, vol. 8, no. 12, pp. 2046–2058, 2013.
[5] R. Bost, R. A. Popa, S. Tu, and S. Goldwasser, “Machine learning classification over encrypted data,” Crypto ePrint Archive, 2014.
[6] R. Curtmola, J. A. Garay, S. Kamara, and R. Ostrovsky, “Searchable Symmetric Encryption: Improved Definitions and Efficient Construction-s,” inProc. of ACM CCS’06, 2006.
[7] O. Goldreich, Foundations of cryptography: volume 2, basic application-s. Cambridge university press, 2004.
[8] P. Paillier, “Public-key cryptosystems based on composite degree residuosity classes.” Springer, 1999, pp. 223–238.
[9] D. Boneh, E.-J. Goh, and K. Nissim, “Evaluating 2-dnf formulas on ciphertexts,” in Theory of cryptography. Springer, 2005, pp. 325–341.
[10] J. Han, M. Kamber, and J. Pei, Data mining: concepts and techniques: concepts and techniques. Elsevier, 2011.

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com