IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Issue-6, June 2019 E-ISSN: 2347-2693

High Confidence Association Rule for Product Selling Strategy

Mamata S. Kalas[1*], Amruta G. Unne[2]

[1,2]Department of Computer Science and Engineering, KITs College of Enginerring Kolhapur

Corresponding Author: mam2kalas@rediffmail.com

DOI: https://doi.org/10.26438/ijcse/v7i6.11841188 | Available online at: www.ijcseonline.org

Accepted: 25/Jun/2019, Published: 30/Jun/2019

***Abstract*—** Mining association rules help data owners to unveil hidden patterns from their data to analyze & predict the operation on application domain. However, mining rules in a distributed environment is not a minor task due to privacy concerns. Data owners are interested in collaborating to mine rules on different levels; however, they are concerned that sensitive information related to somebody involved in their database might get compromised during the mining process. Here formulate the problem to solving association rules queries in a environment such that the mining process is confidential and the outcomes are differentially private. Work proposes a privacy-preserving association rules mining where strong association rules are determined privately, and the results returned satisfy differential privacy. Finally done experiments on real-life data it shows that designed approach can efficiently answer association rules queries and is scalable with increasing data records.

***Keywords*—** Association rules mining, Data Privacy, Data Mining, High confidence.

I. INTRODUCTION

Because of the fast advancement of data collection and storage technologies, extracting knowledge and hidden patterns from stored data has become a major necessity for individuals, companies, and government agencies. In any case, extract information is considered a challenge when the data is distributed over multiple owners, and each data owner is concerned about the privacy of individuals in his data. For instance, companies might be interested in obtaining information concerning the financial status of individuals from different products and sale cost. Privacy-Preserving Data Mining (PPDM) techniques has been utilized in the context of distributed computing to protect the confidentiality of the data of each provider, while still enabling the providers to perform data mining tasks, such as frequent itemsets mining and association rules mining, on the distributed data.

This work describes a protection safeguarding approach for rules mining. Three kinds of members are expected in the proposed model: data providers, master miner, and data consumers. The information being shared is as a table that is on a level plane divided into sub-tables, every one of which is facilitated by one information supplier. Proposed framework preserve the privacy of each provider's selling data while also protecting the query confidentiality against the data providers.

The aim of this work can be summarized as follows:

1. Design a privacy-preserving approach for answering association rules queries with the goal of preserving both data privacy and high confidence.
2. Our proposed approach to provide the differentially private association rules.
3. The proposed method preserves the privacy of the mined data by preventing each data provider from learning sensitive information about other data providers during the mining process.
4. Conduct performance evaluation on real-life data to study the scalability and efficiency of our proposed model.

The rest of this paper is organized as follows. We briefly recall the state-of-the-art methods and problem statement ,objectives and describe the detail of the methodology, respectively, in Sections II,III,IV and V. We then give our conclusion in Section VI.

II. RELATED WORK

Association rule mining is important to getting proper correlations result within large datasets. Association rule mining is related to the frequent item set mining problem which determines sets of items that appear frequently in a dataset. An association rule r is an implication of the form $X \rightarrow Y$, where $X, Y \subseteq I$ are item sets, which captures the concept that a transaction that contains X also contains Y. The strength of an association rule is measured by its confidence, defined as $c(X \rightarrow Y) = \sigma(X \cup Y)/\sigma(X)$.

The support of the rule, defined as $\sigma(X \cup Y)$, is an indicator of the statistical significance of the rule. Typically, for a rule to be representative, its frequency must exceed a minimum support threshold.

Also, Association rule mining has many advantages; mining results make secret of sensitive information about individuals included in the dataset. For instance, with some background knowledge on the items purchased by any person from a super-market customer in a given day, an adversary may be able to narrow down that particular person transaction to a small set, and learn about other items, potentially sensitive, that she may have bought. These problems have been first identified in, and numerous solutions have been proposed since, culminating with the state of the art and provably secures techniques for differentially private data mining.

Differential privacy is a protection model that bounds the probability of an adversary to learn whether a particular individual is present in the dataset or not. To achieve this goal, Differential privacy allows only statistical queries to the data, and the result of each query is problematic with random noise. Existing state of the art differential privacy compliant mining techniques follow a Frequent Item set Mining-centric approach: first, they compute noisy supports for a large number of item sets, and then they identify high-confidence association rules based on item set supports. However, this technique only works well for rules with very large supports. For lower-support item sets, the amount of noise added leads to large errors in the computation of confidence. In fact, to avoid large errors, the state of the art PrivBasis technique does not even compute item set supports for moderate and low frequency item sets. some datasets PrivBasis discards itemsets and corresponding association rules that occur in fewer than 50% of all transactions.

According to authors Mihai Maruseac and Gabriel Ghinita [1], Association rule mining (ARM) was essential in discovering correlations within large datasets. ARM was related to the Frequent Item set Mining (FIM) problem which determines sets of items (i.e., item sets) that appear frequently in a dataset. An association rule r was an implication of the form $X \rightarrow Y$, where $X, Y \subseteq I$ was item sets, which captures the concept that a transaction that contains X also contains Y . The strength of an association rule was measured by its confidence, defined as $c(X \rightarrow Y) = \sigma(X \cup Y)/\sigma(X)$. The support of the rule, defined as $\sigma(X \cup Y)$, was an indicator of the statistical significance of the rule. Typically, for a rule to be representative, its frequency must exceed a minimum support threshold. Although ARM has numerous benefits, mining results may disclose sensitive details about individuals included in the dataset. For instance, with some background knowledge on the items purchased by Alice (a super-market customer) in a given day, an adversary may be able to narrow down Alice transaction to a small set, and learn about other items, potentially sensitive, that she may have bought. This threat has been first identified in, and numerous solutions have been proposed since, culminating with the state of- the-art and provably secures techniques for differentially private data mining.

According to Arik Friedman and Assaf Schuster [2], Differential privacy (DP) was a protection model that bounds the probability of an adversary to learn whether a particular individual is present in the dataset or not. To achieve this goal, DP allows only statistical queries to the data, and the result of each query is perturbed with random noise. Existing state-of-the-art DP-compliant mining techniques follow a FIM-centric approach: first, they compute noisy supports for a large number of item sets, and then they identify high-confidence association rules based on item set supports. However, this method only works well for rules with very large supports. For lower-support item sets, the amount of noise added leads to large errors in the computation of confidence. In fact, to avoid large errors, the state-of-the-art PrivBasis technique does not even compute item set supports for moderate- and low-frequency item sets. Some datasets PrivBasis discards item sets and corresponding association rules that occur in fewer than 50% of all transactions.

According to the Rakesh Agrawal and Ramkrishnat Shrikant [3], the issue of security saving information examination has a long history spreading over various controls. As electronic information about people turns out to be progressively nitty gritty, and as innovation empowers always incredible accumulation and curation of these information, the need increments for a strong, significant, and numerically thorough meaning of protection, together with a computationally rich class of calculations that fulfill this definition. Differential Privacy is such a definition. In the wake of persuading and talking about the significance of differential protection, the dominance of this monograph is given to principal methods for accomplishing differential security, and use of these procedures in inventive mixes, utilizing the inquiry discharge issue as an Ongoing model.

A key point was that, by rethinking the computational goal, one can often obtain far better results than would be achieved b efficiently supplanting each progression of a non-private calculation with a differentially private execution. Regardless of some incredibly amazing computational outcomes, there are as yet principal impediments not simply on what can be accomplished with differential security however on what can be accomplished with any strategy that ensures against a total breakdown in protection. For all intents and purposes every one of the calculations talked about in this keep up differential protection against enemies of self-assertive computational power. Certain calculations are computationally serious, others are productive. Computational multifaceted nature for the foe and the calculation are both talked about.

According to the Raghav Bhaskar, Srivatsan Laxman and Adam Smith [4], Private data analysis in the setting in which a trusted and trustworthy curator, having obtained a large data set containing private information, releases to the public a \sanitization" of the data set that simultaneously protects the privacy of the individual contributors of data and users utility to the data analyst. The sanitization may be in the form of an arbitrary data structure, accompanied by a computational procedure for determining approximate answers to queries on the original data set, or it may be a synthetic data set" consisting of data items drawn from the same universe as items in the original data set; queries are carried out as if the synthetic data set were the actual input. In either case the process is non-interactive; once the sanitization has been released the original data and the curator play no further role.

According to the, Cynthia D work and Aaron Roth [5], the problem of privacy-preserving data analysis had a long history spanning multiple disciplines. As technology enables ever more powerful collection of these data, the need increases for a robust, meaningful, and mathematically rigorous definition of privacy, together with a computationally high class of algorithms that satisfy this definition.

Subsequent to persuading and talking about the importance of differential protection, the prevalence of the book is committed to key procedures for accomplishing differential security, and use of these methods in imaginative mixes, utilizing the inquiry discharge issue as a continuous precedent. A key point is that, by re-evaluating the computational objective, one can regularly acquire much better outcomes than would be accomplished by systematically supplanting each progression of a non-private calculation with a differentially private usage. In spite of some incredibly amazing computational outcomes, there are as yet major restrictions — not simply on what can be accomplished with differential security yet on what can be accomplished with any strategy that ensures against a total breakdown in protection.

According to the Trupti Kenekar1, A. R. Dani [6], Visit sets assume a basic job in numerous Data Mining errands that endeavor to discover intriguing examples from databases, for example, affiliation rules, connections, groupings, scenes, classifiers and bunch. The ID of sets of things, items, side effects and qualities, which frequently happen together in the given database, can be viewed as a standout amongst the most fundamental assignments in Data Mining. The first inspiration for seeking successive sets originated from the need to break down purported market exchange information, that is, to analyze client conduct as far as the obtained items. Visit sets of items portray how regularly things are obtained together.

The existing system had problem of tradeoff between utility and privacy in designing a differentially private FIM algorithm. The existing system does not deal with the high utility transactional item sets. Existing methods has large time complexity. Existing system gives comparatively large size output combination. With communication, data storage technology, a huge amount of information is being collected and stored in the Internet. Data mining, with its promise to efficiently find valuable, non-obvious information from huge databases, was particularly vulnerable to misuse. The condition may become worse when the database contains lots of long transactions or long high utility item sets.

III. PROBLEM DEFINITION

Design privately mining high confidence rules, where each transaction contains a set of items, frequent item set mining tries to find that occur more frequently than a given threshold.

IV. OBJECTIVES

1. A novel technique for differentially private mining of association rules with low and moderate supports.
2. Technique directly samples high confidence rules us ing the exponential mechanism.

V. METHODOLOGY

Followings are some modules here introduced to mining the patterns.

A. HCRMINE

The application of the exponential mechanism requires the computation of the quality functions for each candidate rule. Given a set of n items, the total number of rules that can be generated is $3n - 2n+1 + 1$. Extracting k rules from this set has computational complexity $O(k \times 3n)$ (the quality function must be computed for each candidate in each of the k exponential mechanism execution rounds). This overhead is prohibitive even for moderate values of n. We aim to bring the computational complexity of private high-confidence rules to practical levels. Furthermore, in order to use the privacy budget judiciously, we need to ensure that we do not generate the same rule multiple times. To achieve this, every time we slide the window of eligible items, we always generate rules that contain the newly included item. For instance, if the set of eligible items changes from *{i1, i2, i3, i4}* to *{i2, i3, i4, i5}* all the rules that are generated in the new step must contain item *i5*. As a side effect, the complexity of the rule generation is also reduced.

B. Rule Expansion Optimization

We introduce an optimization that improves the utility of the algorithm by generating more rules than the requested k.

specifically; the optimization uses properties of association rules to infer additional high-confidence rules starting from the set *Rk* of k rules returned by the algorithm.

C. Optimizations of HCR MINING

By combining exponential mechanisms and reservoir sampling, the HCRMine algorithm brings a fundamental improvement compared to existing private rule extraction techniques. However, as the number of requested rules k increases, the privacy budget needs to be divided among more exponential mechanism invocations. As a result, precision will decrease. We formally analyze HCRMine and identify the cause for precision degradation. Next, we introduce two variations of HCRMine that address this problem. We propose the HCRBins method which capitalizes on the parallel composition property of differential privacy, and performs rule extraction on disjoint partitions (or bins) of items.

.

D. Analyzing HCRMine

In order to compare two different exponential mechanisms, *A1* and *A2*, with the same optimal value frothier respective quality function, we only need to analyze the functions *ØA1* and *ØA2*. Moreover, if the quality functions are both defined on subsets of the same set R0 such that the two logarithms are approximately the same.

E. The HCRBins algorithm

To improve mining accuracy, we take advantage of the parallel composition property of differential privacy. Suppose we decompose the set *I* of items into two disjoint subsets, *I1* and *I2*. Furthermore, to obtain *k* rules, we extract *k1* rules from the items in *I1* and *k2* rules from the items in *I2*.Next, we prove that we don't need to split the privacy budget into two components, since we have an instance of parallel composition. Without loss of generality, suppose that *t* is the transaction that is being added or removed from *D* to obtain the neighboring dataset *D′*. We have two possible cases:

a) $t \cap I1 = \emptyset$. In this case, rule r is not affected by the r emoval or addition of *t* since no item in *r* is in *t*. He nce, $q(D, I1, r) = q(D', I1, r)$ and the sampling prob ability doesn't change.

b) $t \cap I1 \neq \emptyset$. The only case when the value of the qu ality function changes is when the rule *r* is formed only by items in $t \cap I1$. Since *I1* and *I2* are disjoint, the rule won't be considered for sampling twice, s o the change in quality function is bounded by its s ensitivity.

F. The HCRPlus algorithm

Given a high-confidence rule $X \rightarrow Y$, HCRBins may not place all the items in $X \cup Y$ in the same bin. In such a case, the rule will not be sampled, decreasing accuracy level in the case when the rule had high confidence. To address this issue, we propose the HCRPlus method, which considers distinct layers of bins. Within each layer, HCRBins is running with budget to extract rules (the sequential composition property applies across layers). Within each layer, the complete set of items *I* is randomly partitioned into m disjoint bins. Hence, the bins will have different composition at each layer. With multiple layers, a set of items is more likely to appear together in the same bin inat least one layer, increasing the probability of sampling quality rules.

VI. CONCLUSION

Work introduces a privacy-preserving approach for answering association rules queries to preserve both data privacy and query confidentiality. The method protects attacks from data consumers by guaranteeing that the returned association rules to the data consumer towards satisfaction. Differential privacy, to preserves the privacy of the mined data by restricting each data provider from learning sensitive information about other data providers during the mining process, next protects the confidentiality of the data consumer's query against the data providers such that the master miner can mine the association rules without revealing the query to the data providers.

ACKNOWLEDGMENT

This paperwork was guided and supported by Mamata S. Kalas. I want to thank the anonymous guide reviewers for their valuable and constructive comments on improving the paper.

REFERENCES

[1] R. Agrawal and R. Srikant. *"Privacy preserving data mining",* InProce edings of International Conference on Management of Data (ACMSIG MOD), **2000.**

[2] R. Bhaskar, S. Laxman, A. Smith, and A. Thakurta. *"Discoveringfrequ ent patterns in sensitive data*". In Proc. of Intl. Conf. on Knowledge Di scovery and Data Mining (KDD), pages **503–512, 2010.**

[3] R. Chen, B. C. Fung, B. C. Desai, and N. M. Sossou. *"Differentiallypri vate transit data publication: a case study on the Montrealtransportati on system"*. In Proc. of Intl. Conf. on Knowledge Discoveryand Data Mining (KDD), pages **213–221, 2012.**

[4] G. Cormode, C. Procopiuc, E. Shen, D. Srivastava, and T. Yu.*"Differe ntially private spatial decompositions."* In ICDE, pages **20–31, 2012.**

[5] C. Dwork, F. McSherry, K. Nissim, and A. Smith. *"Calibrating noiset o sensitivity in private data analysis"*. In TCC, pages **265–284, 2006**.

[6] C. Dwork, M. Naor, O. Reingold, G. N. Rothblum, and S. Vadhan.*"On the complexity of differentially private data release: Efficient algorith ms and hardness results"*. In ACM Symposium on Theory ofComputin g, pages **381–390, 2009.**

[7] C. Dwork and A. Roth. *"The algorithmic foundations of differentialpri vacy"*. Foundations and Trends in Theoretical Computer Science,**9(34) :211–407, 2014.**

[8] A. Friedman and A. Schuster. *"Data mining with differentialprivacy*".

In Proc. of Intl. Conf. on Knowledge Discovery and DataMining (KD D), pages **493–502, 2010.**

[9] A. Ghosh, T. Roughgarden, and M. Sundararajan. *"Universally utility-maximizing privacy mechanisms"*. In ACM Symposium onTheory of Computing, pages **351–360, 2009.**

[10] Omar Abdel Wahab, Moulay Omar Hachami etal *"DARM: A Privacy-preserving Approach for Distributed Association Rules Mining on Hor izontally-partitioned Data"*. Conference Paper · **July 2014 DOI: 10.11 45/2628194.2628206**

[11] Pradeep Chouksey, "*Mining Frequent model Using mass-produced Ap proach*", International Journal of Scientific Research in Computer Scie nce and Engineering, Vol.5, Issue.4, pp.**89-94, 2017**

[12] P.V. Nikam, D.S. Deshpande, "*Different Approaches for Frequent Ite mset Mining*", International Journal of Scientific Research in Computer Science and Engineering, Vol.6, Issue.2, pp.**10-14, 2018**

Authors Profile

Mrs. Mamata S. Kalas , Having been graduate from University Of Mysore, in 1993, From B.I.E.T,Davangere, started her professional carrier there itself. In 1995, came to kolhapur, after her marriage and since then she has worked as lecturer at D.Y.Patil's college of Engg.,Bharati Vidyapeeth's College Of Engg.,Kolhapur.She is M.Tech(CST) Graduate and her dissertation work is based on image segmentation using parametric distributional clustering. She has been awarded with M.TECH (CST) from Shivaji University, Kolhapur of Maharashtra in June 2009. She is persuing Ph. D in computer science and Engineering at Walchand College of Engineering, Reseach center, Shivaji University, under the guidance of Dr.B.F.Momin. She is currently working as an Assistant Professor at KIT'S College of Engg, Kolhapur.She is in her credit, 16 Years of teaching experience, Two Papers presented for international conferences, three papers presented for national conferences, seven papers published in international journals. Her areas of interest are pattern recognition and artificial intelligence, computer architecture, system programming

Miss Amruta G, Unne, having been graduate from Shivaji University, Kolhapur in 2015, from DKTE's Textile and Engineering Institute. She is persuing M.Tech in Computer Science and Engineering at KIT's Collage of Engineering, Kolhapur. Her area of interest are data mining, cloud technology.

