IMPROVING THE PERFORMANCE OF ASSOCIATION RULES HIDING USING HYBRID PARTICLE SWARM OPTIMIZATION ALGORITHM

Eirene Barua[1], Mala Dutta[1], Zafar Jafarov[2]

•

[1]Assam down town University, Assam, India
[2]Azerbaijan Technical University, Baku, Azerbaijan Republic
eirene.adtu@gmail.com, maladuttasid@gmail.com, zafar.cafarov@aztu.edu.az

Abstract

In today's digitized world, data can be taken from many sources like e-marketing sites, social platforms, social networking sites etc. in bulk volumes for usage. Privacy Preserving is a very delicate issue to be looked upon. Hence it becomes necessary to focus on the important privacy preserving parameters. Algorithms for optimization plays an important role in reducing non-sensitive rules in association rule hiding. This paper speaks about a hybrid Particle Swarm Optimization algorithm that requires the properties of all the algorithms which are used for hiding Association Rules and it also highlights the usage of less time.

Keywords: association rules, data privacy-preservation, Particle Swarm Optimization.

I. Introduction

Mining of Association Rules is a very interesting trend which talks about databases on transactions for hiding information which is sensible. There are many types of Rule Hiding Algorithms/Techniques when we talk about sensitive data:

1.Classification Mining Algorithms

2.Decision Region-Based Algorithms

3.Data Perturbation Techniques

Then we have Optimization Algorithms like Particle Swarm Optimization, Cuckoo Optimization Algorithm and Ant Colony Optimization.

This paper describes a hybrid optimization algorithm that gets its characteristics from the mentioned algorithms stated above for hiding association rules, thus producing effective results in less time [1-2].

1.1.Privacy-Preserving Data Mining

Privacy-Preserving Data Mining solely focuses on sensitive information by hiding it based on various rule hiding techniques like association, classification mining, decision region-based and data perturbation techniques.

Here the data is bothered by addition of noise to the set of data used for sensitive data hiding.

There are different models like Sweeney's, Samarati's and Incognito. All the models use multiple secured methods for hiding the data.

1.2. Motivation

At present, there are many applications that are connected with huge amount of data consisting sensible data and information. But when this information is transferred to a third party for extracting the data, there are high chances of the data getting lost. Thus, to solve this problem, data privacy-preservation comes into being.

• Hospital datasets contain a lot of sensible information about patients. Therefore, security should be intact because there are a lot of sensible information related to the patients. Thus, to safeguard this sensible information, there are a lot of privacy-preserving data mining models that can safeguard that sensible information.

• There are cases when robbery takes place in people's houses. Here, the sensible information is the count of the faces appearing in the image, which means however many faces are there in the video has to be saved in order to identify the culprit.

• At present, there exists many companies working on various projects. And that project might contain a lot of information given by the customer. Thus, when the project is sent for processing, there are high chances that the sensible information might be leaked. Thus, for securing such information, privacy-preserving models come into being.

• Also, when talking about bank database, there are a lot of sensible information about the existing customers which should be secured, so that people without proper authorization cannot have access to that information.

1.3. Sweeney's Algorithm:

Sweeney's Algorithm is an algorithm for showing anonymity of Electronic Health Records. Anonymization is accomplished by means of mechanically generalising, substituting, inserting and removing statistics without losing details for research.

1.4. Samarati's Algorithm:

This algorithm scans for the capable k-anonymous explanations by capturing several levels in Domain Generalization Hierarchy. It avails the binary search to get the solution in very less time. Samarati makes the theory that good solutions are the ones where results in a table have minimum generalizations. Thus, her algorithm is considered to look at the concept that determine k-anonymity with minimal suppression. This algorithm fulfils the AGTS model, generalization is applied on column, and suppression is applied on row. MaxSup is the greatest number of tuples that are granted to be suppressed to attain k-anonymity.

1.5. Incognito Algorithm:

Incognito Algorithm produces the set of all possible k-anonymous full domain generalizations of relation T, with an optional tuple suppression threshold. The algorithm consists of iterations of two parts. It begins by examining single-attribute subsets of the quasi-identifier, and afterward repeats scanning k-anonymity with respect to bigger subsets of quasi-identifiers.

II. Literature survey

Using the big itemsets, association rules discover all sets of items with support greater than the minimum support, and then construct the desired rules with confidence greater than the minimal confidence. The lift of a rule is the difference between the actual and predicted support if X and Y were independent. Market basket analysis is a common and widely used application of association rules. Support and confidence are two crucial factors to consider when evaluating association rules. To deal with the sensitivity of association rules, a cyclic technique was adopted (Agrawal et al., 1993; Atallah et al., 1999; Belwal et al., 2013). Similarly, there is confidence reduction (CR), CR2, and generating itemset hiding to improve the hiding process (GIH). For hiding association rules, such type of approaches is being used. (Hahsler et al., 2005; Hong et al., 2011; Kalariya et al., 2015;

Kennedy & Eberhart, 1995; Modi et al., 2010a, 2010b). The work was improved, and algorithm 2b was included to disguise the generating itemset of sensitive rules. (Verykios et al., 2004). An improved algorithm based on Decrease Support and Confidence (DSC) provided better results by hiding predictive association rules. (Wang et al., 2004). Using a genetic algorithm to hide the sensitive rules, privacy-preserving association rule mining over dispersed datasets is possible. (Kesava Murthy & Khan, 2013). Similarly, for the first time, genetic algorithms were used to hide itemsets, and it included a compact pre-large GA-based technique (Goldberg, 1989, 2002) to remove transactions and (Lin et al., 2014) and modified algorithms are introduced to give and insert transactions that are recent.

The most advanced methods are a basic genetic algorithm for transaction deletion and a pre-large genetic algorithm for transaction deletion. For transaction deletion of a fewer variables, and to determine the number of transactions that should be removed in order to reduce negative effects, a modified particle swarm optimization-based algorithm was used.

III. Groundwork for hiding Association rules

The link between the original database D and the sanitised database D is depicted in Figure 1a. The completed cleaned database D is shown in Figure 1h. The coloured area of Figure 1b represents the original database D's frequent itemsets FIs. Figure 1c depicts the sensitive items in D, i.e., itemsets with a Support count of less than Min Support threshold. The non-sensitive data elements are depicted in Figure 1d, where S is in D. Figure 1e–g depicts the three side effects of the sanitization process. The coloured region in Figure 1e represents the hidden failures of sensitive data items that happened. This displays the itemsets that should have been concealed in the sanitised database but were not by the sanitization method. Figure 1f depicts the Sensitive Items that aren't meant to be hidden and must have existed in the sanitization database, but the sanitization algorithm was unable to include all Sis values. The region formed as a result of the production of extra rules that were not available in the original database is seen in Figure 1g.

IV. Provocation for hiding Association rules

Hiding Association Rules poses a number of difficulties, the majority of which are NP-hard problems. This section goes over some of the more common issues [4].

- Failure to Hide: Some sensitive rules were discovered in a sanitised data base that was supposed to be kept concealed. In GA and PSO techniques, this is accomplished through increased computational complexity.
- Rules lost: Some of the non-sensitive frequent item sets are concealed and will not be present in the sanitised database, because the sanitization method failed.
- Artificial rules: In a sanitised database, the effect of sanitization can result in the development of some ghost rules or artificial rules.
- Differences in the database: The ratio between the sanitised database D and the original database must be kept as low as possible, that is, the original database's transactions must be deleted or changed as little as possible. A database similarity ratio of higher than 90 is required.
- Estimated difficulty: The sanitization procedure must be able to give the output database in the shortest amount of time possible, as opposed to other techniques. More efficient the algorithm is, the less complex it is.
- Precision: When data accuracy deteriorates, the apprehension collected from the sanitised database becomes meaningless. The precision is inversely related to the database differences.

V. Defining the problem

In a given a database D, which contains a collection of transactions T1, T2..., Tn, each transaction consists of a set of items I1, I2..., In. Finding the set of rules is the main goal of the association rule mining method, which includes both a priori and FP-Growth algorithms. Based on MST and MCT, these rules are categorised into sensitive and non-sensitive rules. Hiding association rules is mainly associated with hiding association rules which are sensitive. It uses the set of data as input and outputs the set of data which are unrealized. The association rules which are sensitive are hidden, and ghost and lost rules are minimized, thanks to the un-realization dataset used as input to the association rules mining method.

VI. Proposed Hybrid Particle Swarm Optimization Algorithm

This work suggests a hybrid particle swarm optimization approach for proper hiding of association rule that combines the traits of the original particle swarm optimization and other listed optimization methods. Modified C4ARH, Perturbed Dataset, Modified PSO(), ACO() and ARM() are some of the components in this algorithm [3]. The objective function values, as well as their differences and side effect data, are computed by Modified C4ARH. The entire set of data is handled by the Pertubed Dataset() algorithm, which results in the standard set of data S and perturbed set of data P. The values for pbest and gbest are found using the Modified PSO() algorithm. The association rules which are sensitive are defined by gbest, whereas the association rules which are non-sensitive are being defined by pbest.

The association rules were generated using the ARM() technique. Ant Colony Optimization Algorithm is a presumed technique which is used to solve the problems of estimation which can be minimized to finding accurate paths through graphs [4-5].

VII. Hybrid Algorithm

```
Input Original Dataset D, MST, and MCT;
Output A Sanitized Dataset R; begin
do Dataset D
Find item count; call ARM();
end;
begin ARM(D)
Generate association rules
count = number of association rules
Return O, S;
end
begin
for (O, S) apply PSO
for each rule R;
calculate best() for all O,S;
  COA4ARH()
  end
begin
best(R,MST,MCT)
if(R > MST && R > MCT)
```

```
gbest = R;
gbest[]++;
return gbest;
end
begin
COA4ARH() for all O,S
remove gbest;
compute objective function;
R = compare the objective function values of O with S;
R is the dataset with sensitivity;
end
begin
  ACO()
begin pheromone trails;
while (termination condition not satisfied)
do
construct candidate conformations;
perform local search;
update pheromone values;
end
end
```

VIII. Experimental Results

All the experiments are performed on Windows operating system with i3 processor, 16 GB RAM. Corresponding Matlab programs along with necessary datasets are used to test the performance of the association rule hiding algorithm.

There are a variety of well-known techniques for optimization accessible, however Hybrid Particle Swarm Optimization (PSO) and Hybrid Ant Colony Optimization (ACO) are the most used ones. Although we employ a variety of strategies to improve individual advantages, these two optimization approaches outperform them all due to its total effectiveness. In case of best cost function, these two approaches outperform Cuckoo and Generic Optimization techniques. It does not believe in showing a significant change, but it does show a significant decrease in the percentage of optimization that can be viewed using the graphs.

However, in the case of side effect factors, there is a significant difference, which aids in the elimination of side effects to a higher extent than other optimization strategies. This is a big benefit of this technique, and there is also generation of lost rule, which deals with missing values and aids in the synthesis of lost rules; it has a remarkable 98 percent accuracy rate that no other optimization technique can match. This is an important component in deciding whether or not to use Hybrid optimization.

Figure 1 and Figure 2 shows PSO and ACO in terms of cost value.

Figure 3 and 4 shows modified PSO and modified ACO in terms of best cost function.

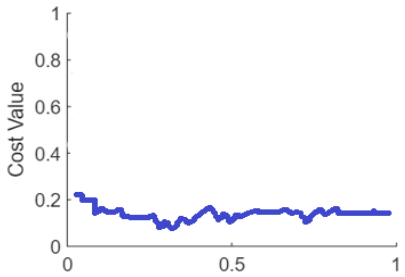

Figure 1: *Experimental result of PSO algorithm*

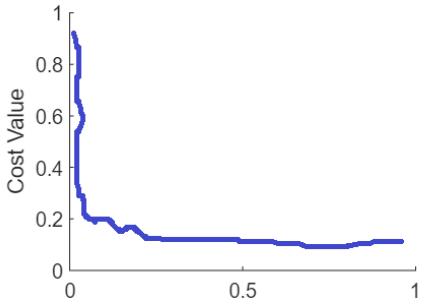

Figure 2: *Experimental result of ACO algorithm*

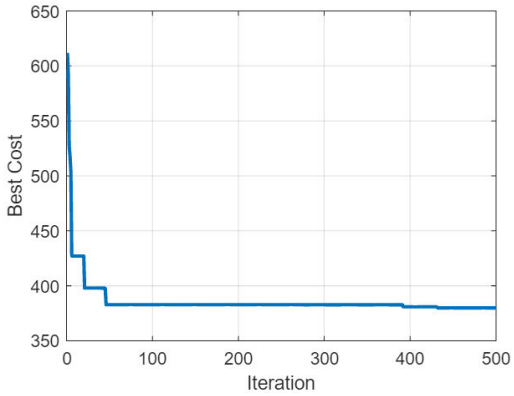

Figure 3: *Experimental result of Modified PSO algorithm*

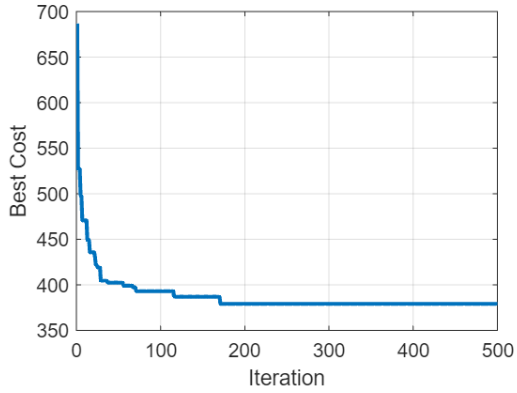

Figure 4: *Experimental result of Modified ACO algorithm*

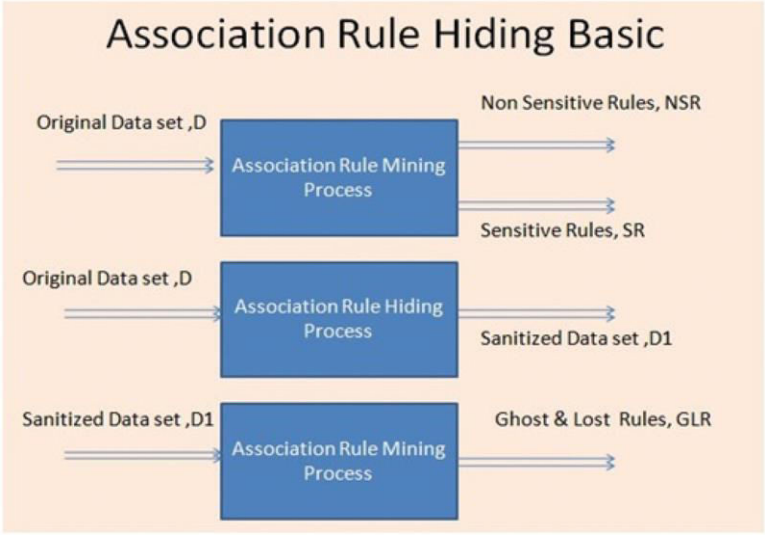

Figure 5: *Framework for Association Rule Hiding*

X. Conclusion

In this article, a hybrid optimization algorithm has been proposed to improve the performance of the mentioned parameters like cost value and best cost function. Experimentally, the result showed better performance than all the other traditional algorithms.

References

[1] Jerry Chun-Wei Lin a,n , Qiankun Liu a , Philippe Fournier-Viger b , Tzung-Pei Hong c , Miroslav Voznak d , Justin Zhan (2021) A sanitization approach for hiding sensitive itemsets based on particle swarm optimization. *Engineering Applications of Artificial Intelligence 53* (2016) 1–18 www.elsevier.com/locate/engappai

[2] Lin, C.-W., Hong, T.-P., & Hsu, H.-C. (2014). Reducing side effects of hiding sensitive item sets in privacy preserving data mining. *The Scientific World Journal*, *2014*, 1–12. https://doi.org/10.1155/2014/235837

[3] Lin, J. C. W., Liu, Q., Fournier-Viger, P., Hong, T. P., Voznak, M., & Zhan, J. (2016). A sanitization approach sfor hiding sensitive itemsets based on particle swarm optimization. *Engineering Applications of Artificial Intelligence*, *53*, 1–18. https://doi.org/10.1016/j. engappai.2016.03.007

[4] Satyanarayana Murthy, T., Gopalan, N. P., & Venkateswarlu, Y. (2018). An efficient method for hiding association rules with additional parameter metrics. *International Electronic Journal of Pure and Applied Mathematics*, *118*(7), 285–290.

[5] Satyanarayana Murthy, T., Gopalan, N. P. (2018). A novel algorithm for association rule hiding. *International Journal of Information Engineering and Electronic Business*, *10*(3), 45–50. https://doi.org/10.5815/ijieeb.2018.03.06.