Computers, Materials & Continua
DOI: 10.32604/cmc.2023.033182
Article

An Adaptive Privacy Preserving Framework for Distributed Association Rule Mining in Healthcare Databases

Hasanien K. Kuba[1], Mustafa A. Azzawi[2], Saad M. Darwish[3,*], Oday A. Hassen[4] and Ansam A. Abdulhussein[5]

[1]University of Information Technology and Communication (UOITC), Baghdad, 10081, Iraq
[2]Accenture, Ling Karan TRX, Kuala Lumpur, 55188, Malaysia
[3]Department of Information Technology, Institute of Graduate Studies and Research, University of Alexandria, 21526, Egypt
[4]Ministry of Education, Wasit Education Directorate, Kut, 52001, Iraq
[5]Information Technology Center, Iraqi Commission for Computers and Informatics, Baghdad, 10081, Iraq
*Corresponding Author: Saad M. Darwish. Email: saad.darwish@alexu.edu.eg
Received: 10 June 2022; Accepted: 02 September 2022

Abstract: It is crucial, while using healthcare data, to assess the advantages of data privacy against the possible drawbacks. Data from several sources must be combined for use in many data mining applications. The medical practitioner may use the results of association rule mining performed on this aggregated data to better personalize patient care and implement preventive measures. Historically, numerous heuristics (e.g., greedy search) and metaheuristics-based techniques (e.g., evolutionary algorithm) have been created for the positive association rule in privacy preserving data mining (PPDM). When it comes to connecting seemingly unrelated diseases and drugs, negative association rules may be more informative than their positive counterparts. It is well-known that during negative association rules mining, a large number of uninteresting rules are formed, making this a difficult problem to tackle. In this research, we offer an adaptive method for negative association rule mining in vertically partitioned healthcare datasets that respects users' privacy. The applied approach dynamically determines the transactions to be interrupted for information hiding, as opposed to predefining them. This study introduces a novel method for addressing the problem of negative association rules in healthcare data mining, one that is based on the Tabu-genetic optimization paradigm. Tabu search is advantageous since it removes a huge number of unnecessary rules and item sets. Experiments using benchmark healthcare datasets prove that the discussed scheme outperforms state-of-the-art solutions in terms of decreasing side effects and data distortions, as measured by the indicator of hiding failure.

Keywords: Distributed data mining; evolutionary computation; sanitization process; healthcare informatics

1 Introduction

Many hospitals and other medical facilities utilize electronic health records (EHRs) to track patient information and streamline administrative tasks in an attempt to better serve their patients. When properly implemented, EHRs aid in the timely and precise diagnosis of illnesses [1]. While patients are obviously the primary beneficiaries of data and EHR integration, healthcare researchers stand to gain as well. The dissemination of HER data is crucial for improving the precision of medical research [2]. Medical researchers now have a new avenue to explore in their quest to improve patient care and diagnose illnesses at their earliest stages via the use of data mining applied to healthcare data. Integrating data from several sources is a crucial part of many data mining applications [3–5]. Patients' privacy has been invaded by the disclosure of this information.

Different approaches to protecting privacy put forward in the literature, including anonymization-based strategies and cryptographic approaches [1,3]. Due to its cheaper communication and computing costs compared to its cryptographic competitors, anonymization is commonly utilized by researchers. The loss of information is a major concern in anonymization-based methods [6–8]. It's possible to classify most approaches into two broad categories: those that protect information during mining, and those that do the same for the outcomes of such an endeavor. To create sanitized datasets that may be securely shared with others, the first group comprises techniques like perturbation, sampling, and modification. The second set comprises of methods for hiding private information uncovered by data mining algorithms, such as decreasing the performance of classifiers in such jobs [8].

The privacy-preserving data mining (PPDM) approach is especially useful in healthcare systems for limiting the disclosure of sensitive patient information. As a result, it's possible to learn about and develop treatments for fatal diseases by analyzing massive data sets from medical research systems without violating patients' privacy [3,4,9–11]. Data mining techniques (database sanitization) that respect users' privacy may be broken down into two groups [9]: (1) algorithms that manage the hiding process by rule support or confidence, and (2) algorithms that modify raw data in such a way as to obscure or distort its original values. The collection of rules that may be mined from the original database may be changed throughout the alteration process, either because benign rules are hidden (lost rules) or because new rules are introduced (not supported by the original database) (ghost rules).

For vertically partitioned data, each participant has its own schema and stores the same set of entities' data. This idea is called vertically partitioned data mining to protect privacy. In a situation like this, it can be hard to make sure that everyone's data is kept private, since sharing private data can lead to privacy problems. As seen in Fig. 1, a shared ID between health examination data and outpatient data may be used to correlate abnormal test findings with specific illnesses. Medical researchers want to figure out how to use the combined data from the health examination record and the outpatient record to find relationships between the data. Association rules found through this collaboration help to find links between some diseases and characteristics of the patients. But sharing private information about a patient is against the law. So, medical researchers have paid a lot of attention to privacy-preserving association rule mining in vertically partitioned healthcare data [12,13]. Privacy-preserving association rule mining in vertically partitioned healthcare data is the focus of the work reported in this article. Because of the sensitive nature of some of the information being shared, it may be difficult to implement such a system safely.

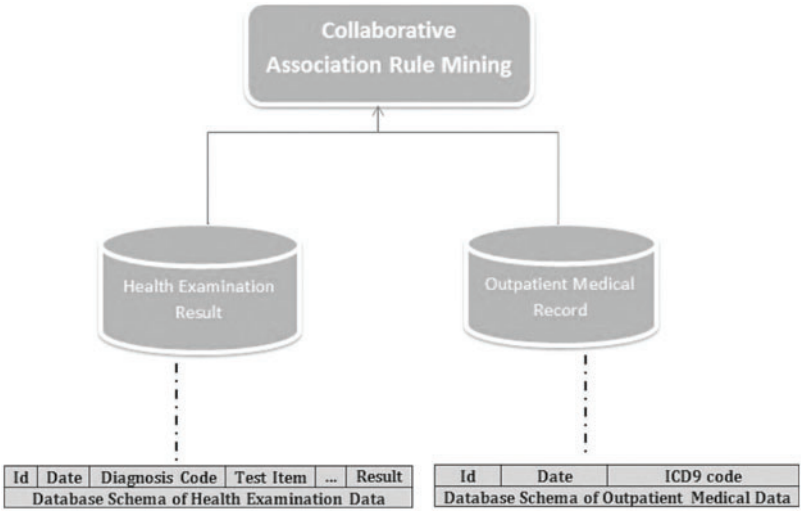

Figure 1: Vertical partition data model [13]

To find the solutions, GAs use a group of points rather than just one. This method is efficient and easy to implement computationally. Each string is treated as a discrete entity in Tabu Search's (TS) solution space. Through local solution enhancement and the ability to escape starving local minima, TS guides iterations from one neighborhood point to another. Global combinatorial optimization problems are more likely to have a workable solution when GA and TS are combined to use each method's strengths. GA starts with a pool of potential solutions and uses a hybrid search strategy to come up with a new set of solutions. Each group of TS's original solutions is developed by means of a local search. In order to maintain the same evolution, GA employs TS's enhanced solution [14].

In healthcare data analysis, negative association criteria may be used to discover coexisting symptoms or medications that have positive interactions with one another. Negative association rule mining is difficult because it requires taking into account the basic differences between positive and negative association rule mining. Finding and filtering the negative association rules are the two main issues that must be addressed by researchers when mining negative association rules. The PPDM issue with negative association rules in healthcare data has never been addressed by using meta-heuristics optimization techniques, despite their huge success [9–11].

In this research, we use a refined approach to the issue of negative association rules in healthcare data mining, one that is based on the Genetic Tabu (GT) optimization framework. This unique methodology uses GA and TS to generate "meta-heuristic" negative association rules, which are more precise and use less memory than previous methods. The Tabu search is helpful since it may get rid of a lot of unnecessary rules and item sets. Without having to predefine a method, sensitive data may be hidden dynamically by using a perturbation strategy based on a delete operation. The algorithm's most notable contributions are as follows: (1) this is the first effort to use a Tabu-Genetic strategy to address the PPDM issue for negative association rules in vertically partitioned healthcare datasets. (2) The approach used dynamically selects which transactions need to be interrupted in order to keep certain pieces of information hidden, as opposed to predefining such transactions in advance.

The article then proceeds as follows. The second section focuses on related studies. The third section explains the suggested strategy for sanitizing distributed healthcare data mining. In Section 4, the experimental results are provided. Section 5 contains a conclusion and a discussion of future work.

2 Related Work

There is a compromise between data privacy and data value in existing data transformation technologies for anonymizing healthcare process data. The academic community has made substantial use of anonymity, data masking, data perturbation, and cryptography to protect private data. There are drawbacks to every method, such as data loss, privacy leaks, and poor analysis results, as discussed in [3,6]. Data mining approaches for protecting personal information might be either general or specific [15,16]. As stated in [16], PPDM-based data linking is achieved by the application of a machine learning technique. In [17], the authors examined the use of hierarchical categorization strategies to solve the problem of PPDM. To reduce trust or support for secret data, Dasseni et al. [18] devised a method based on the Hamming distance. Using a heuristic method, Oliveira et al. [19] developed many sanitization approaches to hide frequently occurring itemsets. Using noise addition, Islam et al. [20] suggested a method to shield and hide personal information while keeping data quality high. In [21], the authors presented a sanitization strategy in which the most frequently utilized item in the transaction serves as the target. For the sake of establishing a fine balance between privacy and communication, a threshold for sharing is also set.

To track how the perturbation procedure is affecting the sanitized data, the boundary of non-sensitive itemsets was suggested by Sun et al. [22]. This was also achieved by the authors in [23], who used the MaxMin technique. Choosing the least disruptive transaction for modification at each step is how the quality of the sanitized database is kept up. Amiri [24] presented three heuristic strategies—aggregate, disaggregate, and hybrid—for hiding sensitive information. The aggregate approach reduces the need for support for a sensitive set of items by filtering out unimportant transactions from the database. Through the elimination of individual items from the list, the disaggregated method minimizes the quantity of support for delicate items. The hybrid strategy is a fusion of two different approaches. Two methods for hiding relevant association rules were devised by Wang et al. [25]. Predicted items are provided openly rather than obscuring crucial association rules. Thus, the resulting database is unusable for mining informative association rules whose antecedents include the expected items. The mechanism for checking up on all possibly relevant changes was devised by Wu et al. [26]. In order to hide any sensitive rules, the database is modified invisibly, with as few unintended consequences as the template will allow. To keep private sets secure, Gkoulalas et al. [27] suggested a strategy based on boundaries. Support for sensitive items may be limited by increasing the size of the original database.

The greedy approximation method and the greedy exhaustion approach were described by Wu et al. [28] with the goal of protecting private association rules. Both methods hide the essential rules by manipulating the database in different ways. Cheng et al. [29] examined the advantages and drawbacks of positive and negative border regulations with the goal of protecting private association rules. Both methods hide the essential rules by manipulating the database in different ways. The least important transaction is changed such that the sensitive association rules are hidden. The term frequency-inverse document frequency approach was developed by Hong et al. [30] as a method for decreasing support for susceptible item sets. In [31], the authors developed a safe system for using

evolutionary algorithms to discover a better set of rules without revealing sensitive information. Similar ideas were presented in [32], although with new chromosomal representations and fitness measures. While the aforementioned algorithms do a good job of choosing which transactions should be deleted, further work is needed in the form of pre-defined weights for side effects, which has the potential to significantly alter the results of the suggested systems. It's a productive method. Once the item has been changed in a random way, the sanitization process will randomly make up the lost and ghost rules.

To address the aforementioned issues, researchers have devised a multi-objective optimization (e.g., the non-dominated sorting genetic algorithm II (NSGA II)) strategy by factoring in both data and knowledge distortion. Despite the fact that this method incorporates several objectives, it risks providing insufficient data for making decisions by eliminating features from databases in a straightforward manner. This assertion is not supported by the sequential dataset [17]. The authors in [33] presented a hierarchical-cluster method for protecting private data sets by using the multi-objective particle swarm optimization (MOPSO) algorithm. Partial transactions may be created, but this might lead to an incorrect conclusion, particularly when dealing with hospital diagnostics. To preserve privacy and facilitate discovery, researchers have recently turned to a deep reinforcement learning approach for database sanitization [34–36]. The interested reader is directed to the recent studies in [37–41] for more reading in this area.

There are discussions of privacy-preserving association rule mining in vertically partitioned data in [12,13,42]. Computing the dot product was one of the methods used. This method, however, is impractical in real-world applications since, it requires a trustworthy third party to send private keys via an encrypted channel. To find the link between sickness and abnormal test results, mining association rules on medical exam data and outpatient records was suggested in different approaches. However, this method does not take into account the sensitive nature of patient confidentiality when integrating data from medical examinations with outpatient medical records. Patients' confidentiality must be protected during the mining of association rules, since this is required by law or regulation.

In conclusion, numerous evolutionary methods have been developed to address the PPDM issue in vertically partitioned healthcare databases. However, these algorithms still rely heavily on sanitization approaches based on data cleansing for reliable association rules. By choosing itemsets using the Tabu-genetic technique, the method suggested in this paper seeks to solve the limitations of current sanitization algorithms for negative association rule mining. The proposed approach minimizes the amount of lost or ghost rules while hiding rules without altering the original database in any way (no fake transactions).

3 Distributed Sanitization Algorithm for Negative Association Rules

This study proposes a method for cooperatively computing association rule mining on a combined database, with two participants A (an EHR with medical examination data of patients) and B (an EHR with outpatient medical records). The database D is vertically partitioned in such a way that $D = \{D_A \cup D_B\}$ with common Id. For problem formulation, see [12–14,42] for more details. Database D is converted into the Boolean form by representing the present and absence with 1 and 0. Frequency of an itemset is the number of transactions where the values of all the attributes in the itemset are 1. Assume

the participant A has l attributes $\langle a_1, a_2, \ldots, a_l\rangle$ and participant B has m attributes $\langle b_1, b_2, \ldots, b_m\rangle$. We are interested to compute the frequency of $(k = l+m)$-itemset $\langle a_1, a_2, \ldots, a_l, b_1, b_2, \ldots, b_m\rangle$. Each element of $\vec{x}$ *and* $\vec{y}$ is calculated as $x_i = \prod_{j=1}^{l} a_j$ and $y_i = \prod_{j=1}^{m} b_j$. The dot product of $\vec{x}$ *and* $\vec{y}$ gives the frequency of k-itemset. In this case, the vector $\vec{x} = \{x_1, x_2, \ldots, x_n\}$ and $\vec{y} = \{y_1, y_2, \ldots, y_n\}$ represent the columns in the database, x_i is 1 if the transaction i has the value 1 for the attribute $\vec{x}$. The dot product of the two vectors $\vec{x}$ and $\vec{y}$ with cardinality n is computes as:

$$\vec{x}.\vec{y} = \sum_{i=1}^{n} x_i.y_i \tag{1}$$

The recommended method for sanitizing vertically partitioned healthcare databases is shown in Fig. 2. In contrast to other attempts, which also hid particular items rather than rules, as part of the suggested sanitization strategy, a genetic algorithm is used to choose the most effective items to modify in order to hide potentially sensitive negative association rules. But when itemsets are hidden, they won't show up in any rules with a confidence level higher than the threshold, whether or not those rules are sensitive. This study expands upon our earlier work [14] by combining GA and TS to address a mining negative association rule problem. Utilizing efficient chromosomal representation and neighborhood strategies, this method outperforms alternatives. Choosing a suitable fitness function in GA is the most important part of this work. Algorithm 1 illustrates the process of identifying the set of frequent negative items in a vertically partitioned database.

Algorithm 1: Distributed negative association rule mining in vertically partitioned database

$L_1 = Find\ infrequent_1 - itemset\ (D)$
for $(k = 2; L_{K-1} \neq \varnothing; k++)$
$C_k = apriori\_gen\ (L_{k-1})$
for *each candidate* $c \in C_k$ ***do***
  If *all the attributes of c are A or B alone*
    c_{count} *is calculated by each participant independently*
  else
    Let c contains p attributes of A and q attributes of B
    Participant A compute the $\vec{x}$ *as* $x_i = \prod_{j=1}^{p} a_j$
    Participant B compute the $\vec{y}$ *as* $y_i = \prod_{j=1}^{q} b_j$
    Collaboratively find the $C_{count} = \vec{x}.\vec{y} = \sum_{i=1}^{n} x_i.y_i$
  end if
  $L_k = \{c \in C_k | c_{count} \geq min\_sup\}$
Return $L_k = \cup_k L_k$

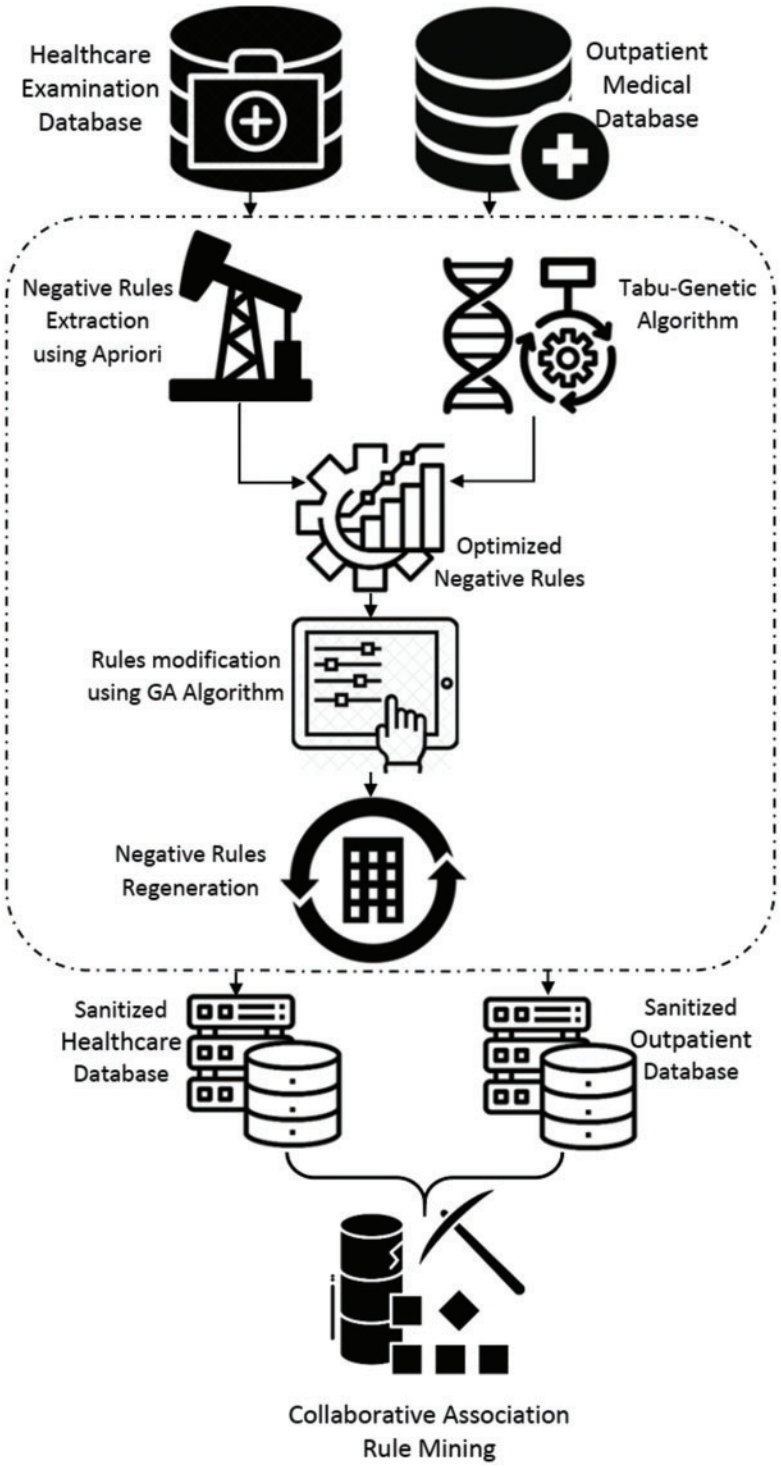

Figure 2: A genetic-Tabu heuristic search-based distributed sanitization technique for negative association rules

3.1 Extracting Optimized Negative Rules

- Transform a dataset comprising a list of attributes and records into a numerical format (coding).
- Algorithm 1 demonstrates how to use the Apriori approach to build all the possible infrequent item sets for a collection of items with any length. The references [12–14,43–46] provide more information.
- Create preliminary negative association rules using the Apriori method, starting with sets of items that occur infrequently.
- The fitness function for several interesting negative association rules should be defined.
- Because of this, the GA is used in the construction of chromosomes that are in accordance with the negative association rules and in the subsequent determination of their respective fitness values. Construct negative association rules based on the average fitness value of each chromosome.
- Using the solutions it has already generated, GA creates new solutions for each hybrid search. Tabu search does a local search to enhance each iteration of solutions. After then, GA employs TS's superior solution to continue its parallel growth (see Fig. 3). Tabu Search treats each string as an own point in the space of possible solutions. By improving the solution locally and avoiding poor local minima, TS directs the recursive process as it moves from one nearby place to the next (see Fig. 4). Because of their complementary strengths, GA and TS have a good shot at solving global combinatorial optimization problems.
- After crossover and mutation, the fitness value should be reset, the final negative rules should be calculated, and any remaining offspring chromosomes should be modified. For this study, we utilized the same measure as in [8] to find novel negative association rules. Here, the system employs a rule encoding format identical to that of [47]. For further information, see [14,46,47].

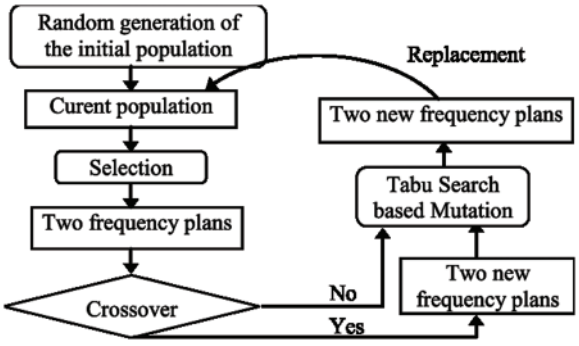

Figure 3: A genetic-Tabu search procedure

3.2 Genetic Algorithm-Based Data Sanitization

The proposed structure for hiding sensitive negative association rules is particularly helpful at this stage. The system tries to hide the rules created in the previous phase by diminishing their confidence to below a user-defined threshold, which it does by boosting support for the antecedent and decreasing support for the consequent via the replacement of 1 s for 0 s in the transactions [18,29,42]. For the technique to work, a substantial amount of computational resources will be required, since it will need to update all sensitive item sets associated to sensitive rules in all database transactions. In the current research, we use GA to choose the best sets of items to modify in order to solve the

aforementioned problem. This means we may safely continue without making many changes to the data in our databases. The elimination of this step allows us to increase the rate of sanitization while also decreasing the number of edits needed for the whole hiding procedure. Furthermore, the method may be used with either small or large data sets. Each chromosome represents a single transaction, and it records whether or not a certain item was present at that time. Several factors and methods determine a chromosome's viability. Numerous chromosomes exist in each population, and only the most advantageous ones are used to create the next generation. Large numbers of random exchanges make up the initial population. Population success at increasing survival fitness will determine the character of the next generation.

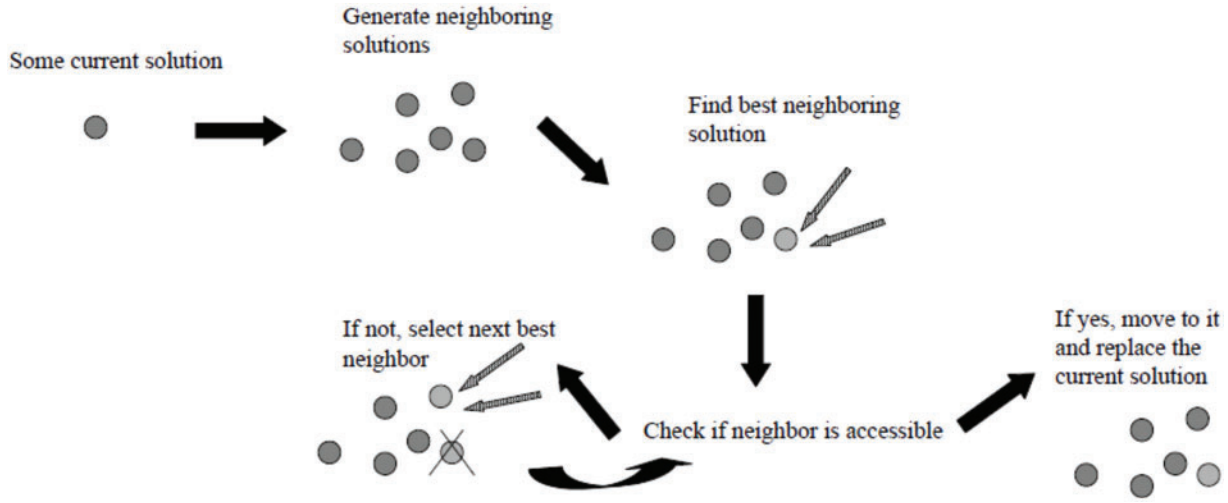

Figure 4: The basic steps of a Tabu search

To minimize the occurrence of lost and ghost rules, the proposed system makes use of two separate fitness functions that adjust only transactions involving a high number of sensitive items and a low number of non-sensitive items. Whatever the scenario, the transaction with the lowest fitness score will be altered. As stated in [48], the first fitness function f_{V_1} is as follows:

$$f_{V_1} = \frac{X_r + Y_r}{2} \tag{2}$$

$$X_r = \sum_{i=1}^{n} (I_i = 1), \quad Y_r = (S_r \text{ in } T_r) \tag{3}$$

$S_r \in I$ defines the set of sensitive items, T_r is the transactions set, n represents the number of items in each transaction, r describes transaction's number, and v is a set of identifiers for elements of $f, f_v = \{f_1, f_2, \ldots, f_n\}$. The second one f_{V_2} is calculated using a weighted sum function, as [49]:

$$W1 * C_1 + W2 * \left(\frac{1}{C_2}\right) \tag{4}$$

$$\forall C_1 \in T_r,\ C_1 = \frac{1}{\sum_1^n Count\ (S_r)} \quad \text{in } T_r + \sum_{i=1}^{n} I_i = 1 \tag{5}$$

$$\forall C_2 \in T_r,\ C_2 = \frac{1}{\sum_1^n Count\ (S_r)} \quad \text{in } T_r + \sum_{i=1}^{n} I_i = 0 \tag{6}$$

$$W1 + W2 = 1 \left(W1 = W2 = \frac{1}{2} \right) \tag{7}$$

$W1$ and $W2$ are weights. By swapping out certain transactions for their offspring that have the most accessible data items, the system reduces the amount of lost rules and ghost rules [46–49]. Additional details on the use of machine learning methods for privacy-preserving data mining in IoT-based healthcare applications and vehicular cloud network settings are provided in the most recent references [50–59]. Furthermore, Refs. [60,61] provide more information regarding how rule based models can be employed for enhancing data privacy. For more information regarding the benchmark medical datasets, readers can refer to Refs. [62–65].

4 Experimental Results

The approach discussed in [66] for preserving privacy of association rule mining in healthcare databases does not take into consideration how to preserve privacy in the case of distributed data mining. The proposed method protects patient privacy without compromising the efficacy of vertically partitioned healthcare databases. Patient medical examination data and outpatient medical data are analyzed using our suggested method for determining a correlation between the illness and the test findings. Regarding the results related to extracting negative correlation rules for each database, based on the same configurations applied for GA and TS and the same measures, the same results were obtained as in [66].

To that end, we've designed a series of experiments to explore the relationship between the number of transactions and the number of hidden negative-sensitive and artificial rules in a sanitized medical examination database. In this experiment, $Min_{sup} = 25\%\%$ and $Min_{conf} = 58\%$, while $Min_{sen-conf}$ is set at 60%, 70%, and 80% for 500, 1000, 2000, 3500, and 5000 transactions, respectively. The negative impacts of the hiding technique on the restricting mode fitness function f_{V_1} and the distorting mode fitness function f_{V_2} are described in Tables 1 and 2, respectively. From what can be seen in both tables, the loss of non-sensitive rules is rather small, increasing with the number of database operation transactions and decreasing with the size of the set of sensitive rules $|R_{sen}|$. It has been estimated that the proposed method has a hiding failure rate of 0%, meaning that no sensitive rules will be revealed to the outside world. Protecting against sensitive rules is always precise.

There seems to be a clear correlation between the number of database transactions and the rate at which new rules are created, as seen in Table 2 (distortion mode). We found that more frequent item sets are introduced when more rules are hidden, and this in turn leads to the generation of more new rules. Once the rules are hidden, no more rules are mined from the database using the restriction mode modification. In other words, when it comes to reducing ghost rules, using f_{V_1} yields better performance. However, the proposed method only modifies a subset of transactions at a time, picking those that meet the maximum modification rules' criteria for changes.

Table 1: Performance Evaluation for $\mathbf{f}_{\mathbf{v_1}}$(%) for sanitized medical examination database

$Min_{sen\text{-}conf}$		60%			70%			80%	
No. of Transactions	LR	HF	AR	LR	HF	AR	LR	HF	AR
1000	0	0	1.17	0	0	1.00	0	0	0.73
2000	0	0	1.34	0	0	1.15	0	0	1.02
3500	0	0	1.68	0	0	1.36	0	0	1.33
5000	0	0	2.09	0	0	2.05	0	0	2.01

Table 2: Performance Evaluation for $\mathbf{f}_{\mathbf{v_2}}$ (%) for sanitized medical examination database

$Min_{sen\text{-}conf}$		60%			70%			80%	
No. of Transactions	LR	HF	AR	LR	HF	AR	LR	HF	AR
1000	0	0.08	1.25	0	0.006	1.03	0	0.003	0.90
2000	0	0.012	1.31	0	0.011	1.21	0	0.08	1.02
3000	0	0.03	1.61	0	0.012	1.34	0	0.010	1.45
5000	0	0.04	2.08	0	0.017	2.09	0	0.013	2.05

In the next set of experiments, we evaluate our method against another that, like the one given here, is concerned with hiding association rules, instead of rules, it uses item sets to do this [48]. In Table 3, we can see the average negative impacts produced by both systems using the sanitized outpatient medical data. Table results show that the proposed system was missing a few rules. Neither method resulted in the generation of any "ghost rules" when the required rules were hidden, and both avoided unintended effects when hiding rules. The examples demonstrate that the suggested approach produces fewer undesirable results and less skewed data than the competing approach. Thus, our method effectively hid a significant number of sensitive association rules in the original database without compromising the integrity of the data mining results.

Table 3: Comparative results (%) for sanitized outpatient medical data

Algorithm	*LR*	*HF*	*AR*
Proposed model	1.98	0	0
Comparative model [48]	6.67	0.17	0

4.1 Limitations of the Proposed Model

Using the proposed method for Tabu search operations takes more time. Furthermore, the performance of the proposed model depends mainly on the chosen parameters for both GA and TS. The process of selecting these parameters must be done through an optimization procedure and not manually.

5 Conclusions

The focus of this research is on the privacy concerns raised by data mining tools for distributed healthcare datasets. To cover up potentially damaging association rules, we developed a genetic Tabu optimization strategy that makes use of a heuristic based on distortion and restriction processes. The proposed approach reduces confidence in sensitive rules. Data sanitization aims to make the fewest database changes and miss the fewest non-sensitive association rules. The proposed algorithm combines Apriori with genetic-Tabu. The proposed method employs negative interestingness to describe and provide an explanation for the effectiveness of negative association rules.

The approach reduces mining's search space by employing genetic-Tabu. The method employs a simple heuristic approach to determine which transactions and items need to be sanitized, a genetic algorithm to influence the victim's item selection, and rules in place of items to hide sensitive information. Several healthcare datasets have been analyzed to test the fitness function's robustness in the face of modifications made to the source data. The simulation results show that the recommended approach has stronger support and confidence while needing less processing time. Future research will study privacy-preserving data transformation approaches, log extensions, and process mining algorithms, as well as how they influence healthcare logs. The recommended strategy will be evaluated on vertically segmented healthcare datasets with varied features to confirm its effectiveness.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] D. Kumari, Y. Vineela, T. Krishna and B. Kumar, "Analyzing and performing privacy preserving data mining on medical databases," *Indian Journal of Science and Technology*, vol. 9, no. 17, pp. 1–9, 2016.

[2] N. Domadiya and U. Rao, "Improving healthcare services using source anonymous scheme with privacy preserving distributed healthcare data collection and mining," *Computing,* vol. 103, no. 1, pp. 155–177, 2021.

[3] L. Abuwardih, W. Shatnawi and A. Aleroud, "Privacy preserving data mining on published data in healthcare: A survey," in *Proc. of the Int. Conf. on Computer Science and Information Technology*, USA, pp. 1–6, 2016.

[4] A. Pika, T. Wynn, S. Budiono, A. Hofstede, W. Aalst *et al.,* "Towards privacy-preserving process mining in healthcare," in *Proc. of the Int. Conf. on Business Process Management*, Austria, pp. 483–495, 2019.

[5] A. Rashid and N. Yasin, "Sharing healthcare information based on privacy preservation," *Scientific Research and Essays*, vol. 10, no. 5, pp. 184–195, 2015.

[6] M. Hassan, M. Butt and M. Zaman, "Privacy preserving data mining for healthcare record: A survey of algorithms," *International Journal of Trend in Scientific Research and Development*, vol. 2, no. 1, pp. 1176–1184, 2017.

[7] A. Pika, T. Wynn and S. Budiono, "Privacy-preserving process mining in healthcare," *International Journal of Environmental Research and Public Health*, vol. 17, no. 5, pp. 1–28, 2020.

[8] T. Gal, G. Kovacs and Z. Kardkovacs, "Survey on privacy preserving data mining techniques in health care databases," *Acta Universitatis Sapientiae and Informatica*, vol. 6, no. 1, pp. 33–55, 2014.

[9] S. Darwish, M. Madbouly and M. El-Hakeem, "A database sanitizing algorithm for hiding sensitive multi-level association rule mining," *International Journal of Computer and Communication Engineering*, vol. 3, no. 4, pp. 285, 2014.

[10] M. Dehkordi, K. Badie and A. Zadeh, "A novel method for privacy preserving in association rule based on genetic algorithms," *Journal of Software*, vol. 4, no. 6, pp. 555–562, 2009.
[11] R. Crawford, M. Bishop, B. Bhumiratana, L. Clark and K. Levitt "Sanitization models and their limitations," in *Proc. of the Workshop on New Security Paradigms*, USA, pp. 41–56, 2006.
[12] J. Liu, Y. Tian, Y. Zhou, Y. Xiao and N. Ansari, "Privacy preserving distributed data mining based on secure multi-party computation," *Computer Communications*, vol. 153, pp. 208–216, 2020.
[13] N. Domadiya and U. Rao, "Privacy preserving distributed association rule mining approach on vertically partitioned healthcare data," *Procedia Computer Science*, vol. 148, pp. 303–312, 2019.
[14] H. Waguih, S. Darwish and M. Osman, "Mining interesting positive and negative association rule based on genetic tabu heuristic search," *Journal of Theoretical and Applied Information Technology*, vol. 96, no. 23, pp. 7834–7845, 2018.
[15] J. Lin, J. Wu, P. Fournier-Viger, Y. Djenouri, C. Chen *et al.,* "A sanitization approach to secure shared data in an IoT environment," *IEEE Access*, vol. 7, pp. 25359–25368, 2019.
[16] D. Toshniwal, "Privacy preserving data mining techniques for hiding sensitive data: A step towards open data," in *Data Science Landscape*, Singapore, pp. 205–212, Springer, 2018.
[17] V. Verykios, E. Bertino, I. Fovino, L. Provenza, Y. Saygin *et al.,* "State-of-the-art in privacy preserving data mining," *ACM Sigmod Record*, vol. 33, no. 1, pp. 50–57, 2004.
[18] E. Dasseni, V. Verykios, A. Elmagarmid and E. Bertino, "Hiding association rules by using confidence and support," in *Proc. of the Int. Workshop on Information Hiding*, USA, pp. 369–383, 2001.
[19] R. Oliveira and O. Zaiane, "Privacy preserving frequent itemset mining," in *Proc. of the IEEE Int. Conf. on Privacy, Secure Data Mining*, Japan, pp. 43–54, 2002.
[20] M. Islam and L. Brankovic, "Privacy preserving data mining: A noise addition framework using a novel clustering technique," *Knowledge Based System*, vol. 24, no. 8, pp. 1214–1223, 2011.
[21] X. Liu, S. Wen and W. Zuo, "Effective sanitization approaches to protect sensitive knowledge in high-utility itemset mining," *Applied Intelligence*, vol. 50, no. 1, pp. 169–191, 2020.
[22] X. Sun and S. Philip, "Hiding sensitive frequent itemsets by a border-based approach," *Journal of Computing Science and Engineering*, vol. 1, no. 1, pp. 74–94, 2007.
[23] M. George and S. Verykios, "A MaxMin approach for hiding frequent itemsets," *Data & Knowledge Engineering*, vol. 65, no. 1, pp. 75–89, 2008.
[24] A. Amiri, "Dare to share: Protecting sensitive knowledge with data sanitization," *Decision Support Systems*, vol. 43, no. 1, pp. 181–191, 2007.
[25] S. Wang, B. Parikh and A. Jafari, "Hiding informative association rule sets," *Expert Systems with Applications*, vol. 33, no. 2, pp. 316–323, 2007.
[26] Y. Wu, C. Chiang and A. Chen, "Hiding sensitive association rules with limited side effects," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 1, pp. 29–42, 2006.
[27] A. Gkoulalas and V. Verykios, "Exact knowledge hiding through database extension," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 5, pp. 699–713, 2008.
[28] C. Wu and Y. Huang, "A cost-efficient and versatile sanitizing algorithm by using a greedy approach," *Soft Computing*, vol. 15, no. 5, pp. 939–952, 2011.
[29] P. Cheng, I. Lee, C. Lin and J. Roddick, "Hide association rules with fewer side effects," *IEICE Transactions on Information and Systems*, vol. 98, no. 10, pp. 1788–1798, 2015.
[30] T. Hong, C. Lin, K. Yang and S. Wang, "Using TF-IDF to hide sensitive itemsets," *Applied Intelligence*, vol. 38, no. 4, pp. 502–510, 2013.
[31] C. Wei, T. Hong, J. Wong, G. Lan and W. Lin, "A GA-based approach to hide sensitive high utility itemsets," *The Scientific World Journal*, vol. 2014, Article ID 804629, pp. 1–13, 2014.
[32] U. Yun and J. Kim, "A fast perturbation algorithm using tree structure for privacy preserving utility mining," *Expert Systems with Applications*, vol. 42, no. 3, pp. 1149–1165, 2015.
[33] T. Wu, J. Zhan and J. Lin, "Ant colony system sanitization approach to hiding sensitive itemsets," *IEEE Access*, vol. 5, pp. 10024–10039, 2017.

[34] T. Wu, J. Lin, Y. Zhang and C. Chen, "A grid-based swarm intelligence algorithm for privacy-preserving data mining," *Applied Sciences*, vol. 9, no. 4, no. 774, pp. 1–20, 2019.
[35] A. Divanis and V. Verykios, "An overview of privacy preserving data mining," *The ACM Magazine for Students*, vol. 15, no. 4, pp. 23–26, 2009.
[36] L. Lekshmy and A. Rahiman, "A sanitization approach for privacy preserving data mining on social distributed environment," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 7, pp. 2761–2777, 2020.
[37] J. Wu, G. Srivastava, A. Jolfaei, P. Fournier-Viger and J. Lin, "Hiding sensitive information in e-Health datasets," *Future Generation Computer Systems*, vol. 117, pp. 169–180, 2021.
[38] J. Wu, G. Srivastava, U. Yun, S. Tayeb and J. Lin, "An evolutionary computation-based privacy-preserving data mining model under a multithreshold constraint," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 3, pp. 1–19, 2021.
[39] A. Hasan, Q. Jiang, J. Luo, C. Li and L. Chen, "An effective value swapping method for privacy preserving data publishing," *Security and Communication Networks*, vol. 9, no. 16, pp. 3219–3228, 2016.
[40] A. Zigomitros, F. Casino, A. Solanas and C. Patsakis, "A survey on privacy properties for data publishing of relational data," *IEEE Access*, vol. 8, pp. 51071–51099, 2020.
[41] K. Ranjith and A. Mary, "Privacy-preserving data mining in spatiotemporal databases based on mining negative association rules," in *Emerging Research in Data Engineering Systems and Computer Communications*, Singapore: Springer, pp. 329–339, 2020.
[42] N. Nanavati and D. Jinwala, "A novel privacy-preserving scheme for collaborative frequent itemset mining across vertically partitioned data," *Security and Communication Networks*, vol. 18, pp. 4407–4420, 2015.
[43] A. Telikani and A. Shahbahrami, "Data sanitization in association rule mining: An analytical review," *Expert Systems with Applications*, vol. 96, pp. 406–26, 2018.
[44] J. Lin, T. Wu, P. Fournier-Viger, G. Lin, T. Hong *et al.,* "A sanitization approach of privacy preserving utility mining," in *Proc. of the Int. Conf. on Genetic and Evolutionary Computing*, USA, pp. 47–57, 2015.
[45] S. Bagui and P. Dhar, "Positive and negative association rule mining in Hadoop's MapReduce environment," *Journal of Big Data*, vol. 6, no. 1, pp. 1–6, 2019.
[46] A. Kadir, A. Bakar and A. Hamdan, "Frequent absence and presence itemset for negative association rule mining," in *Proc. of the Int. Conf. on Intelligent Systems Design and Applications*, Spain, pp. 965–970, 2011.
[47] C. Cornelis, P. Yan, X. Zhan and G. Chen, "Mining positive and negative association rules from large databases," in *Proc. of the IEEE Conf. on Cybernetics and Intelligent Systems*, China, pp. 1–6, 2011.
[48] N. Rai, S. Jain and A. Jain, "Mining interesting positive and negative association rule based on improved genetic algorithm (MIPNAR_GA)," *International Journal of Advanced Computer Science and Applications*, vol. 5, no. 1, pp. 1–10, 2014.
[49] N. Rai, S. Jain and A. Jain, "Mining positive and negative association rule from frequent and infrequent pattern based on IMLMS_GA," *International Journal of Computer Applications*, vol. 77, no. 14, pp. 48–52, 2013.
[50] S. Narmadha and S. Vijayarani, "Protecting sensitive association rules in privacy preserving data mining using genetic algorithms," *International Journal of Computer Applications*, vol. 33, no. 7, pp. 37–34, 2011.
[51] P. Lakshmi, C. Rao, M. Dabbiru and K. Kumar, "Sensitive itemset hiding in multi-level association rule mining," *International Journal of Computer Science & Information Technology*, vol. 2, no. 5, pp. 2124–2126, 2011.
[52] F. Ullah, I. Ullah, A. Khan, M. Uddin, H. Alyami *et al.,* "Enabling clustering for privacy-aware data dissemination based on medical healthcare-IoTS (MH-IoTS) for wireless body area network," *Journal of Healthcare Engineering*, vol. 2020, Article ID 8824907, pp. 1–10, 2020.
[53] U. Ahmed, G. Srivastava and J. Lin, "A machine learning model for data sanitization," *Computer Networks*, vol. 189, pp. 1–6, 2021.
[54] J. Wu, G. Srivastava, M. Pirouz and J. Lin, "A GA-based data sanitization for hiding sensitive information with multi-thresholds constraint," in *Proc. of the Int. Conf. on Pervasive Artificial Intelligence*, Taiwan, pp. 29–34, 2020.

[55] A. Khedr, W. Osamy, A. Salim and A. Salem, "Privacy preserving data mining approach for IoT based WSN in smart city," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 8, pp. 555–563, 2019.
[56] R. Lu, K. Heung, A. Lashkari and A. Ghorbani, "A lightweight privacy-preserving data aggregation scheme for fog computing-enhanced IoT," *IEEE Access*, vol. 5, pp. 3302–3312, 2017.
[57] H. Li, F. Guo, W. Zhang, J. Wang and J. Xing, "(a, k)-Anonymous scheme for privacy-preserving data collection in IoT-based healthcare services systems," *Journal of Medical Systems*, vol. 42, no. 3, pp. 1–9, 2018.
[58] J. Du, C. Jiang, E. Gelenbe, L. Xu, J. Li *et al.,* "Distributed data privacy preservation in IoT applications," *IEEE Wireless Communications*, vol. 25, no. 6, pp. 68–76, 2018.
[59] Z. Almusaylim, N. Zaman and L. Jung, "Proposing a data privacy aware protocol for roadside accident video reporting service using 5G in vehicular cloud networks environment," in *Proc. of the IEEE Int. Conf. on Computer and Information Sciences*, Malaysia, pp. 1–5, 2018.
[60] H. Sun and R. Grishman, "Lexicalized dependency paths based supervised learning for relation extraction," *Computer Systems Science and Engineering*, vol. 43, no. 3, pp. 861–870, 2022.
[61] H. Sun and R. Grishman, "Employing lexicalized dependency paths for active learning of relation extraction," *Intelligent Automation & Soft Computing*, vol. 34, no. 3, pp. 1415–1423, 2022.
[62] A. Rahman, M. Saeed, M. Mohammed, S. Krishnamoorthy, S. Kadry *et al.,* "An integrated algorithmic MADM approach for heart diseases' diagnosis based on neutrosophic hyper soft set with possibility degree-based setting," *Life*, vol. 12, no. 5, 729, pp. 1–13, 2022.
[63] T. Wah, M. Mohammed, U. Iqbal, S. Kadry, A. Majumdar *et al.,* "Novel DERMA fusion technique for ECG heartbeat classification," *Life*, vol. 12, no. 6, 842, pp. 1–13, 2022.
[64] M. Soni, S. Gomathi, P. Kumar, P. Churi, M. Mohammed *et al.,* "Hybridizing convolutional neural network for classification of lung diseases," *International Journal of Swarm Intelligence Research*, vol. 13, no. 2, pp. 1–5, 2022.
[65] M. Mohammed, I. Ali and O. Obaid, "Diagnosing pilgrimage common diseases by interactive multimedia courseware," *Baghdad Science Journal*, vol. 19, no. 1, pp. 168–178, 2022.
[66] S. Darwish, R. Essa, M. Osman and A. Ismail, "Privacy preserving data mining framework for negative association rules: An application to healthcare informatics," *IEEE Access*, vol. 10, pp. 76268–76280, 2022.