[1]Senthilkumar K,

[2]Dr. B. Selvanandhini

Provisioning A Positive Cluster Associative Schema for Privacy Preservation in the Mining Paradigm

JES Journal of Electrical Systems

Abstract: - The considerable data age has seen an exponential rise in personal data due to the widespread use of mobile devices and the Internet. Since important information is extracted throughout the data mining process, there are significant privacy concerns regarding the network user data. Adding random noise to safeguard sensitive data while preserving certain statistical aspects is called privacy preservation, a novel paradigm that operates independently of the attackers' prior knowledge. Expanding upon the proposed technique for network user data and robust privacy, this research proposes a multiple cores positive cluster associative schema (PCAS-PP) with privacy preservation. During the data mining, you can better optimize data clustering and take advantage of the privacy leakage problem. Researchers do a thorough theoretical study and run simulations to assess our schema. The findings demonstrate that our schema is more accurate, efficient, and protects privacy than earlier schemas. Some other metrics like run time, F-measure and cluster ratio of three datasets are evaluated and compared.

Keywords: privacy preservation, data mining, clustering, associative, sensitivity

1. INTRODUCTION

With ever-more-advanced services, more and more elements of the Internet of Things (IoT) are included in our daily lives. In conjunction with social media, the rapidly growing number of intelligent gadgets exponentially increases personal user data [1]. Data collecting is now done by organizations other than the government and statistics departments because of the growth of IoT and database technologies. Data mining allows additional examination and utilization of user information from social media platforms, online stores, and search engines from different backgrounds, individuals and organizations [2]. Regretfully, raw data, including private or sensitive information, may unavoidably be exposed, leading to privacy leaks. On the other hand, sensitive data may leak in many publishing programs that display data to users directly from databases if data publishers fail to take the necessary precautions for data protection [3]. For instance, if the material is not thoroughly vetted before publication, it may be used by commercial rivals in publications such as the annual financial report of a publicly traded company or product details made available by the business [4]. As a result, it is tough to guarantee privacy in data mining while maintaining accuracy using privacy preservation strategies. Safeguarding network user privacy has recently garnered significant attention from the academic community and the broader public [5].

Data encryption, selective data dissemination, distortion, and other methods are privacy preservation strategies. Data encryption is frequently utilized in remote contexts, and encryption techniques conceal sensitive data during data mining [6]. To publish selected data values, generalize or anonymize the data, etc., limited data publication disseminates information subject to specific conditions. By introducing noise, generating exchange and randomization, blocking, and other techniques, sensitive data can be distorted while other data or attributes are preserved thanks to data distortion technology [7]. The treated data likely retains specific statistical features for different procedures, such as data mining.

Privacy is a revolutionary paradigm for statistical databases independent of adversaries' computing capability or prior information. It establishes a strict attack model that lowers the risk of privacy disclosure while successfully guaranteeing data availability [8]. Additionally, it's a form of data distortion method. Other algorithms that employ data mining and privacy techniques comprise three different clustering methods: the Privacy Preservation clustering method, the Privacy Preservation K-means clustering method, and the Improved Privacy Preservation K-means clustering method. With the addition of privacy-compliant noise, these algorithms can cluster data

[1] Research Scholar, Department of Computer Science, Pollachi College of Arts and Science, (Affiliated to Bharathiar University, Coimbatore), Pollachi, senthilssenthil1987@gmail.com
[2]Assistant Professor, Department of Computer Science, Pollachi College of Arts and Science, (Affiliated to Bharathiar University, Coimbatore) Pollachi, selvanandhini.n@gmail.com

efficiently. Still, the clustering effect diminishes when IDP-K-means is applied to a dataset with an uneven density distribution and an undetermined number of clusters [9] – [11]. More significant dataset volumes and less privacy budget constraints require more time and less clustering when using the clustering model. To address these issues, a fresh data mining technique must be developed to protect privacy [12]. The researchers' focus in this study is on protecting privacy when clustering analysis is performed on network user data. It suggests the preservation of privacy [13] – [15]. The proposed technique creates a clustering schema for multiple cores PCAS-PP, effectively addressing privacy leakage in data mining for network user data. Differenced privacy is advantageous since it ensures data privacy regardless of previous information, which makes this possible. For more exacting purposes, select the optimal initial core point selection to produce a collection of initial vital points, and then choose from the output set the best core points for clustering. Here is an overview of the main contributions made by this work.

1) The researchers suggest a model to enhance data security and accuracy through network user data clustering analysis. According to a privacy study, our clustering structure can shield publishers' data from attacks while satisfying their query needs.

2) Using privacy, researchers suggest a proposed clustering approach with multiple cores. By optimizing the first core point selection, the proposed, as opposed to existing methods, effectively overcomes the blindness and unpredictability of PCAS-PP. Additionally, the suggested technique demonstrates clear benefits when applied to datasets with bigger scales, considerable density dispersion, and more flexible budgetary constraints for privacy.

3) It validates our approach through extensive experiments and demonstrates the accuracy of our schema. The findings show that our algorithm outperforms the others regarding accuracy, efficiency, and privacy protection.

The work is drafted as follows: section 2 details privacy preservation in mining; the proposed PCAS-PP is drafted in Section 3. The numerical discussion is shown in section 4, and the outcomes are explained in section 5.

2. RELATED WORK

Privacy protection measures must be integrated into process mining analyses, as demonstrated by the expanding awareness of global data privacy, the application of privacy legislation and the expanding acceptance of the Fairness, Accuracy, Confidentiality, and Transparency (FACT) concept [16]. It will summarise research in this field in this section. Privacy concerns have gained more attention in recent years, especially in light of the need to avoid incidents involving the improper use of personal data. Legislative privacy restrictions, like the GDPR in [17], have been enacted due to this. They forbid the sharing of potentially dangerous information between institutions and encourage employing design strategies such as data minimization, encryption, and pseudonymization to protect privacy [18].

The possibility of privacy breaches has also prompted the creation of novel approaches to data storage and retrieval, blockchain technology, and the exchange and exploration of encrypted data from cloud infrastructure for various commercial purposes, for example. As a result, privacy protection has gradually gained more attention from scientists across multiple disciplines. Many concepts and approaches have been presented in the literature to address privacy problems in process mining [19]. Anonymization, data disruption, and encryption are the three main categories into which these techniques fall. Process mining has comparable privacy protection problems to pattern mining, emphasizing retaining anonymized trace characteristics and order preservation. As part of the Responsible Data Science (RDS) framework [20], the idea of Responsible Process Mining (RPM) was presented to address problems with data abuse analysis and lessen adverse effects. Process mining privacy issues are similar to those involving anonymous trace characteristics and ordered pattern mining. RPM (Responsible Process Mining) was initially presented in the literature as a fresh RDS (Responsible Data Science) challenge to combat data analysis misuse and stop unintended consequences.

The two primary study areas that are now the focus of privacy preservation in process mining are intra-organisational and between organizations. Studies on protecting intra-organisational privacy focus mainly on restricting access to sensitive data to preserve data privacy. On the other hand, research on inter-organizational privacy preservation demands more intricate security protocols to deal with the problem of confidentiality leaking while exchanging data between various companies. Three categories of privacy protection strategies are the

primary focus of intra-organizational research: data perturbation, encryption, and anonymization, together with the algorithm models that result from them—in addition to assessing the risks and proposing a confidentiality framework [21] talked about the outstanding challenges associated with event log encryption, such as the inadequacy of encrypting event data and the drawbacks of relying on one solution. As mentioned by Burattin et al. in their discussion of the outsourcing of PM analytics, sensitive data must be hidden using symmetric or homomorphic encryption to guarantee the privacy of event logs and the procedures subsequently. A trustworthy third party for business process models in the public and commercial sectors was provided by [22]. The author presented an Alpha technique for process discovery using encryption protocols that may safeguard people's and software's privacy. According to the literature, a PM privacy-preserving system architecture based on the authorization model of attribute-based access control (ABAC) should be implemented to impose privacy laws on event logs. Nonetheless, these previously indicated methods still carry some risk of information leakage, assuming the attacker has little prior knowledge.

On the other hand, the author focused on how privacy approaches cause utility loss and provided an ideal method to set ε through utility-based estimation. In response, the notion of localized privacy was made possible by Fahrenkprog-Petersen et al.'s privacy-protected event log release Framework (PRIPEL); it made it possible to maintain privacy at the case level instead of the log level. Privacy based on the $(\varepsilon, '0)$ engine was introduced by [23] for privacy protection to guarantee that personal information cannot be detected even if an attacker knows it beforehand. To safeguard event records, anonymization techniques are used in various investigations. The author suggested that the event log cleaning algorithm utilizes T-closure and K-anonymity as requirements for privacy protection. Comparable traces are found and merged to satisfy privacy protection regulations. Expanding upon the anonymization characterizations of privacy metadata and healthcare data, the author examined the need for data privacy in healthcare process models. To help with PM analysis for healthcare procedures, they provided a theoretical framework for PM.

Additionally, approaches for evaluating how well privacy-preserving techniques work were presented and broad privacy quantification frameworks were proposed. The current study uses these techniques with a comprehensive wen-based PM system to create an open-source, web-based PM application [24]. Group-based anonymization formalizes a TLKC privacy model that guards against attribute-linking attacks in performance analysis and process discovery. In research protecting privacy, performance indicators are the measurable standards that enable the assessment of privacy protection measures in processes. Performances come in a variety of forms. Indications for common PPPM approaches about different research areas. However, because of the uniqueness of the analysis goal and the peculiarities of the log data structure, these current performance indicators are not appropriate for use in general analysis. Table 3 shows that when it comes to utility loss, standard performance measures include timeliness loss, accuracy loss, completeness loss, and usability loss. On the other hand, performance measurements of information loss, degree of trust, degree of assistance, and budget for privacy are often used when it comes to privacy gain in PM [25].

Additionally, it emphasizes how process mining's privacy-preserving study goal differs from other study viewpoints in this area. Another crucial aspect of PPPM is the identification of deviations, the ones that are privacy-secured and the initial event log-based process models [25]. The graph theory-based structural similarity-based method, the group-based strategy with several points of view, and other approaches are examples of the scant research that has been done on this topic. The behaviour viewpoint is crucial regarding privacy-preserving techniques, in addition to the features of the event log related to control, resources, time, and other perspectives. To the best of our knowledge, this problem has yet to be the subject of any prior research; therefore, this paper represents the first to take a behavioural viewpoint on privacy preservation.

3. METHODOLOGY

3.1. Problem description

For each unit of information, let us denote a finite collection of unique objects by $I = \{i_1, i_2, \ldots, i_m\}$. The terms indicate sensitive or non-sensitive items I_S and I_N for each pair of elements, i.e. $I = I_S \cup I_N$ and $I_S \cap I_N = \emptyset$. A transaction is denoted by tid, a distinct subset of I. Here, T_i represents the i^{th} arrived transaction in the infinite series of transactions that make up a transactional data stream and is represented by $TDS = \{T_1, T_2, \ldots, T_n, ..\}$.

Over a data stream TDS, the most recent TDS transactions are stored in a transactional sliding window $TSWd_{\left[\frac{n-w+1}{p}\right]} = [T_{n-w+1}, T_{n-w+2}, \ldots, T_n]$. The most recent p transactions are added to TSW, and the oldest p transactions are eliminated from the window when p new transactions are received from TDS: the window identification and size are represented by $\left[\frac{n-w+1}{p}\right]$ and w, respectively. TSW moves forward at a pace of p steps. Fig 1 displays a sliding window-based transactional data stream with window and step sizes of 6 and 2, respectively. Slightly sensitive things are $\alpha\ and\ \gamma$, whereas sensitive items are $a_1, a_2, b_1, and\ b_2. t_{i+1}, \ldots, t_{i+6}$ are contained in the transactional sliding window $TSW_{\left[\frac{i+1}{2}\right]}$. In the transactional sliding window (TSW), the two oldest transactions ($t_{i+1}\ and\ t_{i+2}$) are removed when fresh ones ($t_{i+7}\ and\ t_{i+8}$) arrive. Assume, as illustrated in Fig 1, that Mike is aware that on a specific day, Lily bought a2 and b1 from a store and that the store allows mining of its purchasing data stream. Assume that Mike knows about Lily's transaction in $TSW_{\left[\frac{i+3}{2}\right]}$ and that a sliding window will enable him to view the data stream. Then, Lily's privacy is in jeopardy because Mike can assume that Lily also purchased γ.

3.2. Privacy model

Researchers expand the privacy paradigm of ρ-uncertainty, created to broadcast transactional static data in data streams. Assume that the attacker can view a data stream using a sliding window paradigm and is aware of specific elements φ in the victim's transaction inside a sliding window of transactions. A violation of privacy occurs if an attacker might reasonably be anticipated to deduce from the transactional sliding window that t includes a sensitive item (α) in addition to χ, where $\alpha \in$ ώ. Taking χ as the antecedent and α as the consequent, by employing the transactional sliding window as a mine, one can obtain the association rule $\chi \rightarrow \alpha$. Sensitive association rules (SARs) are association rules that fall under this category. In a transactional sliding window, TSW_i, The formula to calculate the confidence of a $SAR\ \chi - \alpha\ is\ sup(\chi \cup \alpha)/sup(\chi)$, where $sup(\chi)$ is the total number of transactions in TSW_i containing χ. Every transactional sliding window must be anonymized, and the likelihood that each SAR appears will be smaller than a specific threshold value represented by the symbol ρ in a transactional sliding window.

Lemma 1: The transactional sliding window TSW_i over a transactional data stream TDS is said to meet ρ −uncertainty if and only if, for any $for\ all\ t \in TSW_i, \forall \chi \subset t, and\ \forall \alpha \notin \chi, \alpha \in I_S$, the confidence of the $SAR\ \chi \rightarrow \alpha$ is smaller than a threshold value ρ, where $\rho \in (0, 1)$. TDS fulfils ρ −uncertainty if any transactional sliding window does as well.

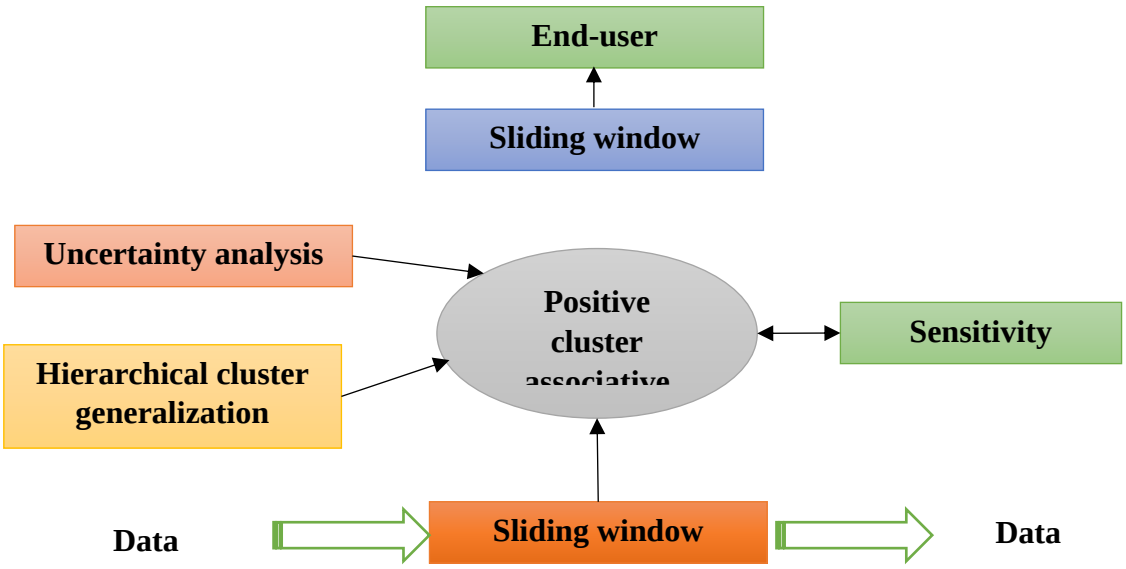

Fig 1 Block diagram of proposed PCAS-PP

3.2. Sliding window model for privacy preservation

Suppose a sliding window meets the requirements for ρ −uncertainty; it advances when p new transactions arrive. In that case, the most recent p transactions are added to the window, and the oldest p transactions are removed. Due to the addition and deletion of transactions, the present sliding window might violate ρ-uncertainty. When a transaction fulfils ρ −uncertainty, researchers examine how changing its addition or deletion affects a sliding window's privacy.

Theorem 1: Assume that a sliding window meets the ρ −uncertainty requirements. It encompasses the gathering of objects. For a transaction called t. The following scenarios could occur in a sliding window:

1) The sliding window does not alter ρ-uncertainty for any $SAR\ \chi \rightarrow \alpha$;

2) Any $SAR\ \chi \rightarrow \alpha$ in the sliding window when $\chi \subseteq I_t\ and\ \alpha \in I_t$;

3) Any $SAR\ \chi \rightarrow \alpha$ in the sliding window.

Proof: As long as the association rule meets ρ-uncertainty, The formulas for calculating the confidence of each $SAR\ \chi \rightarrow \alpha\ are\ sup(\chi \cup \alpha)/sup(\chi)\ and\ sup(\chi \cup \alpha)/sup(\chi) < \rho$. It examines the following scenarios when removed from the sliding window.

1) $Sup(\chi)$ decreases and remains constant for each $SAR\ \chi \rightarrow \alpha\ when\ \chi \subseteq I_t\ and\ \alpha /\in I_t.sup(\chi \cup \alpha)/sup(\chi)$ rises as a consequence. Supp $(\chi \cup \alpha)/sup(\chi) >= \rho$ is an example of this type of result. Thus, $\chi \rightarrow \alpha$ can be thought of as the antithesis of ρ −uncertainty.

2) It is evident that for any $SAR\ \chi \rightarrow \alpha$, the elimination of t does not affect the association rule's degree of confidence, where $\alpha \in I_t\ and\ \chi * I_t$.

3) Put differently, if $\chi \fallingdotseq \alpha$ for every $SAR\ \chi \rightarrow \alpha, then\ \chi \wedge \alpha \subseteq I_t$. It has ownership of it.

$$\frac{\sup(\chi \cup \alpha)}{\sup(\chi)} = \frac{\sup(\chi \cup \alpha)}{\sup(\chi \cup \alpha) + \sup(\chi \cup \neg\alpha)} \tag{1}$$

$$= \frac{\sup(\chi \cup \alpha) + \sup(\chi \cup \neg\alpha)}{\sup(\chi \cup \alpha) + \sup(\chi \cup \neg\alpha)} - \frac{\sup(\chi \cup \neg\alpha)}{\sup(\chi \cup \alpha) + \sup(\chi \cup \neg\alpha)} \tag{2}$$

$$= 1 - \frac{\sup(\chi \cup \neg\alpha)}{\sup(\chi \cup \alpha) + \sup(\chi \cup \neg\alpha)} \tag{3}$$

Where the number of transactions with χ but no α is represented by $\sup(\chi \cup \neg\alpha)$. The deletion of t results in a decrease in $sup(\chi \cup \alpha)$, which consequently results in a drop in $sup(\chi \nVdash \alpha)\ sup(\chi)$. The sensitive association rule does not violate ρ-uncertainty after t is removed.

Theorem 2: Let us assume that ρ −uncertainty is satisfied by a sliding window. It consists of a group of things for a transaction called t. We have the following after adding t to the sliding window. There is no ρ-uncertainty if $\varphi \leq \alpha$ in the sliding window and $\chi * I_t\ and\ \alpha \in I_t$.

1) Any $SAR\ \chi \rightarrow \alpha$ in the sliding window and $\chi *$ It does not violate ρ −uncertainty in this situation; and

2) Any $SAR\ \chi \rightarrow \alpha$ in the sliding window and $\chi \subseteq I_t\ and\ \alpha \in I_t$ might be. Supporting data is provided in the form of similar proof. The sliding window could break ρ −uncertainty. If modifications are made, just these $SARs$, which might contradict ρ −uncertainty, must be considered. Utilizing the static anonymization technique suggested, an analysis is conducted on each SAR inside the current sliding window to determine whether or not ρ −uncertainty is violated.

3.3. Information metrics

Should there be a transactional sliding window TSW_i that does not satisfy ρ −uncertainty within a transactional data stream TDS, then we can use generalization and suppression to change it so that it does. A generalization hierarchy tree is built for the items in I_N, having three nodes: a leaf indicating an item that is not sensitive, an internal node that stands for both the generalized value for every item represented at the root and a generic value for specific objects. This procedure is the same as that described. Items $a_1, a_2, b_1, and\ b_2$ are not sensitive. One can generalize $a_1\ and\ a_2\ to\ A\ and\ b_1\ and\ b_2\ to\ B$. Moreover, $A\ and\ B$ are generalizable to all presume that in the non-sensitive object hierarchy tree H, k is a node. $IL_k = |leaves(k)|/|IN$ is the formula for calculating the information loss of k, where k represents the root node in H and $leaves(k)$ represents the collection of leaves in the sub-tree. In the case of a leaf k, $IL_k = 0$. For instance, node A has an information loss of $\frac{24=1}{2}$. $InfoLoss(a) = IL_k$ defines information loss for a suppressed item; if an item is an item and is generalized to node $k \in H, InfoLoss(a) = 1$. The number of TSW_i transactions that contain an is denoted by $sup(a)$.

$$InfoLoss\ (TSW_i) = \frac{\sum_{a\in I} \sup(a) * InfoLoss\ (a)}{\sum_{a\in I} \sup(a)} \tag{4}$$

This is the information loss incurred when anonymizing two types of transactional sliding windows: $TSW_i\ and\ TDS$ (transactional data stream).

The following formula can be used to determine the average information loss of a sliding window for a transactional data stream (TDS):

$$AveInfoLoss\ (TSW) = \frac{\sum_{i=1}^{\omega} Infoloss\ (TSW_i)}{\omega} \tag{5}$$

The sliding window count is represented by ω.

3.4. Data transactions

To find the confidence of $SAR\ \chi \rightarrow \alpha$, one must ascertain the supports of itemsets $\chi\ and\ \chi\alpha$. It is possible to transform a TSW_i transactional sliding window into a soft set. Assume the following: U is the initial universe set, $P(U)$ is $U's$ power set, $\eta \subseteq E$ is an array of parameters is E. To illustrate a soft set over U, consider the pair (F, η). A mapping F with the notation $F: \eta \rightarrow P(U)$ is represented. Regarding TSW_i, consider U the transactions that occur in TSW_i; I represents the collection of items in E, and η for the subset of objects in TSW_i. Thus, item an in $TSW_i, a \in \eta$, is contained in the collection of transactions denoted by $F(a)$. $|F(a)|$ is a support for a.

$$F(a_1) = \{t_{i+2}, t_{i+5}\} \tag{6}$$
$$F(a_2) = \{t_{i+1}, t_{i+2,}\ t_{i+3}, t_{i+4}\} \tag{7}$$
$$F(b_1) = \{t_{i+1}, t_{i+2,}\ t_{i+3}, t_{i+4}\ , t_{i+5}, t_{i+6}\} \tag{8}$$
$$F(b_2) = \{t_{i+1}, t_{i+2,}\ t_{i+3}, t_{i+4}\ , t_{i+5}, t_{i+6}\} \tag{9}$$
$$F(\propto) = \{t_{i+1}, t_{i+5}\} \tag{10}$$
$$F(\gamma) = \{t_{i+3}, t_{i+4}, t_{i+6}\} \tag{11}$$

Where $a\ and\ b$ are in η, ab's support is $|F(a) \cap F(b)|$. Regarding $TSW_{\left[\frac{i+1}{2}\right]}$ two e, the suitable soft set $(F, \acute{\varepsilon})$ is Supports for a_1 are $|F(a_1)| = |\{t_{i+2}, t_{i+5}\}| = 2$, and support for a_1a_2 are $|F(a_1) \cap F(a_2)| = |\{t_{i+2}, t_{i+5}\} \cap \{t_{i+1}, t_{i+2}, t_{i+3}, t_{i+4}\}| = |\{t_{i+2}\}| = 1$. It aims for the sliding window to anonymize it to meet ρ −uncertainty dynamically. Every transactional sliding window can be anonymized using the static technique. Nevertheless, an effective algorithm is required due to the sliding window's quick and frequent updates. Due to transaction additions and deletions, the adjustment to the sliding window does not eliminate ρ-uncertainty. All it has to do is look for and investigate the impacted SARs that could potentially breach ρ-uncertainty. Before introducing the suppression-added generalization technique, we first give the suppression algorithm.

Algorithm 1:

Input: sliding window, data item set, step size, transactional data stream, window size, soft set and privacy factors;

Output: data streams from the window

1. Slide over TDS to acquire TSW'_{i+1};
2. **for** $a \in I^i_{supp}$ **do**
3. **If** i_{supp} is not visible in the provided raw transactions regarding $TSW'_i \backslash W_{del}$, **then**
4. Remove i_{supp} from I^i_{supp};
5. **else**
6. Reduce i_{supp} from W_{add};
7. Predict I^{i+1}_{supp};
8. **return**

3.5. Generalization

Researchers employ global generalization to avoid spurious association rules from forming, much like suppression. Global generalization assigns a level in a generalization hierarchy tree H to every instance of an item and every item in the identical hierarchy H sub-tree. The control algorithm uses selective global suppression of particular items and worldwide generalization of the non-sensitive items in I_N across the hierarchical tree H to anonymize static transactional data and meet ρ −uncertainty. In this paper, the first sliding window is anonymized using $TDControl$. Changes to the sliding window that remove or add transactions could cause it to no longer meet ρ −uncertainty. To satisfy ρ −uncertainty and control it continuously. The root of H is contained in the particularization tree T, which is a subset of H. The set of its leaves indicates the current window's generalization case. The first TSW'_{i+1} for a transactional data (TDS) stream is $TSW'_{i+1} = TSW'_{i+1}$ which is derived by sliding TSW'_i ahead in step size p. If the items in leaves (i_{supp}) ($i_{supp} \in I^i_{supp}$) have already been eliminated from the raw transactions in previous anonymized processes, then delete them from there regarding $TSW'_i \backslash W_{del}$ to make sure the anonymous sliding window does not contain any erroneous association rules. The generalized values in W_{add} are then saved to Gene once the W_{add} items have been generalized per the particularization tree T currently in use. Since doing so would expose sensitive information, we avoid generalizing about sensitive objects. The BuildASRT and supp function is called to keep some critical information hidden and ensure the SARs don't violate ρ −uncertainty because they only contain sensitive things. Furthermore, we adjust $(F, \eta), TSW'_{i+1}\ W_{del}\ and\ W_{add}$ by SuppASRT. To identify the SARs that do not comply with ρ −uncertainty. Suppression and generalization are then used to address the SARs. At last, we have the sliding window TSW'_{i+1}, which is anonymous.

It consists of the impacted rules that are sensitive and have an antecedent length of 1. To address the SARs that defy ρ −uncertainty, researchers employ generalization and suppression. As stated, identify the SARs broken with prior length lev and deal with them through suppression and generalization. Next, add nodes for each following level should the collection of level nodes not be ∅. TSW'_{i+1}, the anonymous sliding window, is the final result. Is it better to repress it or generalize it? which results in the least amount of information loss. Certain SARs may have an antecedent length smaller than lev due to suppression or generalization. Since these SARs don't violate ρ −uncertainty, they are taken from ASR_{lev}. Based on $SuppASRT$ and T, it updates Gene, $ASRT, TSW'_{i+1}, and\ (F, \eta)$. After that, $SuppASRT$ is returned. The info gain function in the algorithm allows users to choose the Gene item split that yields the highest information gain. If the information benefit from the split is not zero, we specialize in the divided; if not, there is nothing to specialize in. In other words, take a split out of Gene and use its progeny to update $(F, \eta)\ and\ TSW'_{i+1}$.

Algorithm 2:

1. **for** the SAR rule, **do**
2. **if** rule violation leads to uncertainty, **then**
3. Include ASR_{lev};
4. **while** $ASR_{lev} \neq \emptyset$ **do**
5. choose an item based on the maximal payoff ratio;

6. **If** loss ≤ gene, **then**
7. Reduce i_x in ASR_{lev};
8. Set i_x as suppression;
9. **else**
10. generalize i_x;
11. Remove SAR as the length is lesser;
12. Revise ASRT and TSW'_{i+1};
13. **return**

4. NUMERICAL RESULTS AND DISCUSSION

Researchers put the proposal into practice within this segment and conducted several tests to assess its effectiveness.

4.1. Experimental setup

Researchers looked at four datasets from UCI, Wine, Haberman, Waveform Database, and MAGIC, each with varying database features and scale, to assess the effectiveness of our approach. An extensive list of the dataset aliases, data types, attributes, and records is provided in Table []. First, pre-processing is done on the datasets to normalize them and control the values of all the characteristics in the same proximity. During the pre-processing phase, a 0.1 gradient is used to systematically modify the values of Eps and MinPts. To reduce the influence of the Eps and MinPts parameters, this change is only made to 1/25 of the dataset scale. Seeing the clustering effect helps identify each dataset's ideal Eps and MinPts values. As the level of privacy depends on the given Eps and MinPts values, the clustering validity of the method is assessed across a range of privacy constraints. The average results of our studies are presented after 100 separate runs of each test dataset with Windows 10 X64 Ultimate software installed; the Intel (R) Core (TM) i7-4700MQ CPU runs at 3.4GHz and has 8GB of RAM.

4.2. Evaluation metrics

F-Measurement Index: F-measure is a frequently used evaluation index for clustering findings that can be used to assess the accessibility of the grouping outcomes. The F value is proportionate to the degree of similarity between the outcomes when two clustering algorithms' clustering results are computed using the F-measure. The assessment index for the F-measure can be calculated using the following formula:

$$Precision\ (P) = \frac{n_{ij}}{|D_j|} \tag{12}$$

$$Recall\ (C_i, D_j) = \frac{n_{ij}}{|C_i|} \tag{13}$$

$$F_i = \frac{2 * P * R}{P + R} \tag{14}$$

Here, C_i and D_j represent the outcomes of two clustering techniques. Here, $P\ and\ R$ are for precision and recall rate, respectively. The number of components at the intersection of D_j cluster C_i is represented by the symbol N_{ij}. An algorithm is more likely to cluster if its F-measure value is higher. The efficiency of the proposed algorithms for clustering data is examined in this research, and the results are evaluated using the F-measure value. Scientists modify the privacy budget value ε for the two clustering methods. The final results' availability is assessed by comparing the clustering findings with those from the original dataset.

Calinski-Harabasz Index: Another evaluation index for assessing the validity of clustering is Calinski-Harabasz, commonly known as the CH index for short. The deviation matrix inside a class shows how near a class is to itself; the deviation matrix between groups shows how far apart the classes are. This ratio is known as the CH index. The CH index has the following definition:

$$CH\ (k) = \frac{tr\ B(k)/\ (k-1)}{trW(k)/(n-k)}$$

Where, k is the number of classes in use at any one time, n is the number of clusterings, the deviation matrix trace across classes is denoted by $trB(k)$, and the deviation matrix trace inside a class is denoted by $trW(k)$. The dataset's degree of separation is computed as $trB(k)$; this is the sum square of the lengths that separate each cluster's centre from the dataset's centre. "$trW(k)$" is the symbol for the product of squares spanning the class points and the cluster centre, which is used to define how close a cluster is to its core. CH stands for those ratios. The stronger the clustering validity, the higher the CH values correspond to a more distributed cluster within the class.

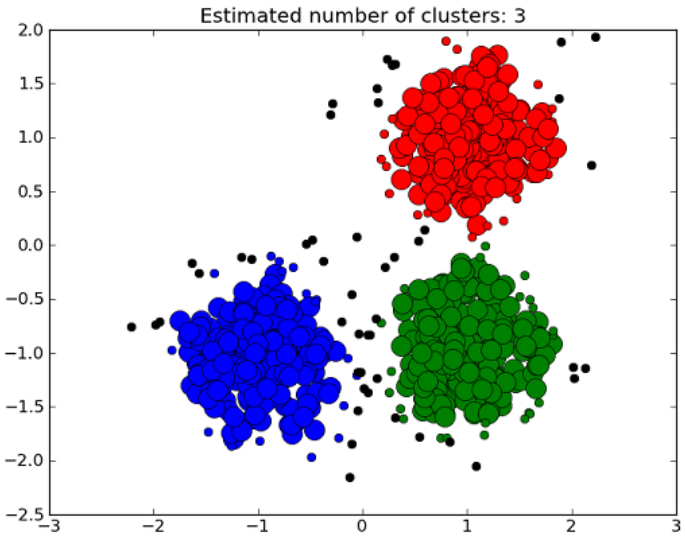

Fig 2 Positive cluster association

4.3. Experimental analysis

Applying the two techniques to dataset D2, it execute privacy preservation clustering; Fig 2 displays the clustering findings, and Fig 3 shows the proposed clustering results. The figures show that both approaches can accurately distinguish each cluster's core, border, and noise points. Nevertheless, because the two approaches' initial core point selection procedures differ, the proposed model produces a different cluster classification in contrast to the existing ones. Because it was initially clustered using multi-core clustering, the proposed yielded more clusters and a more thorough categorization than existing. Despite non-uniform density datasets, the proposed can still precisely locate the noise locations. Because of this, there are times when the distribution seems identical, yet separate clusters are clustering, and it is hard to ignore them. Using the dataset D4 subset with different amounts of data, Fig 3 compares the run efficiency of two distinct algorithms. In D4, the two algorithms are run many times, respectively. When comparing the two methods' time efficiency, it is clear that the proposed model requires a little more time than the smaller data set of the original model. The primary cause is that the first time needed to identify several vital spots is significantly longer than the total time. Nevertheless, the proposed model run time advantage occurs when the dataset hits a specific scale; in this case, $1.67 * 104$ data points. This is because once the initial core points are discovered, the remaining points in each cluster only need to be determined. In contrast, a lot of time is spent using several iterative computing techniques to locate new core points because of the increased data volume and cluster size. From Fig 3, we can infer that our method performs better when dealing with datasets that have more clusters or a larger scale.

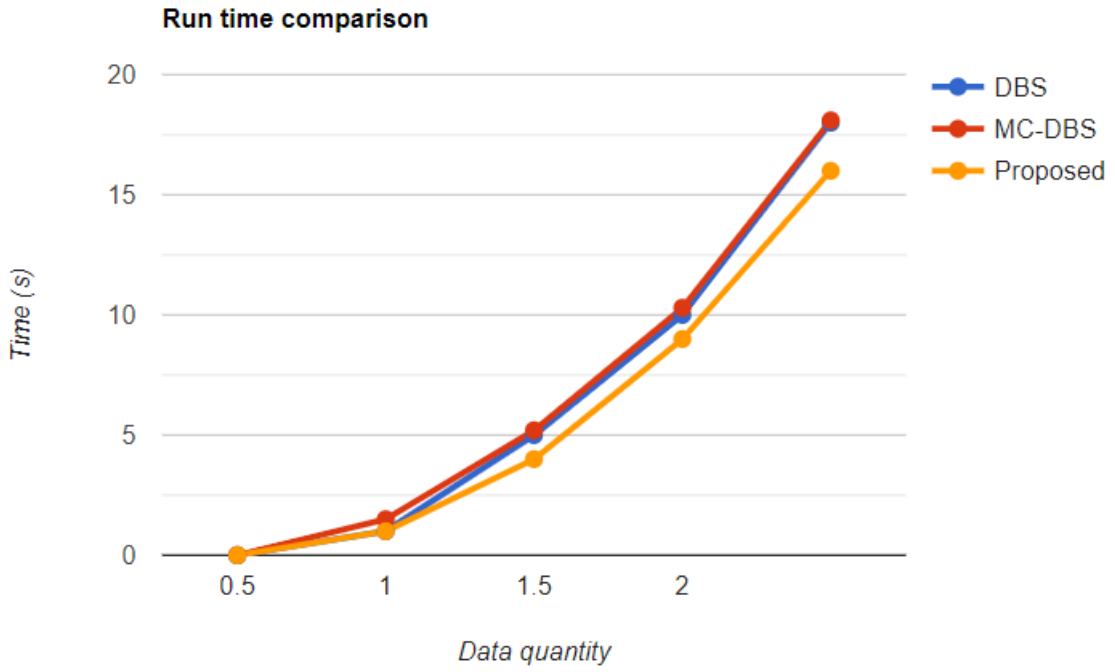

Fig 3 Run time comparison

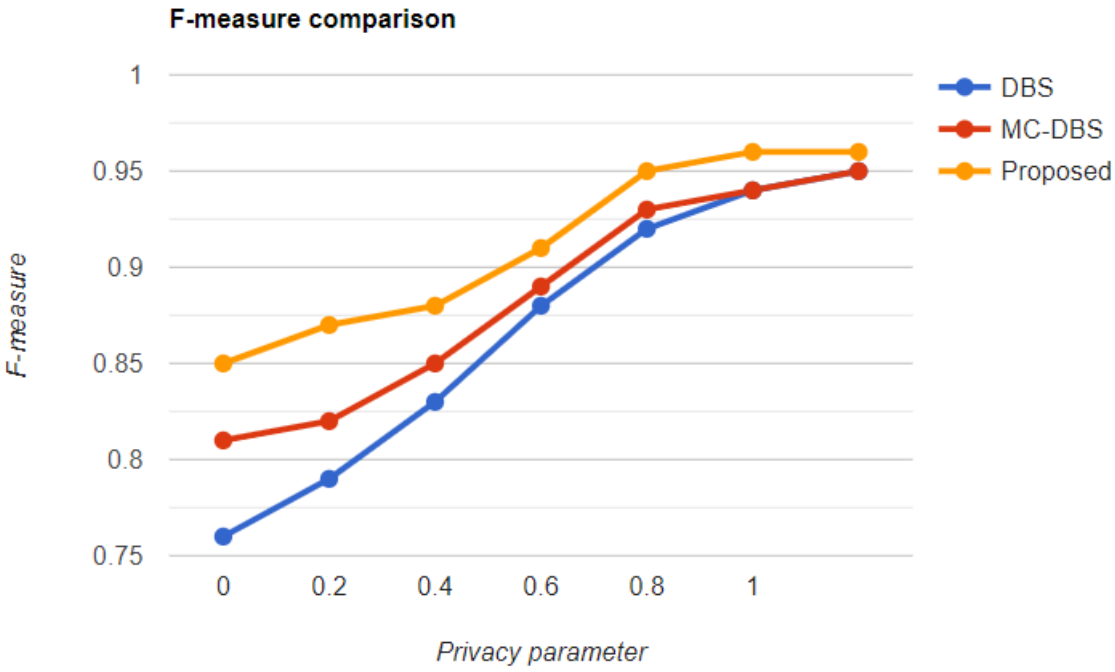

Fig 4 F-measure comparison

Next, the effect of the F-measure on the privacy budget parameter ε, which spans from strict to permissive privacy requirements, is investigated and goes through the following steps: $0.1, 0.2, \ldots, 0.5, \ldots, 1\}$. Fig 4 compares the F-measure values for the two algorithms' clustering outcomes on dataset D3 for a range of ε values. As can be seen, there is little accuracy between the proposed and existing proposals. Both techniques cluster accurately and efficiently when ε is large enough. When ε is reduced, the proposed model is more adept at handling the dataset. However, privacy is beneficial because the extra noise remains constant regardless of dataset size. Consequently, the proposed excels more when it comes to managing large-scale datasets. The proposed models' performance will improve, and their noise immunity will be stronger with more datasets.

Fig 5 Cluster ratio of dataset_1

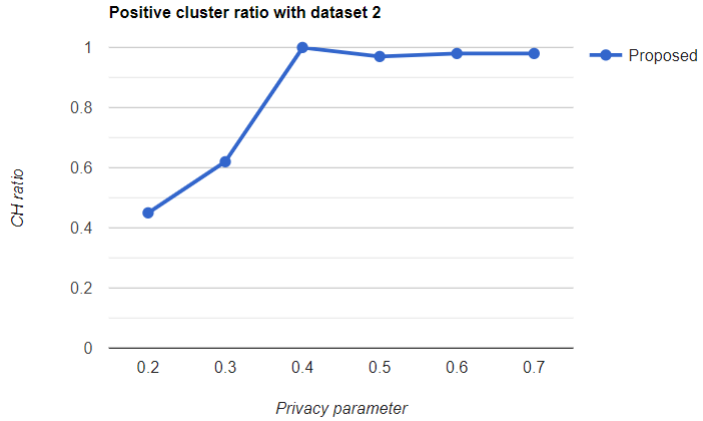

Fig 6 Cluster ratio of dataset_2

Fig 7 Cluster ratio of dataset_3

It used D1, D2, and D3 as our test datasets in the previous experiment. To further assess the clustering quality, compare our method of the Calinski-Harabasz (CH) index with the proposed one. It uses several executions of the two techniques to achieve privacy-preserving clustering and evaluate D1, D2, and D3. We calculate the average CH value to depict the two algorithms' CH ratio curves. The clustering validity of the two techniques is closer the closer the CH ratio is to 1. Fig 5 to Fig 7 display the experiment's results. The statistics show that the proposed may effectively preserve privacy with a modest amount of noise added. According to experts, it ensures that the validity of the clustering produced by the proposed and the traditional clustering approach is comparable. It

suggests that the value of ε affects how much privacy is preserved. The value of ε can be used to control the level of privacy preservation. More excellent noise addition and more robust privacy preservation are associated with smaller ε values. By comparing the three figures' results, upon computing the equivalent privacy preservation level (ε), it becomes feasible to ascertain that the proposed model possesses the subsequent characteristics: greater grouping validity when comparing smaller datasets to larger datasets and higher clustering validity for lower dimensional datasets in comparison to high dimensional datasets.

5. CONCLUSION

One of the significant and most challenging issues in data mining is maintaining the confidentiality and integrity of network user data. Researchers mainly concentrate on privacy preservation in network user data clustering analysis. Unlike previous studies, the researchers effectively tackled the unpredictability and lack of visibility of the proposed PCAS-PP model by selecting many cores at the most significant distance in the clustering outcome. The simulation results show that our method enhances the temporal efficiency of the outcomes and reduces the clustering effect when the added noise level becomes excessive. When the same privacy budget is applied to a larger dataset, the clustering impact is more accurate due to the privacy property, which states that the amount of noise injected is independent of the dataset size. The main goals of this research are to minimize the effect of input parameters on the clustering outcomes and strike a balance between process noise and clustering accuracy.

REFERENCES

[1] Park J, Dong Seong Kim and Hyuk Lim, “Privacy-Preserving Reinforcement Learning Using Homomorphic Encryption in Cloud Computing Infrastructures,.” IEEE Access 8:203564–203579, 2020

[2] Pika, T. Wynn, and S. Budiono, ‘‘Privacy-preserving process mining in healthcare,’’ Int. J. Environ. Res. Public Health, vol. 17, no. 5, pp. 1–28, 2020.

[3] Oishi, Y. Sei, Y. Tahara, and A. Ohsuga, "Semantic diversity: Privacy considering the distance between values of a sensitive attribute," Comput. Secur., vol. 94, Jul. 2020, Art. no. 101823.

[4] Zhang M, Chen Y, Huang J, SE-PPFM: A Searchable Encryption Scheme Supporting Privacy-Preserving Fuzzy Multikeyword in Cloud Systems. IEEE Systems Journal 15(2):2980–2988, 2021

[5] Liu, S. Wen, and W. Zuo, ‘‘Effective sanitization approaches to protect sensitive knowledge in high-utility itemset mining,’’ Int. J. Speech Technol., vol. 50, no. 1, pp. 169–191, Jan. 2020.

[6] Misbha DS , Lightweight key distribution for secured and energy efficient communication in wireless sensor network: An optimization assisted model. High-Confidence Computing 3(2):100126, 2023

[7] Lekshmy and M. A. Rahman, "A sanitization approach for privacy-preserving data mining on a socially distributed environment," J. Ambient Intell. Humanized Computing., vol. 11, no. 7, pp. 2761–2777, Jul. 2020.

[8] Wu, G. Srivastava, A. Jolfaei, P. Fournier-Viger, and J. C.-W. Lin, ‘‘Hiding sensitive information in eHealth datasets,’’ Future Gener. Computing. Syst., vol. 117, pp. 169–180, Apr. 2021.

[9] Wu, G. Srivastava, U. Yun, S. Tayeb, and J. C. Lin, "An evolutionary computation-based privacy-preserving data mining model under a multi-threshold constraint," Trans. Emerg. Telecommunication. Technol., vol. 32, no. 3, Mar. 2021, Art. no. e4209.

[10] 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.

[11] VR Falmari, M. Brindha, Privacy-preserving cloud-based secure digital locker using Paillier based difference function and chaos-based cryptosystem. Journal Inf Security Appl 53. https://doi.org/10.1016/j.jisa.2020.102513, 2020

[12] Bagui and P. C. Dhar, ‘‘Positive and negative association rule mining in Hadoop’s MapReduce environment,’’ J. Big Data, vol. 6, no. 1, pp. 1–6, Dec. 2019.

[13] Ullah, I. Ullah, A. Khan, M. I. Uddin, H. Alyami, and W. Alosaimi, ‘‘Enabling clustering for privacy-aware data dissemination based on medical healthcare-IoTs (MH-IoTs) for wireless body area network,’’ J. Healthcare Eng., vol. 2020, pp. 1–10, Nov. 2020

[14] Madbouly, S. M. Darwish, N. A. Bagi, and M. A. Osman, ‘‘Clustering big data based on distributed fuzzy K-medoids: An application to geospatial informatics,’’ IEEE Access, vol. 10, pp. 20926–20936, 2022

[15] Khedr, W. Osamy, A. Salim, and A. Salem, "Privacy-preserving data mining approach for IoT based WSN in smart city," Int. J. Adv. Comput. Sci. Appl., vol. 10, no. 8, pp. 555–563, 2019.

[16] 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,’’ J. Med. Syst., vol. 42, no. 3, pp. 1–9, Mar. 2018.

[17] Du, C. Jiang, E. Gelenbe, L. Xu, J. Li, and Y. Ren, ‘‘Distributed data privacy preservation in IoT applications,’’ IEEE Wireless Commun., vol. 25, no. 6, pp. 68–76, Dec. 2018.

[18] Nasiri and M. Keyvanpour, ‘‘Classification and evaluation of privacy preserving data mining methods,’’ in Proc. 11th Int. Conf. Inf. Knowl. Technol. (IKT), Dec. 2020, pp. 17–22

[19] Rathod and D. Patel, ‘‘Survey on privacy preserving data mining techniques,’’ Int. J. Eng. Res., vol. V9, no. 6, pp. 832–839, Jun. 2020.

[20] Dhawan S, Chakraborty C, Frnda J, Gupta R, Rana AK, Pani SK, SSII: secured and high-quality steganography using intelligent hybrid optimization algorithms for IoT. IEEE Access 9:87563–87578, 2021

[21] J. KS Das, K Somasundaram, Hybrid optimization-based privacy preservation of database publishing in cloud environment". Concurrency and Computation Practice and Experience 34 (4), 2022

[22] Ranganatha HR , An Enhanced Data Anonymization Approach for Privacy Preserving Data Publishing in Cloud Computing Based on Genetic Chimp Optimization. Int J Inf Secur Privacy (IJISP) 16(1):1–20, 2022

[23] He, Z. Cai, and J. Yu, "Latent-data privacy-preserving with customized data utility for social network data," IEEE Trans. Veh. Technol., vol. 67, no. 1, pp. 665–673, Jan. 2018.

[24] Mohammad, and S. I. S. Al-Hawary, "The effect of supply chain management through social media on the competitiveness of the private hospitals in Jordan," Uncertain Supply Chain Manag., vol. 10, no. 3, pp. 737–746, 2022.

[25] Liu, H. Duan, Q. Zeng, M. Zhou, F. Lu, and J. Cheng, ‘‘Towards comprehensive support for privacy preservation cross-organization business process mining,’’ IEEE Trans. Services Comput., vol. 12, no. 4, pp. 639–653, Jul. 2019.