Slide Window Method Adapted for Privacy-Preserving: Transactional Data Streams

Jayendra Kumar
Assistant professor, CSE Department, Anurag Group of institutions.
Corresponding Author: jayendrakumarcse@cvsr.ac.in

Abstract:
Data streams mining on transactional data is very attractive area for researcher, identical about transactional data makes compromise the privacy of individual so identical information must be removed from transactional data. For publish static transactional streams many privacy-preserving techniques proposed, these methods are not straightforwardly applied on transactional data streams because it have different characteristics. In sliding window addition and removal of transaction leads be unsuccessful to satisfy ρ-uncertainty. To maintain ρ-uncertainty remove the items of window, due this heavy information loss we proposed algorithm which dynamically select items for remove to maintain satisfy ρ-uncertainty with less information loss and continuously make satisfy ρ-uncertainty of slide window with suppression of anonymize sliding window, experimental shows our method is more efficient than batch existing batch processing anonymization sliding window methods.

I. Introduction
Maintain Individual privacy while publishing helpful information many tools and methods needs provide privacy-preserving. A moment ago, industry and academia are pay extensive attention and proposed methods are developed for data publishing situations[1,2,3,4,5,6,7]. Customer transactional data which consist of set of values haggard from list of items [8,9]. With the beginning of big data, data coming continuously with uncontrolled and growing dataset, such called data stream, many data mining patterns can extracted from this this transactional data stream[10], such as frequent item set, which mining from many applications such as online e-commerce, retail chain, web server logs, click stream[11] and computer network. This raw transactional data disclosing person privacy, identical about transactional data makes compromise the privacy of individual so identical information must be removed from transactional data. For publish static transactional streams many privacy-preserving techniques proposed, these methods are not straightforwardly applied on transactional data streams because it have different characteristics. It is more difficult than that for publishing traditional transactional static datasets.

A special data model is used for extraction of transaction from data streams, the data streams is continuously generated, this data model for Data stream mining techniques the damped window model(DW).

Many applications are use sliding window for data stream analysis, so we take privacy problem with sliding window model for data streams. This stream data used by data miners and attacker also, if attacker have knowledge about data in transactional of victim, attacker may find sensitive information of victim, here I takes one example a e-commerce company give this transactional data to a data mining company for finding patterns, let A person from data mining company and his know person B and his brought a item x,y,z from e-commerce company, A find out only one transaction contain item x,y,z, so A easily find its belongs to his friend B. Along with items it have other sensitive data like AIDS test, so B privacy is leaked.

For existing transactional data mining technique are divided into two way that are they organizing data non-sensitive and sensitive [8]. Data stream mining are extract pattern from newly arrived data only. So handling new data is very important in specially data stream environments. Sliding window approach is efficient for data streams mining, the traditionally method for anonymization data streams not suitable for preserving privacy, because the sliding window is frequently updated. Here item set \( S = \{ i_1, i_2, \ldots, i_n \} \) is limited set in which sensitive items are \( S_i \) and non-sensitive are \( S_n \), TID={T\(_1\), T\(_2\),…, T\(_M\)} is infinite transactional, the sliding window represent with \( W_p \) the size is \( x \), where \( p \) represent it updated window with \( p \) size of data. We represent it following fig2.

Fig 1: show data stream mining using sliding widow model
In [18] is introduced a static transactional data privacy model called p-uncertainty, where an attacker knows some item sets \( t \) of transactional data which is monitored, so inferred sensitive data with confidence greater than threshold value. In data stream the sliding window data is deleted and added, to avoid this we require anonymize sliding window transactional data confidence greater than threshold value.

In Fig 2 sliding window step size is 3 and window size is \( x \) which contain non-sensitive data with window is updated mean \( T_1, T_2 \) and \( T_3 \) are deleted and \( T_{x+1}, T_{x+2} \) and \( T_{x+3} \) are added into sliding window, but in newly added transactions have sensitive data so data stream miner can infer sensitive data by making relationship with knowing data so we need to compromise with privacy. Which dynamically and continuously make satisfy p-uncertainty of slide window with suppression and generalization.

### II. Related Work

Preserves privacy in generation of association rules in transactional data addressed by many researchers [2], but they methods based on perturbation [4, 8, 10, 18, 21, 24] techniques, therefore data integrity damaging and false association rules generated [27], but they not stop sensitive rules (inferences), data stream data are continuous and unbounded [25], the Cao et al [25] done some research on relational data streams to privacy-preserving, through I-diversity and k-anonymization [26] cluster framework was presented with delay assurance with maximum between anonymized output and incoming data, in [28] by time constraints on transactional data publication to reflect on k-anonymization move toward for data stream clustering and reuse the cluster, to decrease information loss and speed up the anonymization process in [29] SKY method proposed which constraints for privacy protection, in [30] create cluster on data stream data k-anonymizing approach, [31] delay free anonymization method on health data streams.

Extracting patterns from recent arrived data is very important like stream data, in data mining researcher focus on these tasks. Generated patterns from such kind of data a common approach was sliding window [11], above anonymization methods are applicable to data streams and not satisfies privacy requirements and privacy leakage, in sliding window approach data is updated rapidly so maintain protection of privacy is big problem in [32] wang et al proposed SWAF framework to solve this problem by constantly facilitating k-anonymity on sliding window. In
incrementally anonymize a sliding window under the LK Cprivacy requirements[34] ,with low data loss consider privacy preserving publishing data continuously in sliding window.

III. MinLossSupression Method

With global suppression method void the generation false association rules [37], delete item from all transactions which contain item , in [18] suppress control initially with sliding window method, due to addition and removal data from sliding window it not satisfy the uncertainty,we proposed algorithm to continuously maintain the uncertainty below threshold value of SAR association rules, SAR association are have the sensitive items in either sides rules SAR = \{A \rightarrow \alpha | \alpha \in I_s \}.

In sliding window methods transaction added and some are deleted ,the delete the transaction may give information about newly added transaction so before delete the transaction find the minimum information loss due to suppress the item ,for that find which item are have maximum payoff ,this payoff\(\frac{\text{Count}(b,SAR)}{\text{Sup}(b)\in TSW_{i+1}}\) With respective currently window with new added transactions.Maintain a global suppression item, its frequently updated \(i_{sup}\) with newly find maximum payoff items.

Algorithm: MinLossSupression

Input: TDS stream data, TSW\(_i\), current window transactions, \(\rho\) threshold value

Output: anonymization windown TSW\(_{i+1}\)

1. slide TSW\(_i\) over TDS by add T\(_{add}\)
2. TSW\(_{i+1}\) = TSW\(_i\) + TSW\(_{add}\)
3. Delete \(i_{sup}\) from TSW\(_{i+1}\)
4. SAR = \{A \rightarrow \alpha | \alpha \in I_s \}
5. SAR\(^1\) = \{r | r \in SAR, conf(r) > \rho\}
6. While SAR\(^1\) \neq \\Phi
7. Find max \(\frac{\text{Count}(b,SAR)}{\text{Sup}(b)\in TSW_{i+1}}\) for all \(b \in r\)
8. Delete \(b\) from TSW\(_{i}\), add \(b\) to \(i_{sup}\)
9. Update SAR\(^1\)
10. Return TSW\(_{i+1}\) = TSW\(_{i+1}\) - TSW\(_{del}\)

IV. Results And Discussion

To evaluate our proposed algorithm with SuppressionControlin [1] in terms of efficiency and data quality. Our evaluation with static anonymization with batch processing .we implement our algorithm in Matlab Environment and run on windows7 with 8 GB primary memory(RAM) and four-core 3.2 GHz central processor Unit(CPU).

| Data sets | Size | Items | Max|t| Avg|t|
|-----------|------|-------|----|---|----|

TABLE 1. Two data set Description
In [38] introduce two data set, those are BMS-POS and BMS-WebView-2 these are very point of reference in data mining community. For our experiment we use these data sets. Description of these two data sets characteristics show in table 2, the characteristics are avg|t| and max|t| represents average and maximum size, correspondingly. Items in data sets are not differentiated as non-sensitive and sensitive. We consider 40% of items as sensitive and The default values of the window size w, step size p and privacy requirement are 10000, 50 and 0.5, respectively.

By varying window size w, privacy requirement $\rho$, step size P and two data observe average information losses equation (1) show in Fig 4.

$$\text{AvgInfoLoss} = \frac{\sum_{i=1}^{w} \text{InfoLoss(TSW)}}{w}$$  \hspace{1cm} (1)
Fig 4 The performance by varying window size $w$

(a) Information loss

(b) Execution time
Fig 5 The performance by varying Step Size $p$
in fig 3(a) shows our MinLoss Method is less average information compared to existing high performance SuppressControl Method because payoff of items are calculated after the add new item, with two data sets also. In fig 3(b) execution of algorithm shown, our proposed algorithm give higher performance due to consider the whole sliding window, while our algorithms scan TSW_{del} and TSW_{add} to anonymize the sliding window.

We investing the our algorithm in decrease size of window in fig 4(a) and (b) average information loss and execution time respectively the average information loss is decreased with window size increased because the probability the lower the probability of a SAR violating \( \rho \)-uncertainty is, in fig 5(a) and (b) increase step size \( p \), both algorithms average loss also decreased, compared to existing algorithm our minloss method have less loss because suppression will influence the current sliding window with new add items. In fig 6(a) and (b) shows the performance of algorithms with vary privacy requirement, average information loss is decrease with \( \rho \) increase, and execution time is minimum effect by \( \rho \).

V. CONCLUSION

publish static transactional streams many privacy-preserving techniques methods are not straightforwardly applied on transactional data streams because it have different characteristics. In sliding window addition and removal of transaction leads be unsuccessful to satisfy \( \rho \)-uncertainty which dynamically and continuously make satisfy \( \rho \)-uncertainty of slide window with suppression anonymize sliding window by selecting optimal select items by use of maximal payoff and attacker not able to find knowledge about data and sensitive information of victim.
REFERENCES


PROFILE

Mr. Jayendrakumar is currently working at Anurag Group of Institutions as Assistant professor in Computer Science and Engineering Department. He obtained M.Tech CSE from JNTU Hyderabad. His research interest are Data Mining, Internet of Things and Machine Learning. He is Life Member of Computer Society of India.