DATA PERTURBATION USING RANDOMIZATION APPROACH IN DATA STREAM MINING

UMANG SHUKLA, PARESH SOLANKI

Abstract: Nowadays, lots of data generates from different area of application such as online logs, health care system, telecommunication, electronics and finance. Data mining is associated with finding hidden rule and meaningful relationship that will useful to generalize data in proper end result and users. Releasing and gathering such diverse information belonging to different parties may violate privacy laws and eventually be a threat to individual and business. Privacy preserving data mining strives to provide a solution to this dilemma. It aims to allow useful data patterns to be discovered without compromising privacy. Continuously data arriving that required larger space and more resources. Data Stream arrival rate is high speed which may online or offline, implicitly or explicitly ordered by timestamp, evolving and undetermined in nature. In traditional dataset multiple scan possible but in data stream mining only single scan. Individual and companies concern about releasing sensitive information while sharing between different parties or public domain. In many applications, sharing data is proven to be beneficial for data stream mining application. Our proposed work mainly concern about handling privacy issues in data stream mining. Every time when we execute algorithm for privacy, it will provides different perturbed values with high accuracy result and minimum information loss.

Keywords: Randomization, Data Stream Mining, Privacy Preservation

Introduction: Data Stream in [1] [2] explainsas sequence of continuously arriving data at a very high speed which is online or offline, implicitly or explicitly ordered by timestamp, evolving and undetermined in nature. It has different characteristics from traditional dataset, data steam cannot store because of larger size. It leads to that multiple scan also not possible. Data Stream may keep evolving in data streamsover time. Evolving concepts require data stream processing algorithms to continuously update models to get changes.

As we know data stream change over time because of data arrives from online there is no static storage available and it's never ends, data distribution may change over time, limited or no access to historical data. All changes because of change in environment like financial situation in industry, changing individual characteristics like individual interest rate, complexity of activity.

Application area of data stream is telecommunication, data management, web applications, manufacturing, financial, call records, web page visits, and sensor reading. Traditional data mining techniques are mainly for static and already available datasets. It can be suitable for data warehouses, databases, relational datasets, and transactional datasets. But data streams are continuous, fast, larger volume so we cannot directly apply traditional mining techniques to data steam mining. In recent years, data goes rapidly huge or unlimited rate. It will become hard to store data for mining process because of limited memory storage capability. As per limitation of memory store data steams is also fastinnature so we cannot store past information. It's required to provide privacy before it will available to data mining tasks like clustering, classification and others. Dataset is not available at same time. Algorithm required performing privacy protection to sensitive data with interrupting data flow. Data privacy in data stream in challenging mainly for two reasons [3]: Performance requirement: Flow of coming data stream that cannot be stored in memory. Traditional data mining algorithm is an inapplicable. Time evolution: Data streams are usually changing. Based on this nature, correlations and autocorrelations may change over time. This characteristic makes most traditional algorithms for static data inappropriate. Applications of privacy in data stream mining are finance, health care, retailer and defense. For example two financial companies want to monitor clusters over their real time transaction. Including that none of they want to publish original data. Best solution is that providing data utility and data stream privacy.

Literature Survey: The Perturbation based Privacy preservation techniques are mainly a combination of isometric transformations i.e. translation and rotation transformations. They used with secure random function in order to provide secrecy of user-specified attributes without losing utility in results. Perturbation based techniques are mainly divided into two parts 1) probability distribution approach 2) value distribution approach. In probability approach replaces data with same or distribution itself and in value based approach takes additional noise in data.

Perturbation techniques have two metrics: First is Level of Privacy guarantee and second is Level of model-specific data utility preserved, it is measured by the loss of accuracy for classification and clustering.It may happen that after modify data with perturbation techniques, result data cause misclassification. Motivation for all data perturbation algorithms is to optimize the data transformation process by maximizing both data privacy and data utility achieved. Data privacy is always measured by the level of difficulty in estimating the original data from the perturbed data.

There has been extensive research in the area of statistical databases (SDB) which provides summary statistical information without disclosing individuals' confidential data. The privacy issues arise when the summary statistics are derived from data of very few individuals. A popular disclosure control method is data perturbation which alters individual data in such a way that the summary statistics remains approximately the same as original data. Data mining techniques clustering, classification, prediction and association rule mining are essentially relying on more sophisticated relationships among data records or data attributes but not just simple summary statistics. In the following, we will mainly discuss different perturbation techniques in the data mining area.

Additive data perturbation for building a decision-tree classifier is presented in [4]. In this method random generated noise will be added to data element Yij = Xij + Rij , where Xij= Ith attribute of the jth private data record, and Rij is the corresponding random noise. The resulting distribution of data values is very different from the original distribution. So it becomes tough to accurately estimate original values in individual data records, they proposed a novel reconstruction procedure to accurately estimate the distribution of original data values. The distribution reconstruction process cannot provide exact original values, there is some loss of information, but the authors argue that up to certain level of loss is acceptable in manypractical situations.

Several techniques have been proposed to reconstruct the original data from the perturbed data, in [5] carry out that additional noise can be easily filtered out and privacy will compromise. Proposed random matrix based Spectral Filter (SF) technique to recover original data. Based on data correlations reconstruction is discussed in [6]. Their experiments have shown that when the data correlations are high, the original data can be reconstructed more accurately, so more private information can be disclosed. They proposed two data reconstruction methods based on data correlations: First one is Principal Component Analysis (PCA), and the other used the Bayes Estimate (BE) technique, which in essence is maximum posterior probability estimation.

Another reconstruction method proposes in [7] an improved the Bayesian-based reconstruction procedure by using an Expectation Maximization (EM) algorithm for distribution reconstruction. More specifically, the authors prove that the EM algorithm converges to the maximum likelihood estimate of the original distribution based on the perturbed data. Gaussian technique to generate random noise is proposed in [8] which will be useful for protecting sensitive data with addition of random noise and also used decision tree classification. Proposed technique can provide better privacy gain. Limitation for classification which that method can only applied to numerical data. There are many methods and research works on additive random noise in is random generated noise will be added to original data. Drawback of additive random noise utility for preservation is not clear. There are two different schemas discussed in [9] for Multiplicative random noise is in multiplication relation Yij = Xij * Rij. Also statistical calculation is provided like sum, mean, variance and covariance.

Masking techniques in [10] are mainly classified as perturbation and non- perturbation. Non perturbation techniques applied suppression and add/remove some details. Perturbation based masking techniques are Micro Aggregation, Rounding, Additive Noise and Modified Data Transitive Technique. The first method is Micro Aggregation is one of the perturbation techniques that statistical disclosure control techniques for protecting continuous data. Partition of records in several clusters and one aggregate operator computed for each cluster and replace original values. The second method is rounding: it will replace original value with rounding function value. In a multivariate original dataset, rounding is usually performed one attribute at a time. Rounded values are chosen from a set of rounding points Ri each of which defines a rounding set. The third method is Additive Noise: noise should be added to actual value. Noise being added is typically continuous and with mean zero, which suits well continuous original data. Performance of all above methods is less accurate than proposed method called Modified Data Transitive Technique. Multiplicative randomization technique based perturbationin [11]. Proposed method called Tuple value based multiplicative perturbation. Proposed method consider single row as input which required for data stream approach.

Preliminaries: Data Stream mining: A data stream contains web searches, sensor data, healthcare system data and also network traffic in ordered sequence data. Extracting knowledge from rapid and real time data is called as Data stream mining.

Utility and Privacy: In data mining, preserving data privacy and extracting useful information is difficult task. Perfect privacy can only be achieved with the absolute loss ofthe utility of information. Objective has been to maximize the utility of a data set while minimizing the risk of privacy loss due to the use of the data set.

Randomization: In proposed technique to generate noise based on randomization values are selected for noise generation are randomly each time based on user's range and generated average of the noise will be multiply with respectively sensitive data. Proposed method will select randomly any value for each row individually as per given range by user.

Precision and recall:*Precision*is the ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved. It is usually expressed as a percentage. *Recall* is the ratio of the number of relevant records retrieved to the total number of relevant records in the database. It is usually expressed as a percentage.

Privacy Preservation: Sharing sensitive data in public domain for mining approach threaten for everyone. Apply such an algorithm that can provide privacy to sensitive information and also maintain output result of mining approach same as original data.

Proposed Method: In this concept, considers all attribute except sensitive attribute value and class label to calculate noise. We have proposed technique to generate noise based onrandomization. Values selected for noise generation are random each time based on user's range. Range value random values selected for noise generation and generated noise will be multiply with sensitive data. Now we can apply clustering simple k-means clustering algorithm in data stream mining for performance analysis. This method will select randomly any value for each row individually as per given range by user.Every timesensitive value replaced by selected noise range is not fixed so, we can say that this method is provides better result every runtime.

Perturbed datastreamsshouldgenerateidentical result as of original data stream.

$\frac{1}{W}\sum_{n=1}^{W}(Ds_{n\quad T\times N}) = Ds'_{T\times N}$where, w = size of total $Ds_{T\times N}$consider for replacement according user define Drifting value.

Procedure: Random Selection Based Multiplicative Data Perturbation Technique (RSBMDP)

Input: Data Stream D.

Intermediate Result: Perturbed data stream D'.

Output: Clustering Result R' of Data stream D'.

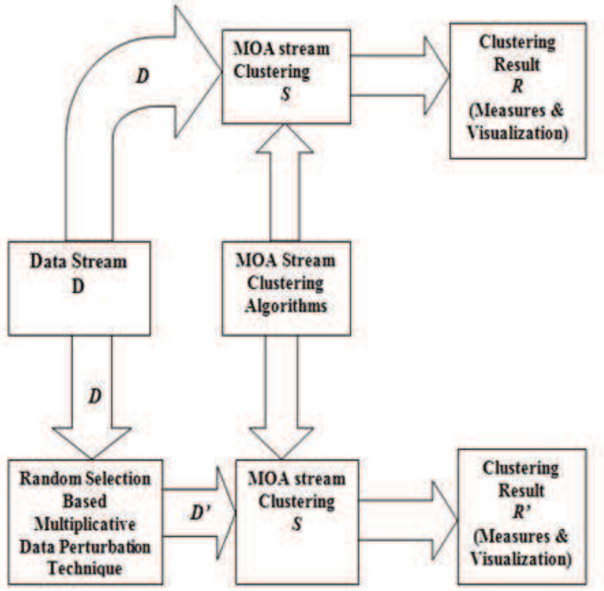
MOA stream Clustering S
Clustering Result R (Measures & Visualization)
D
Data Stream D
MOA Stream Clustering Algorithms
D
Random Selection Based Multiplicative Data Perturbation Technique
D'
MOA stream Clustering S
Clustering Result R' (Measures & Visualization)

Figure1: Proposed Framework

Steps:

```
1. For each instance I in Dataset D
2. Row Value=0
3. Sensitive Data=0
4. user Range=0
5. For each Attribute Aj in Instance (I), where J= 1, 2, 3, 4, . . . . . ., TotalAttributes
   • Store attribute values in array result []
   • rowValue= randomAccess(results[],userRange)
6. End for
7. rowValue= rowVlaue * sensitiveData
8. Clustering (I)
9. End for
```

Experiments and Results: Evaluation approach focused on the overall quality of generated clusters based on dataset perturbation. In this section, steps defined how data stream mining with proposed perturbation technique works.

- Setup each dataset as stream in MOA (Massive Online Analysis) framework.
- Define sliding window (W) over the data stream to evaluate measures and cluster membership matrix.
- Apply our proposed data perturbation method to protect the sensitive attribute value to provide privacy.

- K-Means clustering algorithm is used to find the clusters for our performance evaluation. K-Means is one of the best known clustering Algorithms and also scalable. Number of cluster to be find from original and perturbed dataset was taken same as number of cluster.
- Compare how closely each cluster in the perturbed dataset matches its corresponding cluster in the original dataset. Quality of the generated clusters by computing the F-measure.
- Experiments were performed to measure accuracy while protecting sensitive data. We here presents two different results, one is corresponding to clustering accuracy in terms of membership matrix which was manually derived from clustering result and another represent corresponding graph for F1_P(precision) and F1_R(Recall) measures provided by MOA framework.
- Datasets shows in table1 with configuration to determine the accuracy of our proposed method. In results, each dataset to determine 5 clusters using K-Means clustering algorithm.

Dataset	Total	#	Nominal	Attribute Protected
Electric norm[13]	45312	45k	Ignored	Nswprice
Bank Marketing[14]	45,211	45k	Ignored	Balance, Duration

Table1: Dataset

In evaluation approach, we focused on overall quality of generated clusters. Compared how closely each cluster in the perturbed dataset matches its corresponding cluster in the original Dataset. First need to identify the matching of cluster by computing the matrix of frequencies from shown in below table. We refer to such a matrix as the Clustering Membership Matrix (CMM), where the rows represent the clusters in the original dataset, the columns represent the clusters in the perturbed dataset, and Freqi , j is the number of points in cluster Ci that falls in cluster Ci’ in the perturbed dataset. After computing the frequencies Freqi, j, we scan the CMM to calculate precision, recall, and F-measure for each cluster Ci’ with respect to Ci in the original dataset.

	C1’	C2’		Cn’
C1	Freq 1 , 1	Freq 1 , 2	......	Freq 1 , n
C2	Freq 2 , 1	Freq 2, 2	......	Freq 2 , n
:	:	:	......	:
Cn	Freq n, 1	Freq n , 2	......	Freq n , n

Table2: CMM Matrix

Original dataset D clustering results compared to perturbed dataset D’ at sliding window W=3000.

Below reading considering different range and sensitive_driftvalue based accuracy comparison with original dataset D and perturbed dataset D’. We also considers different window size W =3000.

User Percentage	Trials	
	First	Second
40	99.46	99.45
50	99.83	99.84
60	99.84	99.84
80	99.84	99.84
100	99.84	99.93

Table3: Balance sensitive attribute D to D’ accuracy at window size w=3000

User Percentage	Trials	
	First	Second
40	94.98	93.6
50	95.3	95.3
60	94.2	97.6
80	94.9	96.71
100	96.51	98.64

Table4: Duration sensitive attribute D to D’ accuracy at window size w=3000

User Percentage	Trials	
	First	Second
40	94.44	95.7
50	94.56	96.6
60	95.78	93.84
80	95.69	97.32
100	91.55	97.21

Table5: NswPrice sensitive attribute D to D’ accuracy at window size w=3000

For Accuracy MOA (Massive Online Analysis) framework provides different options. But we mainly focused on two important measures F1_P and F1_R. F1_P determine the precision of system by considering the precision of individual cluster. F1_R determine the recall of system, which take into account the recall of each cluster. Results are presented in terms of graphs for each modified attribute and also for different range & sensitivity values. K-Means is applied in order to evaluate with number of clusters fix (K=5). Instances are processed in defined sliding window size.

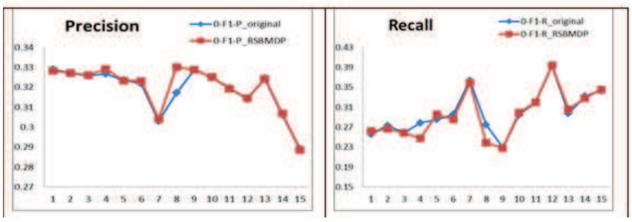

Figure2: Accuracy on attribute Balance with userPercentage=40

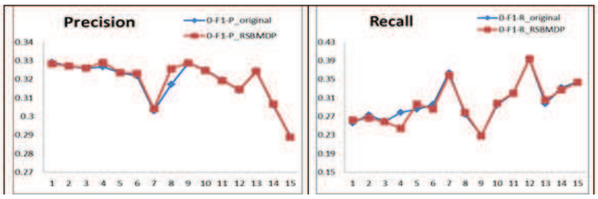

Figure3: Accuracy on attribute Balance with userPercentage=60

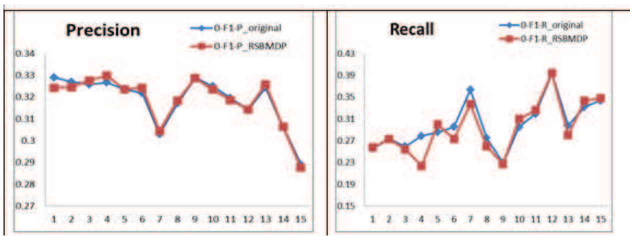

Figure4: Accuracy on attribute Duration with userPercentage=80

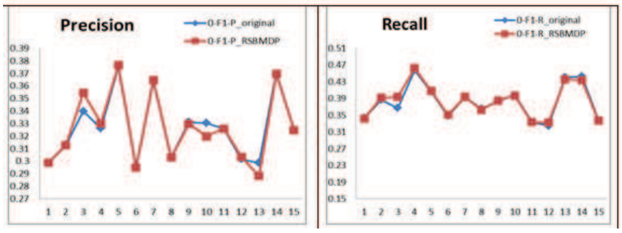

Figure5: Accuracy on attribute NswPrice with userPercentage=40

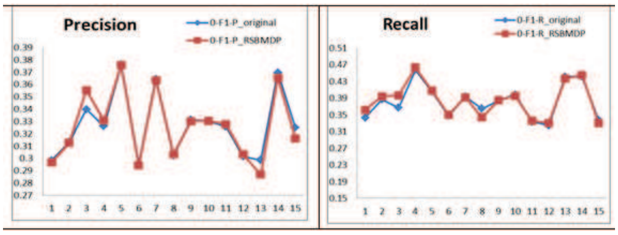

Figure6: Accuracy on attribute NswPrice with userPercentage=80

We have different results of precision and recall measurements on experimental datasets. From above analysis results, sensitive attributes which can lead us to good proposed algorithm work. For other sensitive attributes in given dataset it will follow same characteristics.

Conclusion: Proposed work based on prior privacy preservation which applied in stage of knowledge extraction process. For calculation of accuracy, we applied K-Means clustering algorithm with sliding window size on original and perturbed data. As per our experiments, we gain a fairly good level of privacy and reasonable accuracy in almost all cases. We used three different datasets for our experiments. It also shows that proposed method is an efficient, flexible and easy-to- use method in PPDM. We quantified the privacy of our scheme using the concept of misclassification error. Information loss due to data perturbation was quantified by a loss of accuracy, which can be quantified by percentage of instances of data stream are misclassified using cluster membership matrix. It can also be evaluated against classification algorithms. Our experiments are limited to only numeric type attributes. This work can be extended for nominal type attributes and different datasets.

References:

1. M. Kholghi and M. Keyvanpour “An Analytical Framework For Data Stream Mining Techniques Based On Challenges And Requirements”, International Journal of Engineering Science and Technology (IJEST), Vol. 3, No. 3, pp.2507-2513, Mar 2011.
2. K. Patel “Privacy-Preserving Data Stream Classification: An approach using MOA
3. F. Li, J. Sun, S. Papadimitriou, G. Mihala and I. Stanoi, “Hiding in the Crowd: Privacy Preservation on Evolving Streams through Correlation Tracking”, in Proc.23rd IEEE International Conference on Data Engineering, pp. 686-695, 2007.
4. R. Agrawal and R. Srikant “Privacy- Preserving Data Mining” In Proceedings of the ACM SIGMOD International Conference on Management of Data. Dallas, Texas, pp.439-450, 2000.
5. Dhaneswar Kalita, Effect of Deposition Temperature on Structural and Carrier Mobility of thermally Evaporated Nanocrystalline Cdse Thin Films; Engineering Sciences international Research Journal: ISSN 2320-4338 Volume 3 Issue 1 (2015), Pg 48-55
6. H. Kargupta, S. Datta, Q. Wang and K. Sivakumar “Random data perturbation techniques and

privacy preserving data mining". Knowledge and Information Systems, 7:387-414, 2005.

7. Z. Huang, W. Du, and B. Chen "Deriving private information from randomized data" In Proc. of ACM SIGMOD Conference, pp. 37-48, 2005.
8. D. Agrawal and C. Aggarwal "On the design and quantification of privacy preserving data mining algorithms", In Proceedings of the 20th ACM SIGACT SIGMOD SIGART Symposium on Principles of Database Systems, pp 247-254, 2001.
9. Naresh Kandakatla, Geetha Ramakrishnan, Rajasekhar Chekkara, Pharmacophore, 3d-Qsar Modeling Studies of Hdac2 Ligands and insilico Search for New Hits; Engineering Sciences international Research Journal: ISSN 2320-4338 Volume 3 Issue 1 (2015), Pg 56-67
10. P. Kamakhi and A. VinayyaBabu "Preserving the privacy and sharing the data using classification on perturbed data", International Journal of Computer Science and Engineering, Vol.2, No. 3, pp. 860–864,2010.
11. B .Pandya, U.K. Singh, "An Overview of Traditional Multiplicative Data Perturbation", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Vol. 2, Issue 3, pp.424-429, March 2012.
12. Dr.A.Tamilarasi and S. Vijayarani "A New Technique for Protecting Sensitive Data and Evaluating Clustering Performance" International Journal of Information Technology Convergence and Services (IJITCS), Vol.1, No.2, pp.7-18 April 2011.
13. G Rajendar , Dr. Basavaraja Banakara , Obadhya John Raj.K, New Algorithm for Capacitor Placement to Improve Voltage Stability Using L-index Sensitivity Matrix; Engineering Sciences international Research Journal: ISSN 2320-4338 Volume 3 Issue 1 (2015), Pg 68-76
14. H. Chhinkaniwala and S.Garg "Tuple Value Based Multiplicative Data Perturbation Approach To Preserve Privacy In Data Stream Mining" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.3, pp. 53-61, May 2013.
15. MOA-Datasets, http://moa.cs.waikato.ac.nz/datasets
16. A.Sofi, B.R.Phanikumar, Tarun Sama, Mechanical Properties of Concrete Containing High Volume Pond-Ash and Steel Fibre; Engineering Sciences international Research Journal: ISSN 2320-4338 Volume 3 Issue 1 (2015), Pg 77-81
17. Bank-Marketing-Dataset, http://archive.ics.uci.edu/ml/datasets/Bank+Marketing

***

Umang Shukla/Assistant Professor/Department of Information Technology/
SAL College of Engineering/Ahmedabad/ India/
Paresh Solanki/ Assistant Professor/ Department of Computer Engineering/
U. V. Patel College of Engineering/ Mehsana/ India/