Optimizing the Trade-off Between Classification Accuracy and Data Privacy in the Area of Data Stream Mining

Ullusu Hewage Waruni Amali Hewage [1] Russel Pears [2] and M. Asif Naeem [3]

[1]School of Engineering, Computer and Mathematical Sciences
Auckland University of Technology
Auckland, New Zealand
waruni.hewage@aut.ac.nz

[2]School of Engineering, Computer and Mathematical Sciences
Auckland University of Technology
Auckland, New Zealand
russel.pears@aut.ac.nz

[3]Department of Computer Science
National University of Computer and Emerging Sciences
Islamabad, Pakistan
asif.naeem@nu.edu.pk

ABSTRACT

Data perturbation has grabbed the attention of data mining, as preserving the privacy of the data is crucial, especially in sensitive data. But the perturbation process negatively affects the accuracy of predictions, generating a trade-off between privacy and accuracy. We propose seven different cumulative noise addition based perturbation methods combining a set of techniques such as logistic function, use of absolute noise values, and cycle-wise noise addition as possible solutions for this accuracy-privacy trade-off issue. These techniques are introduced to optimize the trade-off between classification accuracy and data privacy by controlling the maximum noise level.Moreover, we evaluate the performance of the proposed methods compared to the state-of-art of the noise addition-based perturbation methods to select the best of them.

Keywords: perturbation, random projection, cumulative noise, logistic, privacy accuracy trade-off.

Computing Classification System (CCS):
[500]Security and privacy Privacy protections [100]Security and privacy Social aspects of security and privacy [500]Information systems Data stream mining
Mathematics Subject Classification (MSC): 68T30-Knowledge representation, 68U35-Computing methodologies for information systems

1 Introduction

People release their personal data in various situations such as medical check-ups, requesting a bank loan, and applying for employment in their day-to-day lives. The organizations use these data for

increasing performance by making predictions. Assume that Sam has done a medical check-up a few months back which revealed that he has AIDS which he wanted to keep as a secret from society. But suddenly his neighborhood gets to know about this situation because a person who personally knows Sam has participated in the data analyzing process of that health organization has identified Sam from his personal details and revealed about his situation. Our data is not protected any more, and people can access this data to attack our personal lives. Therefore, organizations need a method to protect their customers' personal data when they use those to make predictions/ analyze performance.

Privacy-Preserving Data Mining (PPDM) helps to protect the privacy of data when it is being used for data mining purposes. Data perturbation is one of the techniques that fall under PPDM which is suitable for data streams, and it alters the original data values which makes it difficult to recover by unauthorized people using recovery techniques but still manages to maintain the relevant properties of the data set which are useful for the data mining purposes. It converts data to another form, so anyone cannot identify individuals by looking at their personal data.

One of the critical success factors of Data Mining is the availability of high-quality data that will support the generation of accurate models. On the other hand, sensitive data cannot be published in its original form and thus different types of data perturbation methods have been proposed to maintain privacy. However, data perturbation can negatively affect the accuracy of prediction models. When data perturbation techniques are applied to increase the privacy of data, it decreases the accuracy of the classification models as perturbation distorts the original data values, and this trade-off between data privacy and classification accuracy is an inherent problem that needs to be investigated in this area.

When we try to increase privacy it decreases the accuracy and vice versa which is the most common issue in PPDM. We tried to find out a suitable perturbation method that minimizes the difference between privacy and accuracy values. This ultimately optimizes the accuracy-privacy trade-off. The objective of our research work is to propose a method to optimize the trade-off between accuracy and privacy to enhance the performance of data mining tasks. To achieve the mentioned objective, this paper proposes seven variations of cumulative noise addition methods which combines novel techniques such as logistic function, cycle wise noise addition, noise resetting, and absolute noise values into an existing perturbation method called "Random Projection-based Cumulative Noise Addition" proposed by (Denham, Pears and Naeem, 2020).

Random projection-based cumulative noise addition (Denham et al., 2020) is the most recent work done in the area of data perturbation using noise addition. It adds noise cumulatively to the data stream, while traditional noise addition adds the noise independently to each record. This method was able to achieve high privacy and accuracy values, which optimizes the accuracy-privacy trade-off(More details will be discussed in the "Existing State-of-the-Art Work" section). But there are some practical issues with this method in a long-term run, as this has been designed to apply to data streams. The main concern is that we cannot continue adding the noise cumulatively for a long time, since noise can overpower the actual data values after some time. If this happens, it can drastically decrease the accuracy because data values become highly distorted in the long-term run and the classifier tends to learn the noise instead of data. Our work is motivated by this issue,

and we carried out this research to find out possible ways to improve the work done by (Denham et al., 2020), by controlling the maximum noise added to the stream.
It is worth emphasizing that we mainly focused on developing an advanced perturbation method that can be used with any classification algorithm, which is suitable for data stream mining. We used Adaptive Random Forest(ARF) as the classifier for the experiments and measure the accuracy and privacy for cumulative noise addition in cooperated with different techniques to control the total noise. The main contribution of this research is an improved noise addition-based perturbation method that can be used to optimize the accuracy privacy trade-off in a data streaming environment. As a summary, the paper makes the following contributions to the field:

- Introducing seven different variations of cumulative noise addition methods which can be used as noise addition-based data perturbation techniques.
- Developing algorithms for Linear Cumulative and Logistic Cumulative Methods.
- An effective cumulative noise addition-based perturbation method, which can be used to optimize the trade-off between data privacy and classification accuracy.
- An evaluation of the performance of seven different cumulative noise addition methods concerning the state-of-the-art, using relative error and breach probability for different cycle sizes and growth rate values.
- A vulnerability analysis of the best performing perturbation method from the experimented seven variations of methods.

The remainder of this paper is organized as follows; Section 2 provides a review of the existing data perturbation methods, highlighting the strengths and limitations of each method. The existing work of art is being presented in section 3 and this is the baseline of our research work. Section 4 consists of the proposed methodology, that describes two main types of cumulative noise addition methods. The experimental setup including data sets is being described in section 5, and section 6 presents an extensive analysis of the results. Section 7 and 8 provide a discussion and conclusion and future work of this work, respectively.

2 Related Work

Data mining is a prominent area in today's world which uses data to make predictions to improve the performance of organizations by making correct decisions. This includes building learning models using supervised (classification and regression) and unsupervised learning (clustering and association) approaches. Modeling using different learning algorithms has been discussed in research work such as (Raziyeh Zall, 2019) which uses Multi-Relational Classifier Based on Canonical Correlation Analysis, (Pozna and Precup, 2014) which applies the signatures to expert systems modeling using rules and (Ahmed, Brickman, Dengg, Fasth, Mihajlovic and Norman, 2019) that investigates on using different machine learning approaches to classify pedestrians' events based on IMU and GPS. These works prove that the data mining models can be successfully used in various application areas. We only focus on classification and more specifically classification algorithms which can be used to learn

from data streams. Massive Online Analysis (MOA) (Bifet, Kirkby, Kranen and Reutemann, 2012) has developed a set of algorithms such as Hoeffding Tree, Hoeffding Adaptive Tree, and Adaptive Random Forest (Gomes et al., 2017) which can be used in data streaming environments.

Different types of data perturbation methods have been proposed in the literature to enhance the privacy of data in the past few decades. Data perturbation methods such as random rotation (Chen and Liu, 2005a), random projection (Liu, Kargupta and Ryan, 2006) and geometric perturbation (Chen, Sun and Liu, 2007) maintain the pair-wise distances of the records which are helpful for data mining tasks while methods such as additive noise (Agrawal and Srikant, 2000) and condensation (Aggarwal and Yu, 2004) maintain the data set properties but sacrifice the record-wise properties.

In additive noise or randomization, a high privacy level can be witnessed when the noise variance is increased since original values are highly distorted, but consequently results in a low accuracy level. Though this method distorts the original data values massively, it is still vulnerable to different types of privacy breaching attacks. In (Giannella, Kargupta and Liu, 2008) authors have discussed five attack techniques that can be used against randomization. This includes Spectral Filtering (Kargupta, Datta, Wang and Sivakumar, 2004) Singular Value Decomposition (Guo, Wu and Li, 2006) and Principal Component Analysis (PCA) (Huang, Du and Chen, 2005) which are based on Eigen analysis to recover the original data from perturbed data. In addition to these, Maximum A Posteriori attack (MAP) (Huang et al., 2005) and Distribution attack (Giannella et al., 2008) were also discussed as possible attack types to filter out original values from additive noise.

Random rotation or rotation perturbation was proposed to maintain the record-wise properties in a data set to obtain a high accuracy/ utility level while preserving the privacy of the data. In (Chen and Liu, 2005b) rotation perturbation was defined as a matrix multiplication that multiplies the original data matrix by a rotation matrix which results in a perturbation matrix with the same number of records and features as the original data matrix. This transformation is orthogonal and hence the distance between perturbed records is similar to the distance between original records, and this implies that the perturbed data set gives similar classification accuracy to the original classification results. The accuracy of the classification results is high, but several privacy-breach attacks have worked well with this method too. Distance inference attacks (Chen et al., 2007), Independent Component Analysis (ICA) (Chen et al., 2007) and (Liu, Giannella and Kargupta, 2006), Known Input/ Output attack (Liu, Giannella and Kargupta, 2006), (Giannella, Liu and Kargupta, 2013) and PCA (Liu, Giannella and Kargupta, 2006) can be used to reconstruct the original data from rotation perturbation. It has been proved that the random rotation can be perfectly reversed using a few known input/output pairs (Liu, Giannella and Kargupta, 2006).

Modified and combined versions of rotation perturbation have been proposed to overcome the privacy issue of rotation perturbation. A combination of random rotation and randomization that addresses the distance inference attack known as the general linear transformation was proposed by (Guo and Wu, 2006). Authors (Liu, Wang and Zhang, 2009) investigated privacy vulnerabilities and found out the proposed method is vulnerable to attacks in case of available background information. Random rotation, followed by a translation that addresses attacks to the rotation center, is proposed by (Chen et al., 2007). Attacks based on background knowledge have been developed for this method by (Giannella et al., 2013).

Random projection uses the technique of matrix multiplication and has been proposed by (Liu, Kargupta and Ryan, 2006) to address the privacy issues that arose from random rotation while maintaining the distance between records to achieve a high accuracy level of the classification results. "Random projection refers to the technique of projecting a set of data points from a high-dimensional space to a randomly chosen lower-dimensional subspace" (Liu, Kargupta and Ryan, 2006) by multiplying with a random matrix. The main idea of random projection, motivated by the Johnson-Linden Strauss Lemma (Liu, Kargupta and Ryan, 2006). This Lemma allows decreasing the dimensionality while maintaining the pair-wise distance of two points within an arbitrary small factor (Liu, Kargupta and Ryan, 2006). Though the random projection is not vulnerable to an Independent Components Analysis (ICA) based attack because the reduced dimensionality of the perturbed data set results in an under-determined system of linear equations, it is still vulnerable in cases where the attacker is equipped with prior knowledge about the original data set.

Attacks to the random projection method were discussed in surveys such as (Giannella et al., 2008) and (Okkalioglu, Okkalioglu, Koc and Polat, 2015) and these attacks are based on some level of prior knowledge of the original data. Known input/output attack and known projection matrix attack are two attack types that can be used against random projection. The general idea of the known input/output attack is that the attacker has prior knowledge of a few original records and their respective perturbed records, and using those known record pairs the rest of the original data records can be recovered. Authors of (Liu, 2007) have discussed the known input/output-based MAP attack that can be used to attack random rotation perturbation even when a collection of input/output pairs is less than the number of features of the original data set is known. In (Sang, Shen and Tian, 2012) authors have proposed to shuffle records before publishing to rectify this problem, but unfortunately, the method cannot easily be adapted to a data streaming environment.

A novel research "Random-projection based cumulative noise addition" which combines random projection, noise addition, and translation was proposed by (Denham et al., 2020) to enhance the privacy of the perturbed data using random projection. Instead of traditional additive noise, authors have proposed a novel noise addition method that is called cumulative noise addition. It has been proved in this research work that it is possible to achieve a considerable level of privacy and accuracy by combining these perturbation techniques. But the system has to face a huge amount of noise that may overtake the original data values when the data stream unfolds and that can lead to reducing the accuracy is the main issue of this scenario.

Though the perturbation methods in the literature have proposed different techniques to increase the privacy level of data while maintaining the accuracy of the classification results, it does not seem to be completely achieved by the existing related work. The techniques which provide better privacy level are lacking the expected accuracy of the data mining results, and the methods which provide significance accuracy is vulnerable to different types of privacy attacks. Therefore, it still requires a method that can optimize the trade-off between classification accuracy and data privacy for the betterment of the field.

2.1 Existing State-of-the-Art Work

After deeply analyzing the experimental results of the method, random projection-based cumulative noise addition (referred to as LRW) proposed by (Denham et al., 2020) which is the most advanced technique developed in this area is used as the base perturbation method for this study.

In the process of random projection, the original data matrix $X(mxn)$ is multiplied by a random matrix R (k x m) to generate a perturbation matrix $Y(kxn)$. Each element of R is independent and identically distributed (i.i.d.) and is generated from a Gaussian distribution with mean 0 and variance σ_r^2 (Liu, Kargupta and Ryan, 2006). The projection process is represented as, $Y = \frac{1}{\sqrt{k\sigma r}}RX$ and the multiplication by the factor $\frac{1}{\sqrt{k\sigma r}}$ ensures that the column-wise inner product is preserved (records are represented as columns). According to the Johnson-Linden Strauss Lemma, random projection can be considered as an approximately distance preserving perturbation. The problem with the distance preserving perturbation is that the records closer to the origin are less perturbed than records far away from the origin (Chen et al., 2007). This allows uncovering some original records easily, even without a complex attack. To avoid this vulnerability issue, (Chen et al., 2007) have proposed a random translation method, which applies the same translation to each record. This extends the perturbation method to $Y = \frac{1}{\sqrt{k\sigma r}}RX + \Psi$. "Applying a constant translation to all records has no effect on many data mining tasks, but an attacker must sacrifice one known input/output pair to account for it" (Denham et al., 2020).

To add an additional degree of uncertainty to any recovery attempt that attempts to reverse random projection, the authors of (Denham et al., 2020) have introduced two types of noise addition, namely independent noise (randomization) and cumulative noise. For our study, we are focusing on the cumulative noise addition, since it has been successful in reducing the trade-off between privacy and accuracy to some extent. In the process of cumulative noise addition, i.i.d. Gaussian noise values are added to each record, but additionally, each of these random values is also added to every subsequent record in the stream and can be represented using $Y = \frac{1}{\sqrt{k\sigma r}}RX + \Psi + \lceil$. The cumulative noise addition is useful for resisting known input/output attacks, since the attacker has to face increasing levels of noise when the data stream unfolds. This allows creating the same impact with a small variance of cumulative noise as opposed to a large variance setting with independent noise. "Cumulative noise is designed to achieve a similar privacy benefit as independent noise, but with less impact on the accuracy of data stream mining algorithms" (Denham et al., 2020).

By analyzing the experimental results, when compared to independent noise addition, cumulative noise addition has provided low classification error with a marginally higher breach probability. Getting a low classification error is a result of cumulative noise addition which adds a small variance of noise and hence the distortion of original values is also very small. However, the privacy level of cumulative noise addition is expected to increase with time as the data stream unfolds. When considering the privacy accuracy trade-off, cumulative noise addition outperformed the independent noise addition method.

As we consider the data streaming environment, the stream has to face a huge amount of noise in a cumulative noise addition environment when the data stream unfolds. It is a good approach when considering the privacy aspect, but can negatively affect the accuracy as the classifier tries to learn the noise values rather than the original values because of the high distortion of the original

data values. Therefore, a method that can control the maximum amount of noise added to the stream needs to be combined with the cumulative noise addition. Then it will help to achieve a high accuracy level by controlling the maximum noise level while still adding it cumulatively to maintain the high privacy level.

3 Proposed Approach

Facing an enormous amount of noise with the time when the data stream unfolds is an inherent issue that can be experienced in any kind of cumulative noise additive environment. Therefore, a technique that controls the maximum noise level added to the stream needs to be incorporated to enhance the accuracy without making a considerable effect on privacy.

To achieve this objective, two main categories of cumulative noise addition named Linear Cumulative Noise Addition and Logistic Cumulative Noise Addition were introduced. Under these two main approaches, different techniques such as cycle-wise noise addition, use of absolute noise values, and noise resetting method were performed to investigate the behavior of privacy and accuracy. The ultimate objective of using these different techniques is to control the maximum noise level, but still adding it cumulatively. By doing so, it is expected to increase the accuracy level while maintaining the high privacy level provided by the cumulative noise addition and hence optimizing the trade-off between privacy and accuracy. We used the state-of-the-art (Denham et al., 2020) as the base of our research and expanded and improved it by cooperating with possible techniques that help to minimize the accuracy-privacy trade-off. Table 1 summarizes all the symbols we used to build up the algorithms in section 3.1 and 3.2.

Table 1: Symbol Table

Symbol	Meaning
Y	Perturbed Dataset
R	Random Projection MAtrix
X	Original Dataset
Ψ	Cumulative Noise
$\lceil$	translation Matrix
ϖ	Logistic Cumulative Noise
cs	cycle Size

3.1 Linear Cumulative Noise Addition Methods

Techniques such as cycle-wise noise addition, noise resetting, and using absolute noise values have cooperated with the linear cumulative noise addition method to control the noise level of the stream. Following are the detailed description of four different variations experimented with, including the base method.

- LRW (Linear Random Walk without Resetting) - Cumulative noise is added in a random walk fashion. No absolute values and no noise value resetting are used. Note - This is the base

method of the research (Denham et al., 2020)

  * $Y = \frac{1}{\sqrt{k\sigma r}} RX + \Psi + \lceil$

- LRWR (Linear Random Walk with Resetting) - Cumulative noise is added cycle-wise in a random walk. No absolute values are used. At the end of each cycle, the noise level is reset to zero. (cs is the cycle size.)

  * $Y = \frac{1}{\sqrt{k\sigma r}} RX + \sum_{i=-cs}^{=cs}(\Psi) + \lceil$;when i=cs, $\Psi = 0$

- LAR (Linear Absolute with Resetting) - Cumulative noise is added in cycles using absolute noise values. In the first half and second halves of the cycle, noise values were added and subtracted, respectively. At the end of each cycle, the noise level is reset to zero.

  * $Y = \frac{1}{\sqrt{k\sigma r}} RX + (\sum_{i=-cs}^{=0} + abs(\Psi), \sum_{i=0}^{=cs} - abs(\Psi)) + \lceil$;when i=cs, $\Psi = 0$

- LA (Linear Absolute without Resetting) - Cumulative noise is added in cycles using absolute values. As with the scheme above in LAR, but with no noise resetting at the end of the cycle.

  * $Y = \frac{1}{\sqrt{k\sigma r}} RX + (\sum_{i=-cs}^{=0} + abs(\Psi), \sum_{i=0}^{=cs} - abs(\Psi)) + \lceil$

Figure 1 graphically illustrates the four variations of linear cumulative noise addition methods.

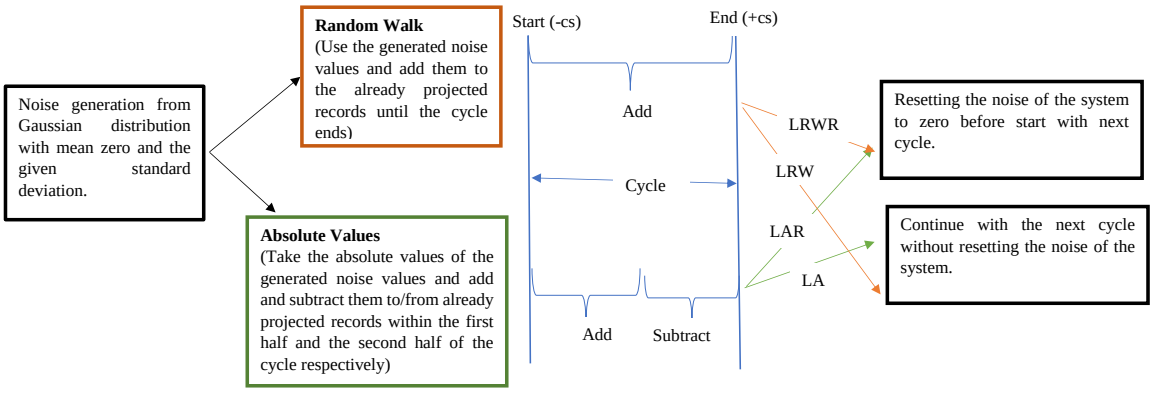

Figure 1: Process of Linear Cumulative Noise Addition

The entire process involved with these four variations of linear cumulative noise addition can be explained using the following pseudo-code (See Algorithm 1).

3.2 Logistic Cumulative Noise Addition Methods

Logistic cumulative noise addition is influenced by the well-known logistic function (Verhulst, 1838) (illustrated in Figure 2) and the expectation of using this concept in a cumulative noise addition environment is to further control the noise in the system but still adding it cumulatively. The mathematical representation of the logistic function is shown below. L = maximum value of the

Algorithm 1 Pseudo Code - Linear Cumulative Noise Addition Methods

Input: Cycle Size (cs), Cumulative Noise Sigma (σ), Projected Dataset (X)

Process: **While** End of X

for each Record i in each cycle (defined by cs)

do

Generate noise (n) from Gaussian distribution $N(0, \sigma^2)$

Add n cumulatively to the records $(y_i = x_i + n_i)$

Use n_i **OR** absolute n_i

End Cycle

System Noise = 0 **OR** System Noise = Current Cumulative Noise

Continue Next Cycle

Output: Perturbed Dataset (Y_i)

function, e = Euler's number, k = logistic growth rate and x_0 = x value of the mid-point has been used here.

$$\dot{\mathbf{f}}(x) = \frac{L}{1 + e^{-k(x-x_0)}} \tag{3.1}$$

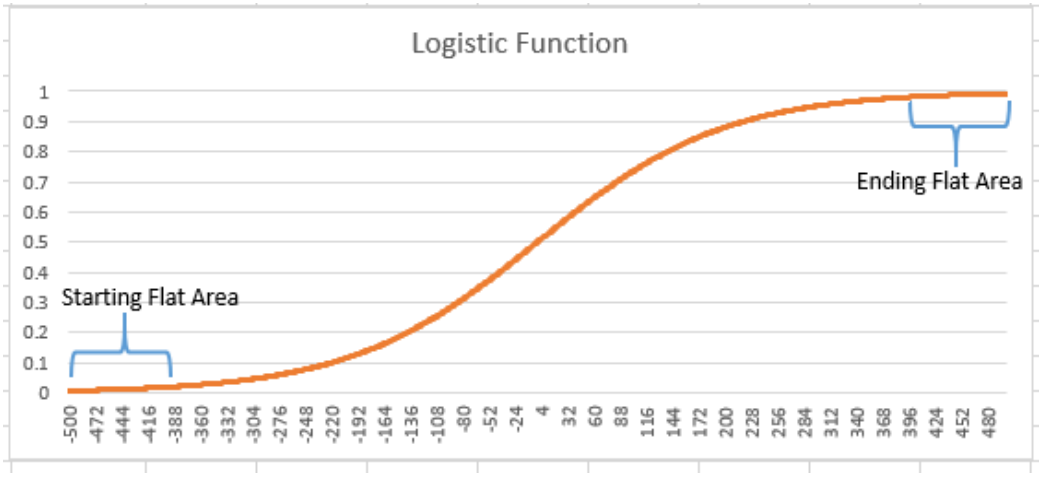

Figure 2: Logistic Curve
(Verhulst, 1838)

In this logistic cumulative noise addition environment, the noise was generated from a Gaussian distribution with mean zero and the standard deviation $f(x) \times \sigma$ (return value of the logistic function multiplied by the standard deviation of cumulative noise). From here onward, we refer to the noise generated according to the logistic process as ϖ. Since the logistic function allows controlling the growth rate and the maximum value returns from the function, it provides an effective way to control the noise addition rate and the maximum noise level, respectively. It also provides a good platform for cycle-wise noise addition. Therefore, in cooperating logistic function with cumulative noise addition appears to be a promising technique to control cumulative noise level that can positively affect to the accuracy hence allows optimizing the trade-off between privacy and accuracy. Here are the four variations of the logistic cumulative noise addition method which was experimented with.

- SRW (Logistic Random Walk without Resetting) - Noise is added cycle-wise in a random walk fashion. No absolute values and no noise value resetting is used

  * $Y = \frac{1}{\sqrt{k\sigma r}} RX + \sum_{i=-cs}^{=cs}(\varpi) + \lceil$

- SRWR (Logistic Random Walk with Resetting) - Logistic noise is added cycle-wise in a random walk fashion. No absolute values are used. At the end of each cycle, the noise level is reset to zero.

  * $Y = \frac{1}{\sqrt{k\sigma r}} RX + \sum_{i=-cs}^{=cs}(\varpi) + \lceil$;when i=cs, $\varpi = 0$

- SAR (Logistic Absolute with Resetting) - Logistic noise is added in cycles using absolute values. As with LAR, at the end of each cycle, the noise level is reset to zero.

  * $Y = \frac{1}{\sqrt{k\sigma r}} RX + (\sum_{i=-cs}^{=0} + abs(\varpi), \sum_{i=0}^{=cs} - abs(\varpi)) + \lceil$;when i=cs, $\varpi = 0$

- SA (Logistic Absolute without Resetting) - Logistic Noise Addition in cycles using absolute values. As with LA, no noise value resetting is used.

  * $Y = \frac{1}{\sqrt{k\sigma r}} RX + (\sum_{i=-cs}^{=0} + abs(\varpi), \sum_{i=0}^{=cs} - abs(\varpi)) + \lceil$

A graphical representation of the steps conducted in logistic cumulative noise addition methods are shown in Figure 3

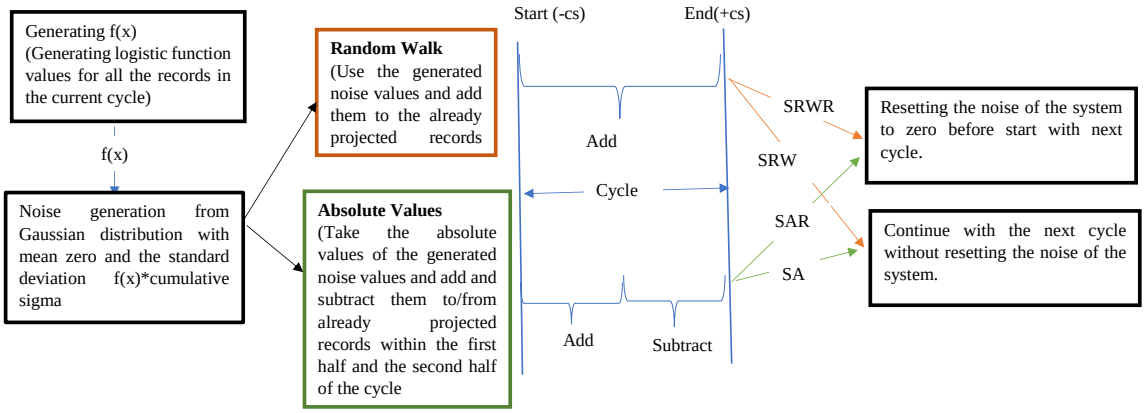

Figure 3: Process of Logistic Cumulative Noise Addition Methods

The following pseudo-code summarizes the process followed with logistic cumulative noise addition methods (See Algorithm 2).

3.3 Classification and Evaluation Process

The proposed methodology uses ARF (Gomes et al., 2017) as its learning algorithm. Since we use this for a data streaming/online learning environment, there are some specific requirements of data streams that need to be considered. Processing one example at a time for at most one time, uses

Algorithm 2 Pseudo Code - Logistic Cumulative Noise Addition Methods

Input: Cycle Size (cs), Cumulative Noise Sigma (σ), Projected Dataset (X), Maximum Value of Logistic Function (L), Growth Rate (k)

Process: **While** End of X

for each Record i in each cycle (defined by cs)

do

Calculate the value of Logistic Function $f(x) = \frac{L}{1+e^{-k(x-x_0)}}$

Generate noise (n) from Gaussian distribution $N(0, f(x) \times \sigma^2)$

Add n cumulatively to the records $(y_i = x_i + n_i)$

Use n_i **OR** absolute n_i

End Cycle

System Noise $= 0$ **OR** System Noise $=$ Current Cumulative Noise

Continue Next Cycle

Output: Perturbed Dataset (Y_i)

a limited amount of time and memory, and should be ready to predict at any time are the most significant requirements of a data stream (Bifet et al., 2012), (Gomes et al., 2017) and (Bifet, 2011). By considering all these requirements, training and testing of the model can be carried out in two possible ways namely holdout method and interleaved test-then-train or prequential method (Bifet et al., 2012), (Gomes et al., 2017) and (Bifet, 2011).

The holdout method measures the performance of a single holdout set and is most useful when the division between train and test sets has been pre-defined. On the other hand, in the prequential method, each example is used to test the model before it is used to train the model, and accuracy is incrementally updated. When this is performed in the correct order, the model is always being tested on the samples it has not seen. The prequential method does not need a holdout set or pre-defined training and testing sets, that take the maximum use of data available (Bifet et al., 2012). These properties make this method more suitable for a data streaming environment that evolves and learns incrementally, and we do not have a clear idea about the amount of data or the availability rate of the data. This arises the need for accuracy to be measured over time. After considering all these factors and the usability of these methods in a practical environment, we decided to use the prequential evaluation setting implemented for our model. Therefore, this method does not produce accuracy measures for training and testing separately.

3.4 Accuracy-Privacy Trade-off Optimization

Optimization is a serious matter that needs to be handled carefully, as it depends on what kind of problem you are working with. Achieving optimization in different scenarios has been discussed in the literature. For example, (Precup, Teban, Albu, Borlea, Zamfirache and Petriu, 2020) discuss optimizing fuzzy models for prosthetic hand myoelectric-Based control, (Roman, Precup, Bojan-Dragos and Szedlak-Stinean, 2019) have investigated on optimization problem in virtual reference feedback tuning for tower crane systems and (Yuhana, Fanani, Yuniarno, Rochimah, Kóczy and Purnomo, 2020) discuss on predicting the minimum passing level of competency achievement which

is also an optimization problem. Though there is a considerable amount of work done in PPDM no discussion can be found in terms of optimizing the accuracy-privacy trade-off.

The ideal optimization scenario for our study is that where we can achieve perfect values for both privacy and accuracy. In other words, zero classification error and zero breach probability(Liu, 2007) which is impossible to achieve. A possible way to make this work is, trying to minimize both classification error and breach probability. Therefore, we define the optimization problem using classification error and breach probability according to the Privacy Accuracy Magnitude(PAM) formula *(PAM = (error)^2 + P(€-privacy breach)^2)* proposed by (Denham et al., 2020).

We can say that if the PAM is less than a given error threshold (ϑ), the accuracy-privacy trade-off has been optimized.

$$\text{If } PAM < \vartheta; \textit{ Then the trade-off is optimized.}$$

Deciding a suitable value for ϑ is a complex and important issue. Especially, because the accuracy and privacy values depend on the characteristics of the data set. Moreover, the optimal level of privacy and accuracy highly depends on the user's requirements, and an improved framework should be needed to handle these scenarios. Therefore, at this stage of the research, we compare the PAM values for proposed PPDM methods to identify the method which produces the minimum PAM, and how each method performs in terms of privacy and accuracy.

By considering all these techniques together, our proposed methodology can be briefed as in Figure 4. Data enters into the system sequentially one sample at a time. These data go through the perturbation process to preserve privacy by hiding their true values. Inside the perturbation module, we experimented with eight different variations of the cumulative noise addition methods to find out the best method of them. Then the perturbed data is being classified using Adaptive Random Forest in a prequential evaluation setting and the accuracy of the classified data is updated for each record. Simultaneously, the privacy of the data is also measured by performing Known I/O attacks on the perturbed data. Finally, the overall error score is calculated using error and breach probability in terms of accuracy and privacy, respectively.

4 Experiments and Results

4.1 Data sets

Two freely available data sets which can be considered as data streams are being used for the experiments, and these can be considered as streams because the records were ordered according to the time they were produced. Only the numerical features of the data sets were considered for the experiments, and both data sets contain potentially sensitive data. Data is being pre-processed accordingly. Details of the experimented data sets are represented below.

- Activity Recognition system based on Multisensory data fusion (AReM data set from UCI Machine Learning Repository) – This contains real-world data which includes 35,999 records, 6 features, and 5 classes.
- New York City Taxi Trip Duration (Taxi data set from Kaggle, 2017) - This contains real-world data which includes 50,000 records, 7 features, and 3 classes.

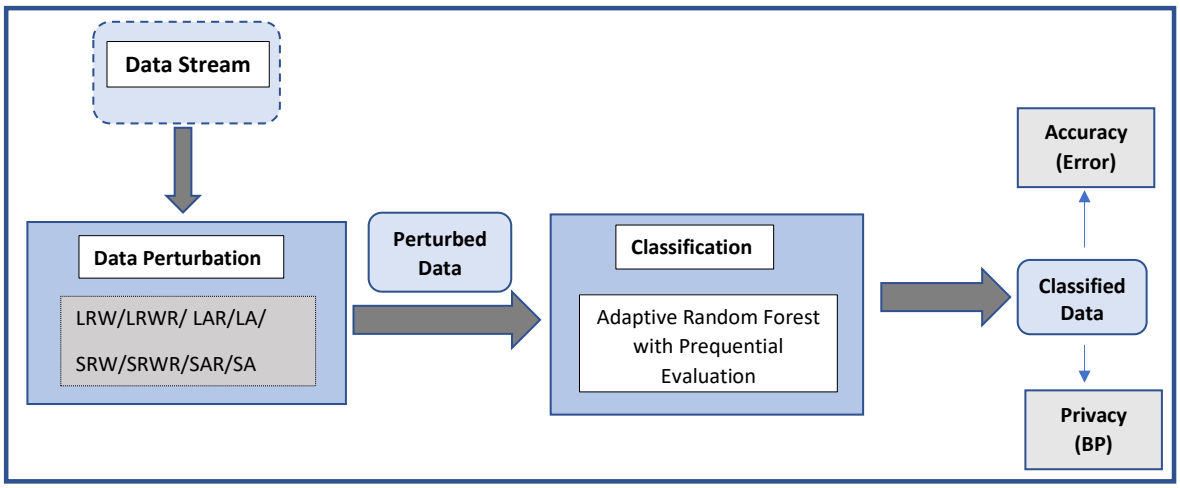

Figure 4: Proposed Methodology

4.2 Experimental Setup

Experiments are carried out to measure the privacy and accuracy of the above-mentioned data sets. Privacy and accuracy measures used in (Denham et al., 2020) are adopted. The following Configurations are used for the experiments.

The noise was generated from a Gaussian distribution with a mean zero and a variance of 3.90×10^{-6}. Noise variance for the cumulative noise addition was selected according to the method proposed in (Denham et al., 2020) which generates a similar amount of noise when adding independent noise with a variance value of 0.0625. Cycles mean the virtually broken-down sets of the entire data set, defined by cycle size. Experiments were carried out for the cycle sizes 300, 600, 1000 and 2000. Here is the configuration of other parameters used in the experiment.

General Configuration

- Classification Method – Adaptive Random Forest (ARF) with Naive Bayes leaf Prediction in a prequential evaluation setting
- Attacking Method – Known Input/ output Attack
- Cycle Sizes – 300, 600, 1000, 2000
- Variance of Cumulative Noise – 3.90 x 10^{-6}
- Number of Known Input/ output pairs – 4 per attack
- Number of Attacks – 5% of total record count
- € - epsilon – 0.2

Configurations of logistic function

- Maximum Value – 1
- Growth Rate – 0.01, 0.02

4.3 Measuring Privacy and Accuracy

In (Denham et al., 2020) for the privacy perspective known input/output MAP attacks were conducted and the €- privacy breach probability was measured while for the accuracy perspective relative error was calculated. The same methods have been used in our research work also. In (Denham et al., 2020) authors have extended the known input/output attacking method to random projection-based cumulative noise addition environment which was originally proposed by (Giannella et al., 2008) for multiplicative data perturbation. To measure the accuracy of the data stream, Adaptive Random Forest (ARF) proposed by (Gomes et al., 2017) is used as the classifier and relative error represents the degree of success achieved by a record recovery attempt is measures. It is defined as the magnitude of the difference vector between the original record and its recovered counterpart, normalized by the magnitude of the original record vector (Denham et al., 2020). An "€-privacy breach" of a record occurs if the relative error of the recovered record is less than a specified threshold (€; € $>$ 0) (Liu, 2007).

Overall error score/performance was evaluated according to the Privacy-Accuracy Magnitude (PAM) proposed in (Denham et al., 2020) and we modified the process by using normalized values of relative error and breach probability to maintain fairness.

4.4 Results

This section explains the results of the experiments carried out, including overall performance, relative error, and breach probability behavior with different cycle sizes. Results of the cycle size 300 and 2000 of AReM data set are represented here and a similar trend of results was displayed by the other cycle sizes also. The code for the experiments can be found in https://github.com/whewage/Variations-of-Cumulative-Noise-Addition.git

Table 2: Overall Performance using PAM (AReM data set, Cycle size 300)[1]

Method	Relative Error		Breach Probability		Overall Error Score	
	k = 0.01	k=0.02	k = 0.01	k=0.02	k = 0.01	k=0.02
LAR	0.6174	0.5502	0.1111	0.2857	0.3936	0.1349
LA	1.0000	0.9110	0.4444	0.7143	1.1975	0.1600
LRWR	0.3462	0.2945	0.4444	0.7143	0.3174	0.1199
LRW	0.6053	0.5388	0.0000	0.1429	0.3664	0.1339
SAR	0.2833	0.3482	1.0000	0.5714	1.0803	0.1228
SA	0.9540	1.0000	0.2222	0.0000	0.9595	0.1649
SRWR	0.0000	0.0000	0.8889	1.0000	0.7901	0.1038
SRW	0.2845	0.2877	0.4444	0.4286	0.2785	0.1188

[1]Normalized values of relative error and breach probability were used for the convenience of understanding.

Table 3: Overall Performance using PAM (AReM data set, Cycle size 2000)[1]

Method	Relative Error		Breach Probability		Overall Error Score	
	k = 0.01	k=0.02	k = 0.01	k=0.02	k = 0.01	k=0.02
LAR	0.8056	0.8127	0.1111	0.1111	0.6614	0.6728
LA	1.0000	1.0000	0.2222	0.2222	1.0494	1.0494
LRWR	0.3136	0.3385	1.0000	1.0000	1.0984	1.1146
LRW	0.2479	0.2751	0.4444	0.4444	0.2590	0.2732
SAR	0.3568	0.4606	0.8889	0.7778	0.9174	0.8171
SA	0.7606	0.7167	0.0000	0.0000	0.5785	0.5137
SRWR	0.0000	0.0000	0.5556	0.6667	0.3086	0.4444
SRW	0.0329	0.0914	0.1111	0.4444	0.0134	0.2059

According to the results of Tables 1 and 2, we can observe that for k=0.01, the lowest error score (the highest performance) was achieved by SRW for both cycle sizes. For k=0.02, with cycle sizes of 300 and 2000, the lowest error scores were given by SRWR and SRW, respectively. But for a cycle size of 300, we can see that SRW returns the second-lowest error score, which is quite close to the error score of the SRWR. Initially, it was assumed that the noise resetting at the end of each cycle would make a considerable change in the performance, but from the results, we can see that it only makes a marginal improvement in accuracy ($1 - relativeerror$) but does not have a significant impact on the overall error score. Therefore, the perturbation method SRW (Logistic random walk without noise resetting) was selected as the best performer as it gives the minimum overall error. When we compare the results with the state-of-the-art work (LRW), SRW has outperformed in both cycle sizes considering both growth rate values.

Analyzing the behavior of accuracy and privacy with different cycle sizes and growth rates is important to understand the effects of those two parameters, and also in selecting optimal values for these parameters. Relative error and the breach probability of the "AReM" data set were analyzed with four different cycle sizes ($cs = 300, 600, 1000, 2000$) and two different growth rates ($k = 0.01, 0.02$). Note that the growth rate only affects the logistic noise addition methods (highlighted in both Table 3 and 4).

According to the results in Table 3, we can see that relative error increases with the cycle size in both linear and logistic cumulative noise additions. A possible reason for this behavior can be that the noise contained in the system is high when it comes to the higher cycle sizes. On the other hand, increasing the growth rate negatively affects the accuracy, since the relative error increases with the growth rate. In such cases, the classifier is not able to adapt its model fast enough to cope with the rate of increase of noise, which in turn leads to a greater error. But as you can see in the table, SRWR and SRW have low relative errors in all the cases compared to the state-of-the-art work (LRW) and LRWR and SAR have outperformed LRW in low cycle sizes.

According to the results represented in Table 4. in most of the cases breach probability decreases

Table 4: Behavior of relative error with different cycle sizes and growth rate values.

Method	Cycle Size							
	300		600		1000		2000	
	k=0.01	k=0.02	k=0.01	k=0.02	k=0.01	k=0.02	k=0.01	k=0.02
LAR	0.3664		0.3907		0.404		0.4248	
LA	0.3980		0.4040		0.4224		0.4455	
LRWR	0.3440		0.3545		0.3657		0.3724	
LRW	0.3654		0.3654		0.3654		0.3654	
SAR	0.3388	0.3487	0.3596	0.3607	0.3708	0.3671	0.3770	0.3859
SA	0.3942	0.4058	0.3920	0.3973	0.4203	0.4245	0.4200	0.4142
SRWR	0.3154	0.3182	0.3308	0.3340	0.3293	0.3345	0.3390	0.3350
SRW	0.3389	0.3434	0.3403	0.3466	0.3418	0.3443	0.3425	0.3451

Table 5: Behavior of breach probability with different cycle sizes and growth rate values

Method	Cycle Size							
	300		600		1000		2000	
	k=0.01	k=0.02	k=0.01	k=0.02	k=0.01	k=0.02	k=0.01	k=0.02
LAR	0.025		0.025		0.005		0.005	
LA	0.040		0.033		0.015		0.010	
LRWR	0.040		0.017		0.025		0.045	
LRW	0.020		0.020		0.020		0.020	
SAR	0.065	0.035	0.017	0.033	0.030	0.020	0.040	0.035
SA	0.030	0.015	0.033	0.008	0.010	0.015	0.000	0.000
SRWR	0.060	0.050	0.025	0.042	0.055	0.045	0.025	0.030
SRW	0.040	0.03	0.017	0.058	0.015	0.045	0.005	0.020

with the cycle size and hence the privacy increases. Unlike the classification error, we cannot see a clear movement of the behavior of the breach probability with the growth rate. But we can see that SA and SRW (logistic cumulative noise additions without noise resetting) show an approximately similar behavior and on the other hand, SAR and SRWR show a similar trend with different k values. But it is not sufficient to explain the behavior of privacy with different growth rates. When we compare results with LRW a clear pattern cannot be seen but SRW has produced equal/low breach probability values for $k = 0.01$ in $cyclesize = 600$ and 1000 and both k values in $cyclesize = 2000$.

Moreover, we have conducted the attacks to the starting and ending flat areas of the logistic cycle instead of performing random attacks to the data stream. Starting flat area of the logistic cycle is the most vulnerable location because of two main reasons. The first reason is the noise level added is very low in that area, and the second one is the noise addition rate is very low. On the other hand, when we consider the ending flat area it is also vulnerable because noise is added at a constant noise addition rate which makes it easier for the attackers to breach the privacy regardless of the high noise level of that area. Therefore, starting and ending flat areas have been identified as the most vulnerable to attacks and to prove our claim we have conducted the attacks on those identified areas of the best performing perturbation method SRW.

Attacks were performed in two different manners namely, attacks to the flat areas of randomly selected cycles and attacks to the flat areas of each cycle and compared the results with the breach probability of performing random attacks to the data stream. Table 5 displays the average breach probabilities of AReM and Taxi data sets respectively after conducting 50 rounds of attacks. We conducted 50 rounds of attacks to reduce the possible bias when conducting a single round of attacks.

Table 6: Comparison of breach probabilities after performing attacks to the different locations of the data stream

Data sets	**Attacks-starting flat period**		**Attacks-ending flat period**		**Random attacks**
	Random cycles	Each cycle	Random cycles	Each cycle	
AReM	0.0310	0.0312	0.0296	0.0319	0.0270
Taxi	0.0273	0.0341	0.0250	0.0366	0.0200

The results prove that the selected perturbation method SRW is relatively more vulnerable when the attacks are performed to the flat areas of the logistic cycle. Both data sets have shown that it is easy to breach privacy when attacking the starting flat period of randomly selected cycles with breach probabilities of 0.002 and 0.0023 than the breach probabilities of ending flat areas of AReM and Taxi data sets, respectively. When the attacks were performed to the flat areas of each cycle, the starting flat area seems to be less/equally vulnerable with the ending flat area. However, even with those changes, SRW has been succeeded in maintaining more than 97% of privacy in all the scenarios, which proves the reliability of the method.

5 Discussion

By analyzing the results of the experiments, the perturbation method SRW can be identified as the best performer considering both privacy and accuracy perspectives. In cooperating logistic function with the cumulative noise addition certainly shows a positive impact towards optimizing the trade-off between privacy and accuracy. The nature of the logistic function helps to control the maximum noise level of the cumulative noise addition, which avoids noise from dominating the data. This

leads to increasing the accuracy level, hence provides an opportunity to minimize the trade-off.

Noise addition in cycles seems effective since relative error decreases, hence accuracy increases in smaller cycle size (cs). On the other hand, beach probability decreases hence privacy increases when cycle size increases. The growth rate(k) of the logistic function also makes an impact, since relative error increases with it. But the impact of growth rate on breach probability should be investigated further, since results do not show a clear pattern. However, it is vital to select the appropriate cycle size and the growth rate values. Noise resetting does not give the expected results as it only does a marginal improvement of the accuracy but does not have any significant change to the overall score. Using absolute noise values is not a good technique to control the noise level because noise injection distorts the data irrespective of addition or subtraction.

Flat areas of the logistic cycle are the most vulnerable to the attacks, and if the attacker can find out the cycle size, he can succeed in attacking those most vulnerable areas. Therefore, cycle size is an important parameter that needs to be protected.

6 Conclusion and Future Work

In summary, we experimented with different techniques which can be combined with cumulative noise addition to controlling the maximum noise added to the data stream. Controlling the maximum noise level is essential in the cumulative noise addition environment, as the system has to face a huge amount of noise with the time when the data stream unfolds, which can highly decrease the classification accuracy. By doing so, our objective was to maintain the maximum data privacy benefits which receives from cumulative noise addition while doing a minimum negative impact on the classification accuracy that leads to optimizing the trade-off between privacy and accuracy. According to our experiments, cumulative noise addition in cycles combined with the concept of logistic function turned to be a promising method to control the maximum noise level of the stream. Therefore, cumulative noise addition combined with logistic function has been proved a better approach to optimize the trade-off between privacy and accuracy by controlling the maximum noise level of the system.

As for future work, the use of noise resetting should be investigated further conceptually and experimentally with different data sets, as it appears to be a promising method to control the noise level of the system. Moreover, instead of using a fixed cycle size noise can be added in randomly selected cycle sizes to ensure high privacy by avoiding attacks to the flat areas and this can be considered as an important improvement to this work. In addition to that, a well-formulated privacy-accuracy framework using the selected method can be the next step of the work. This framework should be able to answer the optimization of accuracy privacy trade-off issue according to the user's accuracy and privacy requirements.

Acknowledgement

We greatly acknowledge the authors of the state of the artwork (Denham et al., 2020), Ben Denham, Russel Pears and M. Asif Naeem for all the support given to improve this work.

References

Aggarwal, C. C. and Yu, P. S. 2004. A Condensation Approach to Privacy Preserving Data Mining, *Advances in Database Technology - EDBT 2004*, Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 183–199.

Agrawal, R. and Srikant, R. 2000. Privacy-preserving data mining, *SIGMOD Rec.* **29**(2): 439–450.

Ahmed, M. U., Brickman, S., Dengg, A., Fasth, N., Mihajlovic, M. and Norman, J. 2019. A machine learning approach to classify pedestrians' events based on imu and gps, *International Journal of Artificial Intelligence* **17**(2): 154–167.
URL: *http://www.ceser.in/ceserp/index.php/ijai/article/view/6260/6207*

Bifet, A. 2011. MOA Data Stream Mining: A practical approach, *Methodology* **8**(May): 127–141.
URL: *http://dspace.cusat.ac.in/jspui/handle/123456789/3616*

Bifet, A., Kirkby, R., Kranen, P. and Reutemann, P. 2012. Massive Online Analysis (MOA) Manual, (March).
URL: *http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Massive+Online+Analysis+Manual#*

Chen, K. and Liu, L. 2005a. A random rotation perturbation approach to privacy preserving data classification, *International Conference on Data Mining* .

Chen, K. and Liu, L. 2005b. Privacy preserving data classification with rotation perturbation, *Proceedings - IEEE International Conference on Data Mining, ICDM* pp. 589–592.

Chen, K., Sun, G. and Liu, L. 2007. Towards Attack-Resilient Geometric Data Perturbation, *SIAM International Conference on Data Mining*, pp. 78–89.

Denham, B., Pears, R. and Naeem, M. A. 2020. Enhancing random projection with independent and cumulative additive noise for privacy-preserving data stream mining, *Expert Systems with Applications* **152**.

Giannella, C., Kargupta, H. and Liu, K. 2008. A Survey of Attack Techniques on Privacy-Preserving Data Perturbation Methods Chapter.

Giannella, C. R., Liu, K. and Kargupta, H. 2013. Breaching Euclidean distance-preserving data perturbation using few known inputs, *Data and Knowledge Engineering* **83**: 93–110.
URL: *http://dx.doi.org/10.1016/j.datak.2012.10.004*

Gomes, H. M. et al. 2017. Adaptive random forests for evolving data stream classification, *Machine Learning* **106**(9-10): 1469–1495.

Guo, S. and Wu, X. 2006. Deriving private information from general linear transformation perturbed data, *Proceedings of the 2006 SIAM International Conference on Data Mining* pp. 59–69.

Guo, S., Wu, X. and Li, Y. 2006. On the lower bound of reconstruction error for spectral filtering based privacy preserving data mining, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **4213 LNAI**: 520–527.

Huang, Z., Du, W. and Chen, B. 2005. Deriving private information from randomized data, *Proceedings of the ACM SIGMOD International Conference on Management of Data* pp. 37–48.

Kargupta, H., Datta, S., Wang, Q. and Sivakumar, K. 2004. On the privacy preserving properties of random data perturbation techniques, pp. 99–106.

Liu, K. 2007. *Multiplicative Data Perturbation for Privacy Preserving Data Mining*, Phd thesis, University of Maryland, Baltimore County (UMBC).

Liu, K., Giannella, C. and Kargupta, H. 2006. An Attacker's View of Distance Preserving Maps for Privacy Preserving Data Mining, *In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **Vol. 4213**: 297–308.
URL: *https://doi.org/10.1007/11871637_30*

Liu, K., Kargupta, H. and Ryan, J. 2006. Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 18, pp. 92–106.

Liu, L., Wang, J. and Zhang, J. 2009. Privacy vulnerabilities with background information in data perturbation, *Society for Industrial and Applied Mathematics - 9th SIAM International Conference on Data Mining 2009, Proceedings in Applied Mathematics*, Vol. 3, Te?hni?al Report CMIDA-HiPSCCS 005-08, Department of Co mputer S?ien?e, University of Ken- tu?ky, KY, pp. 1268–1277.

Okkalioglu, B. D., Okkalioglu, M., Koc, M. and Polat, H. 2015. A survey: deriving private information from perturbed data, *Artificial Intelligence Review* **44**(4): 547–569.

Pozna, C. and Precup, R. E. 2014. Applications of signatures to expert systems modelling, *Acta Polytechnica Hungarica* **11**(2): 21–39.

Precup, R. E., Teban, T. A., Albu, A., Borlea, A. B., Zamfirache, I. A. and Petriu, E. M. 2020. Evolving Fuzzy Models for Prosthetic Hand Myoelectric-Based Control, *IEEE Transactions on Instrumentation and Measurement* **69**(7): 4625–4636.

Raziyeh Zall, M. R. K. 2019. On the Construction of Multi-Relational Classifier Based on Canonical Correlation Analysis, *International Journal of Artificial Intelligence* **17**.
URL: *http://www.ceserp.com/cp-jour/index.php/ijai/article/view/5274*

Roman, R. C., Precup, R. E., Bojan-Dragos, C. A. and Szedlak-Stinean, A. I. 2019. Combined Model-Free Adaptive Control with Fuzzy Component by Virtual Reference Feedback Tuning for Tower Crane Systems, *Procedia Computer Science* **162**(Itqm 2019): 267–274.
URL: *https://doi.org/10.1016/j.procs.2019.11.284*

Sang, Y., Shen, H. and Tian, H. 2012. Effective reconstruction of data perturbed by random projections, *IEEE Transactions on Computers* **61**(1): 101–117.

Verhulst, P. F. 1838. Logistic Function.
URL: *https://en.wikipedia.org/wiki/Logistic_function*

Yuhana, U. L., Fanani, N., Yuniarno, E. M., Rochimah, S., Kóczy, L. and Purnomo, M. H. 2020. Combining fuzzy signature and rough sets approach for predicting the minimum passing level of competency achievement, *International journal of artificial intelligence* **18**: 237–249.