EUSRM
A Universe of Research
ISSN(ONLINE) : 2319-3069

Engineering Universe for Scientific Research and Management
(International Journal)
Impact Factor: 3.7

Vol. 6 Issue 4
April 2014

Geometric Data Perturbation Approach for Privacy Preserving in Data Stream Mining

Twinkle Ankleshwaria[1], Prof .J.S.Dhobi[2]
Department of Computer Science & Engineering , Gujarat Technological University, Gujarat, India[1,2]
twi1902@gmail.com[1], jsdhobi@yahoo.com[2]

Abstract

Today as we are living in the era of information explosion, it has become very important to find out useful information from large massive data. Also advances in internet, communication and hardware technology has lead to an increase in the capability of storing personal data of individuals. Massive amount of data streams are generated from different applications like medical, shopping record, network traffic, etc. Sharing such data is very important asset to business decision making but the fear is that once the personal data is leaked it can be misused for a variety of purposes. Hence some amount of privacy preserving needs to be done on the data before it is released to others. Traditional methods of privacy preserving data mining (PPDM) are designed for static data sets which makes it unsuitable for dynamic data streams. In this paper an efficient and effective data perturbation method is proposed that aims to protect privacy of sensitive attribute and obtaining data clustering with minimum information loss.

Keywords: *Data Perturbation, Data Stream, Clustering, Privacy Preserving, Geometric Data Perturbation.*

1. Introduction

Data streams can be conceived as a continuous and changing sequence of data that continuously arrive at a system to store or process [2].Tremendous and potentially infinite volumes of data streams are often generated by real – time surveillance systems, sensor networks, stock market, on-line transaction in financial market or retail industry, scientific and engineering experiments, power supply, meteorological research, internet traffic, telecommunication, power supply, etc. These data sets need to be analyzed for identifying patterns which can be used to predict future behavior. However, data owners may not be willing to share the real values of their data due to privacy reason. Hence, some amount of privacy preservation needs to be done on data before it is released.

Privacy preserving data mining (PPDM) tends to transform original data, so that sensitive data are preserved. Verykios et.al [9] classified privacy-preserving data mining techniques based on five dimensions which are, data distribution, data modification, data mining algorithm, data or rule hiding , and privacy preserving.

The first dimension is related to distribution of data: Centralized or Distributed. Further Distributed can be divided into horizontally or vertically distributed, where horizontally means different records reside in different places and vertically means all values of different attributes reside in different places. The second dimension is related to modification of original values of data that are to be released for data mining task. Different methods in which modifications can be carried out are perturbation, blocking, aggregation, swapping, merging, sampling, etc. The third dimension is related to data mining algorithm like classification, clustering, association rule mining etc .The fourth dimension is related to whether the raw data or aggregated data should be hidden. Finally the fifth dimension is related to the techniques that are used for protecting privacy. From the review of previous research it can be seen that existing techniques for privacy preserving data mining are designed for static databases, which makes it unsuitable for dynamic data streams [1].

Based on these dimensions, different PPDM techniques may be classified into following five categories.

Anonymization based PPDM, Perturbation based PPDM,

Randomized Response based PPDM, Condensation approach based PPDM, and Cryptography based PPDM.

Detailed description of the above PPDM categories is described in [3]. Motivated by PPDM, a new research area called Privacy Preserving in Data Stream Mining has been emerged. The initial idea of it was to extend traditional data mining techniques to work with the stream data.

In this paper we have proposed a new method which uses geometric data perturbation approach for privacy preserving in data stream mining. The remaining section of this paper is organized as follows. In section 2, we discuss different related work in this area. In section 3, we describe the proposed privacy preserving method using geometric data perturbation. Experimental strategy set up, results and result analysis is described in section 4. Finally, we conclude in section 5.

2. Related Work

[3] Privacy and accuracy in case of data mining is a pair of contradiction .Achieving one can lead to adverse effect on other. In this, an effort to review a good number of existing PPDM techniques is reviewed. Finally it concludes, there does not exists a single privacy preserving data mining algorithm that outperforms all other algorithms on all possible criteria like performance ,utility ,cost , complexity, tolerance against data mining algorithms etc. Various evaluation parameters for PPDM algorithms are discussed.

[1]In this various challenges & Issues related to PPDM are discussed. Existing algorithms offering privacy need further investigation for possible improvements. Common framework for different PPDM is still an issue that will unify more advanced measures for the evaluation. Privacy preserving for fuzzy sets and Data stream mining are most recent and prominent directions of PPDM.

[8]In this data modification –based framework is presented .At first, these techniques classified into two classes of anonymization and perturbation approaches and after analyzing each approach, their significant characteristics are given. The main challenge of anonymization approach was insufficient protection of critical values against different attacks, whereas issue in perturbation approach is of creating suitable & stable balance between privacy and data mining

[10] In this a method of Privacy Preserving Clustering of Data Streams (PPCDS) is proposed stressing the privacy –preserving process in a data stream environment while maintaining a certain degree of excellent mining accuracy. PPCDS is mainly used to combine Rotation –Based Perturbation , optimization of cluster enters and the concept of nearest neighbour, in order to solve the privacy –preserving clustering of mining issues in a data stream environment .In the phase of Rotation –Based Perturbation , rotation transformation matrix is employed to rapidly perturb with data streams in order to preserve data privacy. In the phase of cluster mining, perturbed data is primarily used to establish a micro-cluster through the optimization of a cluster centre, then applying statistic calculation to update the micro-cluster

In [11] the random response technology and the geometric data transformation method, That is called random response method of geometric transformation .It can protect the privacy of numerical data .Theoretical analysis and experimental results showed that at the same time cost, the algorithm can get better privacy protection than the previous algorithms, and would not affect the accuracy of mining results.

In [12] a proposed approach based on geometric data perturbation and data mining –service oriented framework is introduced.GDP had shown to be an effective perturbation method in single –party privacy preserving data publishing. The multiparty framework and the problem of perturbation unification under this framework is presented. Three protocols are proposed which is first effort on applying GDP to multiparty privacy –preserving mining.

In [13] it is shown how several types of well –known data mining models will deliver a comparable level of model quality over geometrically perturbed data set as over the original data set.GDP ,includes the linear combination of three components: rotation perturbation ,translation perturbation, and distance perturbation .GDP perturbs multiple column in one transformation. A multi-column privacy evaluation model is proposed and analysis against three types of inference attacks: naïve-inference, ICA –based and distance –inference is done.

In [14] tuple value based multiplicative data perturbation approach is proposed in which author has tried to keep statistical relationship among the tuple attributes intact. It considers sensitive attribute as dependent attribute and others as independent .Independent attributes of tuple has been used to calculate tuple specific random noise .K-means clustering algorithm over defined sliding window size on perturbed data stream has been used in order to estimate the accuracy and effectiveness of clustering results over two standard datasets. Result is evaluated against various measures like Precision, Recall, and CMM.

In [15] proposed data perturbation method for privacy preserving classification of data streams, which consists of two steps: data streams pre-processing and data stream s mining. In first step algorithm for perturbation is proposed and in second step Hoeffding tree algorithm is applied on perturbed dataset. Experiments are conducted and classification model is generated which shows minimal information loss from original dataset.

We discussed wide area for privacy-preserving data mining techniques. Privacy Preserving in Data stream mining is recently emerged research field in response to the issues and challenges related with continuous data stream.

3. Proposed Method.

3.1 Problem Description.

The main objective of this proposed method is to provide privacy before release of data. Perturbation method can be used for privacy preserving in data stream mining.Geometric Data perturbations method is totally on

distance base for estimating original value from the perturbed data, with addition of Gaussian noise. Geometric perturbation is an enhancement to rotation perturbation by incorporating additional components such as random translation perturbation and noise addition to the basic form of multiplicative perturbation Y = R £ X.
The goal is to transform a given data set into perturbed dataset that satisfies a given privacy requirement with minimum information loss for the intended data analysis task. Two step process: data stream preprocessing, data stream cluster mining. In the first step the objective is to perturb data stream to preserve data privacy. In the second step the objective is to mine perturbed data stream to cluster the data using sliding window mechanism.

3.2 Framework

Figure 1, shows all the theoretical aspects are represented graphically. Data sets can be generated using stream generators or can be used from any data repository like UCI which is available. Upper part shows stream clustering on true dataset $\mathcal{D}$, where as bottom half shows the extended framework in order to implement our proposed method to perturb true dataset and applying same clustering operation on modified data stream $\mathcal{D}'$. Both datasets are provided to standard clustering stream learning algorithm in Massive Online Analysis (MOA)[16] software to obtain result $\mathcal{R}$ and $\mathcal{R}'$ respectively. The main focus of proposed work is to obtain close approximation between clustering results $\mathcal{R}$ and $\mathcal{R}'$ to balance tradeoff between privacy gain and information loss.

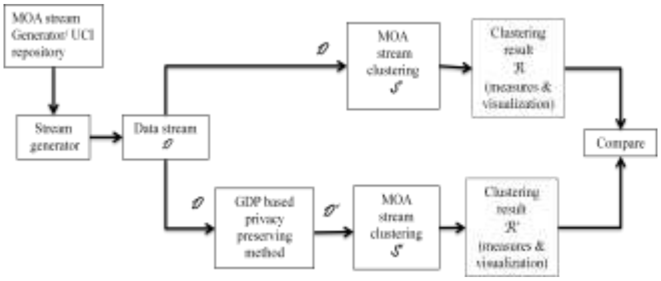

Fig. 1 Framework for Privacy Preserving in Data Stream Clustering.

3.3 Algorithm

To protect the sensitive attribute, geometric data perturbation method is used. We use sliding window over the incoming data stream. The algorithm is repeated over each window. After rotation, translation Guassian noise is added to the sensitive attribute to generate perturbed dataset. To calculate noise, all attributes except class attribute are considered. Guassian noise is uncorrelated so, it is distributed and provide good privacy using Guassian distribution function .Finally both datasets , original and perturbed is processed using a predefined clustering algorithm.

Algorithm: Privacy Preserving using Geometric Data Perturbation.
Input: Data Stream **D,** Sensitive attribute **S**.
Intermediate Result: Perturbed data stream **D ’**.
Output: Clustering results **R** and **R’** of Data stream **D** and **D’** respectively**.**
Steps:

1. **Given** input data **D** with tuple size **n**, extract sensitive attribute $[S]_{n\times1}$.
2. Rotate $[\mathbf{S}]_{\mathbf{n\times1}}$ into 180o clock-wise direction and generate$[\mathbf{R_S}]_{\mathbf{n\times1}}$.
3. Multiply elements of $[\mathbf{S}]$ with $[\mathbf{R_S}]$, transformed sensitive attribute values will be
   $[\mathbf{X}]_{\mathbf{n\times1}}= [\mathbf{S}]_{\mathbf{n\times1}} \times [\mathbf{R_S}]_{\mathbf{n\times1}}$
4. Calculate translation T as mean of sensitive attribute $[\mathbf{S}]_{\mathbf{n\times1}}$.
5. Generate transformation $[\mathbf{St}]_{\mathbf{n\times1}}$ by applying translation **T** to $[\mathbf{S}]_{\mathbf{n\times1}}$.
6. Calculate Gaussian distribution **P(x)** as a probability density function for **Gaussian** noise $\mathbf{P(x)} = \frac{1}{\sigma\sqrt{2\pi}}\mathbf{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$
   where, μ=Mean, σ=Standard deviation**.**
7. Geometric data perturbation of sensitive attribute $[\mathbf{Gs}]_{\mathbf{n\times1}} = [\mathbf{X}]_{\mathbf{n\times1}} + [\mathbf{St}]_{\mathbf{nx1}} + \mathbf{P(x)}$.
8. Create perturbed dataset **D’** by replacing sensitive attribute $[\mathbf{S}]_{\mathbf{n\times1}}$ in original dataset **D** with $[\mathbf{Gs}]_{\mathbf{n\times1}}$.
9. Apply Any clustering algorithm with different values of k on original dataset **D** having sensitive attribute **S.**
10. Apply Any clustering algorithm with different values of k on perturbed dataset **D’** having perturbed sensitive attribute **Gs.**
11. Create cluster membership matrix of results from step 9 and step 10 and analyze.

Fig. 2 .Privacy Preserving Algorithm.

4. Experiments & Results

4.1 Experimental setup

To test framework as shown in Fig.1 Massive Online analysis (MOA) has been used. MOA framework is an open source framework for implementing algorithms and running experiments for online learning from evolving data streams [17]. It contains collection of offline and online algorithm for both classification and clustering as well as tools for evaluation [18]. In addition to this it supports bi-directional with WEKA machine learning algorithms.

To evaluate the effectiveness of proposed privacy preserving method, Experiments have been carried out on Intel Core I3 processor with 2GB memory on Windows 7.

The proposed technique is implemented in Java. Simulation has been done in data stream clustering environment. The experiments were processed on two different datasets available from the UCI Machine Learning Repository [19]. The brief information of chosen datasets is described in Table 1.

Table 1: Dataset information

Dataset	Total instances	Instances processed	Nominal attribute	Attributes protected
Bank Management	45211	45k	Ignored	Income, Duration
Letter Recognition	15327	15k	Ignored	Lno

K-mean Clustering algorithm using WEKA data mining tool in MOA framework has been simulated to evaluate the accuracy of proposed privacy preserving approach.

4.2 Results

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.

4.2.1 Cluster Membership Matrix (CMM)

Cluster Membership Matrix identifies how closely each cluster in the perturbed dataset matches its corresponding cluster in the original dataset as shown in Table 2. Rows represent the clusters in the original dataset, while columns represent the clusters in the perturbed dataset, and Freq $_{i,j}$ is the number of points in cluster C_i that falls in cluster C_i’ in the perturbed dataset.

Table 2: Cluster Membership Matrix

	C_1’	C_2’	……	C_n’
C_1	Freq $_{1,1}$	Freq $_{1,2}$	……	Freq $_{1,n}$
C_2	Freq $_{2,1}$	Freq $_{2,2}$	……	Freq $_{2,n}$
⋮	⋮	⋮	…	⋮
C_n	Freq $_{n,1}$	Freq $_{n,2}$	……	Freq $_{n,n}$

Using CMM accuracy can be obtained. Table 3, shows the percentage of accuracy obtained when selected attributes are perturbed using our algorithm in each datasets.

Table 3: Accuracy Obtained

Dataset	Attributes Perturbed	% Accuracy
Bank Management	Income Duration	97.64 75.05
Letter Recognition	Lno.	91.73

4.2.2 Precision & Recall Measures

Precision and Recall are two important measures to determine the effectiveness and accuracy of any information retrieval system. These measures are provided with MOA framework [17] where, F1_P determine the precision of system and F1_R determine the recall of system. Accuracy using these two measures is represented using line graph from Fig.3 to Fig.8. Each graph contains the measure we obtained when original data is processed without applying privacy preserving method and when data is undergone through our proposed privacy preserving method. *K-Means* is applied in order to evaluate both cases by keeping number of clusters fix (K=5 and K=3). Instances are processed in defined sliding window size.

F1-P, F1-R measures are the average measures based on precision and recall of individual clusters.

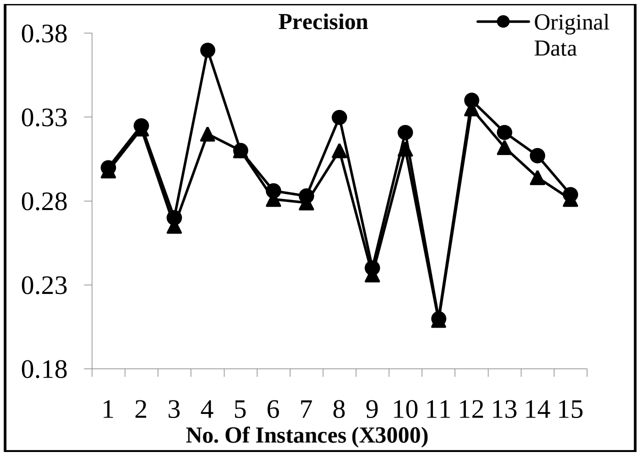

Fig.3. Precision Accuracy on attribute Income in Bank Management dataset (w=3000 and k=3)

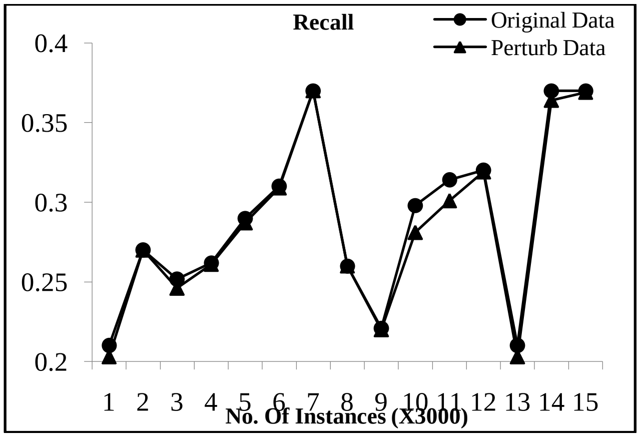

Fig.4. Recall Accuracy on attribute Income in Bank Management dataset (w=3000 and k=3)

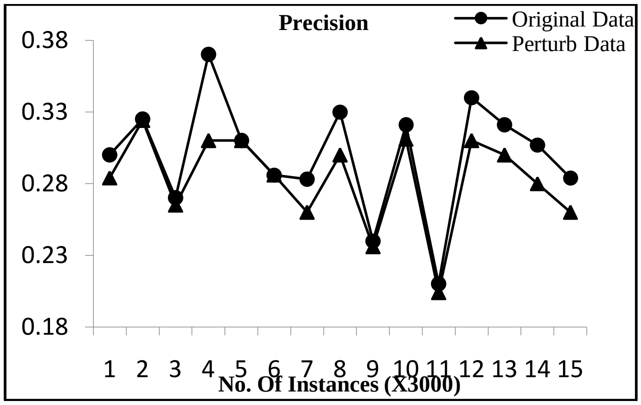

Fig.5. Precision Accuracy on attribute Duration in Bank Management dataset (w=3000 and k=3)

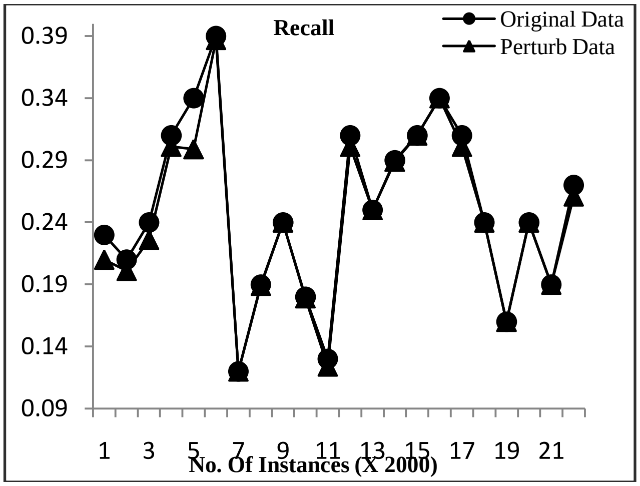

Fig.6. Recall Accuracy on attribute Duration in Bank Management dataset (w=3000 and k=3)

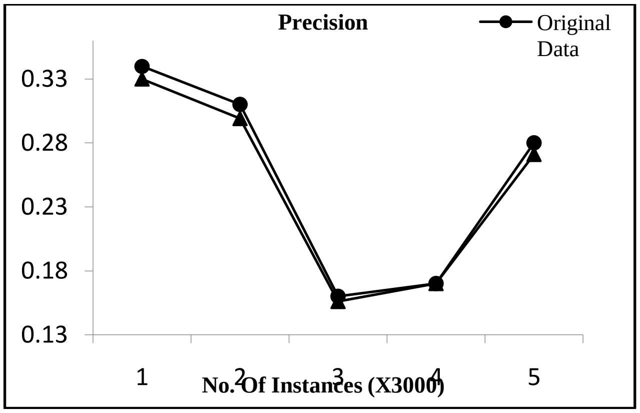

Fig.7. Precision Accuracy on attribute Lno in Letter Recognition dataset (w=3000 and k=3)

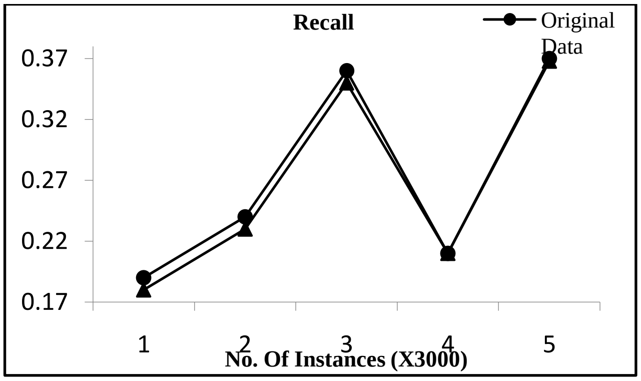

Fig.8. Recall Accuracy on attribute Lno in Letter Recognition dataset (w=3000 and k=3)

5. Conclusion

An ultimate goal for all data perturbation algorithm is to optimize the data transformation process by maximizing both data privacy and data utility achieved. Proposed approach focused on data perturbation by geometric transformation and noise addition to preserve privacy of sensitive attributes. We extended existing MOA framework in which, each tuple of data stream is independently treated. We considered single attribute as sensitive attribute (dependent attribute) and rest are as non confidential attributes (independent attributes) ignoring class attribute .We evaluate the experiment result in terms of correctly classified instance, Misclassification error. Results show fairly good level of privacy has been achieved with reasonable accuracy in almost all tested cases against evaluation measures- Precision , Recall and Cluster Membership Matrix..We limited experiments to protect numeric attributes only but work can be extended to nominal type attribute also. Experimental results and result analysis showed that proposed method can not only preserve data privacy but also can mine data streams accurately.

References

[1] Chhinkaniwala H. and Garg S., "Privacy Preserving Data Mining Techniques: Challenges and Issues", CSIT, 2011.

[2] L.Golab and M.T.Ozsu, Data Stream Management issues-"A Survey Technical Report", 2003.

[3] Majid, M.Asger, Rashid Ali, "Privacy preserving Data Mining Techniques: Current Scenario and Future Prospects", IEEE 2012.

[4]. Aggrawal, C.C, and Yu.PS.," A condensation approach to privacy preserving data mining". Proc. Of In .conf. on extending Database Technology (EDBT)(2004).

[5] Chen K, and Liu, "Privacy Preserving Data Classification with Rotation Perturbation", proc.ICDM, 2005, pp.589-592.

[6] K.Liu, H Kargupta, and J.Ryan," Random projection –based multiplicative data perturbation for privacy preserving distributed data mining." IEEE Transaction on knowledge and Data Engg,Jan 2006, pp 92-106.

[7] Keke Chen, Gordon Sun, and Ling Liu. Towards attack-resilient geometric data perturbation." In proceedings of the 2007 SIAM international conference on Data mining, April 2007.

[8] M. Reza, Somayyeh Seifi," Classification and Evaluation the PPDM Techniques by using a data Modification -based framework", IJCSE, Vol3.No2 Feb 2011.

[9] Vassilios S.Verykios, E.Bertino, Igor N," State –of-the art in Privacy preserving Data Mining", published in SIGMOD 2004 pp.121-154.

[10] Ching-Ming, Po-Zung & Chu-Hao," Privacy Preserving Clustering of Data streams", Tamkang Journal of Sc. & Engg,Vol.13 no. 3 pp.349-358

[11] Jie Liu, Yifeng XU, "Privacy Preserving Clustering by Random Response Method of GeometricTransformation", IEEE 2010

[12] Keke Chen, Ling lui, Privacy Preserving Multiparty Collaborative Mining with Geometric Data Perturbation, IEEE, January 2009

[13] Keke Chen, Ling Liu," Geometric data perturbation for privacy preserving outsourced data mining", Springer, 2010.

[14] H.Chhinkaniwala & S.Garg," Tuple -Value Based Multiplicative Data Perturbation Approach to preserve privacy in data stream mining", IJDKP, Vol3, No.3 May 2013.

[15] Mr.Kiran Patel," Privacy Preserving Data Stream Classification: An Approach using MOA framework",GIT Vol-6,2013

[16] A. Bifet, R. Kirkby, P. Kranen and P. Reutemann, *Massive Online Analysis Manual*. May 2011.

[17] Bifet A., Holmes G., Kirkby R. and Pfahringer B., MOA: Massive Online Analysis http://moa.cs.waikato.ac.nz/, Journal of Machine Learning Research (JMLR) ,2010.

[18]Kranen, Kremer, Jansen, Seidl, Bifet, Holmer, pfhanringer," clustering performance on evolving data streams: Assessing algorithms & Evaluation Measures within MOA", published in ieee international conference, data mining workshops (ICDMW), 2010, pp.1400-1403.

[19]UCI data Repository
http://archive.ics.uci.edu/ml/datasets.

.