Indian Journal of Science and Technology, *Vol 9(47), DOI: 10.17485/ijst/2016/v9i47/101081, December 2016*

ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

CATs-Clustered k-Anonymization of Time Series Data with Minimal Information Loss and Optimal Re-identification Risk

J. S. Adeline Johnsana[1]*, A. Rajesh[2] and S. Kishore Verma[3]

[1]Department of Computer Science and Engineering, St. Peter's University, Avadi, Chennai - 600054, Tamil Nadu, India; adeline.j.s@gmail.com

[2]Department of Computer Science and Engineering, C. Abdul Hakeem College of Engineering and Technology, Melvisharam - 632509, Tamil Nadu, India; amrajesh73@gmail.com

[3]Department of Computer Science and Engineering, SCSVMV University, Kanichipuram - 631561, Tamil Nadu, India; kishore.saj3@gmail.com

Abstract

Background/Objectives: Time series is a significant type of data, widely used in diverse application such as financial, medical, and weather analyses, which in-turn contain personal privacy to a great extent. **Methods/Statistical Analysis:** The perquisite to protect privacy of time series data is to bolster the data holder to get involved in the above applications without any privacy threats. The k-anonymization approach of time series data has picked up consideration over late years, a key requirement of such an approach is to guarantee anonymization of time series data while minimizing the information loss caused from that approach. **Findings:** In this article, we implemented a novel methodology called CATs (Clustered k-Anonymization of Time Series Data) that applies the idea of clustering on time series data and ensure anonymization by gaining minimized information loss within venerable utility. The fundamental perception here is that the time series data tuples that are alike, ought to be a part of one cluster, and de-identification of these tuples is furnished. We thus formulate and proposed this approach as CATs, implemented through mishmash of WEKA and ARX anonymization tool. We have executed the solution on two benchmark time series data set available in UCR, Our experimental result strives that CATs confirms to have minimal information loss ranging from 18% to 24% reduction rate when compared with existing TSA (Time Series Anonymization) approaches. **Applications/Improvements:** As result of our experimentation, we express that our approach can play a remarkable role in the field of financial management, Online Medical process monitoring and management etc.

Keywords: Clustering, Information Loss, k-Anonymization, Privacy Preserving Data Mining, Re-Identification Risks, Time Series Data Mining

1. Introduction

At present, emotional bring up in instrumentation of the physical and virtual field has given us an exceptional chance to gather and mine helpful information from various source to comprehend essential marvels**.** Thus, in each scientific study, the estimations are performed over time instants, these perceptions prompted to an accumulation of composed information called time series data. The motivation behind time series data mining is to derive dependable learning from the state of information. Significantly, as a rule, time series errand

*Author for correspondence

depends on these basic four parts, in particular, dimensionality reduction, representation systems, distance measures, and indexing strategies. Since time series data are favoured in an extensive variety of space and applications, for example financial, retail, environmental and process monitoring, and defence and health care. Time series data-mining are going to be a vital process in extracting the useful knowledge from data collected over different time instance, which prompts to security dangers. Guarding the protection risk of time series had welcomed the late scientist to consolidate Privacy Preserving Data Mining (PPDM) Towards Time Series Data Mining (TSDM). Privacy protection of time series data for publication is a troublesome assignment because of the mind blogging nature of the data and the space they involve. PPDM algorithm can be broadly classified into four categories: Randomization, Group-Based Anonymization, Distributed Privacy Preservation, and Privacy Preservation of mining result.

Specifically, we consider the necessary problems of anonymizing the time series with negligible information loss to support the above-mentioned line-up, we devised the problem as k-anonymization of time series with minor information loss and optimal re-identification risks.

Thus, k-anonymization process attracted the focus of recent research due to its reduced complexity and ease of use, further we framed the problems to be utilized under these categories of PPDM. A critical requirement of k-anonymization of time series data is to ensure anonymization of data and promptly minimize the information loss incurred due to data distortion.

[1,2] delivered traditional approach that stood against linkage attacks of relational data.[3–5] introduced partition-based approach on trajectory and finite item set progresses by applying k-anonymization and suffer from severe pattern loss. l-diversity[6] and t-closeness[7] are proposed to overcome the two critical issues that reside with k-anonymization: (i) Back ground knowledge attack and (ii) Homogeneity attack.

[8,9]approaches are used to prevent linkage attack on time series but still suffers from change in data size after application of the process.[10–15] utilize the tactics of applying clustering to minimize information loss and subsequently guarantee good data quality that is relevant for relational data and not to time series data.[16, 17]analysed the k-anonymization process with minimum information loss over bio-medical data with different parametrical setup using ARX tools.

The objective of our strategy is to devise a novel k-anonymization system that reports minimal information loss and optimal re-identification risk over time series data utilizing the two well-known tools called WEKA and ARX. The major thought residing in our methodology is to apply k-anonymization process on the clustering issue. The k-anonymization process shall be easily treated as clustering issue where we need to locate a set of clusters and apply k-anonymization process on those clusters individually latter append the individual k-anonymized clusters ck1, ck2, ck3 into Ck and publish the Anonymized Time Series Database (ATDB).

With a specific end goal to amplify the time series data quality, we need the time series record in a group to be as like each different as could be expected under circumstances. This guarantees less distortion is required when time series records in a group are changed to have same quasi identifier values, with respect to k-anonymization by offering a least information loss.

We endeavored to formulate this specific problem as CATs (Clustered k-Anonymization of Time series Data). We applied k-means clustering on time series data and framed time series data set table into collection of clusters, where C = $(C_1,C_{2,}C_3.\ldots.C_m)$ and applied k-anonymization process on those collection of clusters to obtain k-anonymized clusters ck1, ck2, ck3. Consecutively calculated the information loss which termed to minimum when compared to the prevailing Time Series k-Anonymization (TSA) process.

We note that k-anonymization with cluster structure i.e. CATs seems to have terrible decreased information loss when compared to the k-anonymization without cluster structure i.e. TSA.

The rest of the paper is sorted out as tails; we survey the essential ideas of k-anonymity method and the most significant approach to handle time series data i.e. Dimensionality Reduction and Clustering in section(2). We discussed the necessitate initiations that are required to our approach in section(3).We formally characterize the concern of CATs (Clustered Anonymization of Time series Data) and TSA(Time Series Anonymization) and presented our approach in section(4).Then, we assessed our procedure in view of test results in section(5).We close and convey future thoughts in section(6).

2. Literature Survey

Now, we will review the existing approaches of privacy preserving data mining, especially related to time series data. The prevailing PPDM approaches broadly fall under four categories: (i) Randomization (ii) Group-Based anonymization (iii) Distributed privacy preservation (iv) Privacy preservation of mining results.

(i) Randomization procedures, utilises data twisting techniques as part of request to make private representation of the record (i.e.) by adding noise to the original data.
(ii) Group-Based Anonymization, focuses on reducing the granularity of representation of the data.
(iii) Distributed privacy preservation concentrates on employing cryptographic approaches to preserve the data, since the data is being handled in distributed environment.
(iv) Privacy preservation of mining results relies on mine- anonymize strategy which attempts to apply anonymization (i.e. granularity reduction) on the mined results.

[1]proposed a primary solution called k-anonymization to protect the databases from linkage attacks, that works on the principle of distorting each records QI attributes to be identical to at least k-1 other record's QI attributes and this method loses its effectiveness in preventing pattern linkage attacks when it is being applied on time series data.

[2,5]had implemented anonymity approaches on sequences or trajectories by k-anonymization.[5] schemed a technique that reveals the data in anonymized form, after applying generalization and then regenerating arbitrarily the trajectory from the revealed dataset, but it is liable to significant pattern loss.[2] introduced a solution to anonymize the series of determinate item sets by incorporating a prefix-tree to enrich anonymization, but this solution is limited to the strings that have exact match. Information hiding approaches are categorized into two broad classification based on the existing works: (i) Perturbation based approach and (ii) Partitioned based approach.

Perturbation based methods preserve data's by inserting some noisy data, such that the noisy data should be as much as equivalent to original data's properties, but this is seeming to be mysterious to be attained at all circumstances. Usually perturbation based method won't concentrate on linkage attacks, but which is going to be the primary concern of our work.[1,18] are the two well-versed methods that fall under partitioned based approaches.

k-anonymity with generalization plays a significant role in privacy preserving data publishing, however this approach is not well suitable to time series data. Thus here time series data patterns are severely malformed during quasi identifier's anonymization, which will be over carried by our Clustered k-Anonymization of Time Series (CATs). Certain earlier approaches[2, 5] used k-anonymization on series determinate item sets and trajectories. Often[2] able to preserve values effectively but cannot preserve time series patterns. The scheme in[5] will not be suitable with normal time series data, since it is restricted to symbolic sequences and not good enough in processing the definite values of time series data. Methodologies in[3,4] are similar to[5] but method in[3] attains privacy on sensitive patterns by randomly altering the pattern sequences, which are prone to significant pattern loss under uncertain circumstances. The approach in[4] believes that adversary is having partial background knowledge, which may lead to have privacy compromise at some extent. l-diversity[6] and t-closeness[7] are proposed as upward sequel of k-anonymization to overcome attacks that arise on account of background knowledge and homogeneous attacks on sensitive attributes. l-diversity[6] eliminates background knowledge and homogeneity attacks threats by maintaining sensitive attributes of each equivalence class, k – group to a threshold of holding minimum l representation. The semantic relationship between the attributes of the record are lost to certain extent during l-diversity process, thus t-closeness[7] strives to maintain this semantic issues. The authors of [18] derived an approach called condensation alike k-anonymity, which attempt to process the pseudo data not on the original data. This method fails in preserving the attribute's correlation in terms of individual records. Attribute correlation is an important factor when it is subjected to time series data, thus this approach is not suitable to be processed on time series data. Micro-aggregation[3] can be applied on time series data to avoid linkage attacks. However, this approach attempts to alter the size of the data set, since it groups the original data and calculates centroid for each group and replaces the original values with that of centroid that leads to dataset size reduction. Thus this method is not suitable for anonymizing time series data, which are more sensible to size of the data because each data represents a time instants and moreover micro aggregation is also subject to have severe pattern loss.

[19–21]propose distinct dimensionality reduction methodologies that are applicable to handle sparse time series data and proposed methods applicable to time series searching and indexing,[22] gives better feasibility insight about Principal Component Analysis and Random Projection dimensionality reduction techniques that are applied on time series data.[16,17] proposed ARX anonymization with generalization and subsequently with suppression, they executed the process on biomedical data and tuned monotonic parameters to attain minimal information loss. These are subject to biomedical data set that analyses the information loss through various parametrical setups not appropriate for time series data.

[10–15]proposed methods, which concentrates on blending the problem of clustering and classification into anonymization process.[11] utilizes clustering to minimize information loss and subsequently guarantee great information quality. Foremost perception here is the data tuples that are normally alike each other, ought to be a part of same equivalence class and called as k-member cluster problem applicable to person – specific data (i.e.) relational data. k-means clustering is applied to form clusters and k-anonymization is applied to each specific cluster, through which information loss is minimized. Different perspective of information loss calculation is followed in handling numeric attributes and categorical attributes. Experiments are conducted on comparisons with k-values and information loss, cardinality and execution time, classification metric.[10] views anonymization problem in terms of clustering problem, where the relational table is clustered and k-anonymization is applied by having concern on different information loss metric.[12] decides the crude cluster structure in the crude table and marks every tuple in T by a class name. This named table is signified by T_e, has a cluster attribute that contains class name for every record. Basically, protecting the crude cluster structure is to protect the way towards separating such class-name during generalization, his strategy able to get good cluster quality.

[13]endeavoured to analyse the exactness of classification through clustering using k-means & density based with and without anonymization data (Diabetic dataset).[14] proposed DRBS method which does data relocation based on sub clustering isometric transformation for maintaining similarity and correlation of the data.[15] proposed FCM clustering based perturbation for preserving the relational data with optimum trade-off. [16] proposed the scheme of k –anonymization with minimal information loss using ARX tool, the course of objective is attained by means of regulating the suppression threshold value according to feasibility requirement and[17] proposed the framework to k-anonymize the patient's health care records using ARX's k-anonymization and analysed the re-identification risk with respect to various levels of k-value.

3. Preliminaries

3.1 k-Anonymization

k-anonymization model assumes a personal data are kept in a relation (i.e.) table of attributes (i.e. column) and records (i.e. row). The definite objective of k-anonymization is to change the table as that any tuple in a table is unclear from in any event (k-1) other records. To progress in the process of k-anonymization begins with expelling all the identifiers for example S.S.N, Name, User id's etc. from the table, after removing the identifying attribute the relation is termed to have privacy threat when other attributes being linked with the publicly available databases. So the other linking attributes are derived as quasi identifiers and generalization is created on these attributes to anonymize them.

3.2 k-Anonymization of Time Series Data

k-anonymization with respect to time series data believes that each and every time series tuples in a database will constitute below mentioned three things

- An identity attribute id;
- A collection of quasi identity attribute(QIT), where each quasi identity attribute represents n time series values derived at each time instants that are different and consecutive, meant by QIT=($A_{t1,}A_{t2,}A_{t3}\dots A_{tn}$)
- A collection of sensitive attributes that are represented by A_{st} as a whole.

The sensitive attributes are the values that must not be revealed and also restricted from direct access. During anonymization, the identity of all-time series tuples is removed and each quasi identity attributes are generalised to avert linkage attacks (i.e. re-identification of A_{st}) Specifically each QIT is generalized using ARX anonymization tool, subsequently the Information loss and Re-identification risk are calculated.

3.3 Dimensionality Reduction

Time series are of high dimensions, with a specific goal to effortlessly control and oversee time series data, dimensionality reduction is utilised i.e. to lessen the quantity of random variable utilising different scientific strategies. Random Projection is a basic computationally effective approach to lessen the dimensionality of the time series data, therefore random projection dimensionality reduction method is performed using WEKA tool.

3.3.1 *Random Projection Definition*

The raw d-dimensional data is anticipated to a k-dimensional (k<<d) subspace, using an arbitrarily k × d-dimensional matrix R whose rows have unit length. Utilising matrix symbolisation: - if $X_{d \times N}$ is the original collection of N-dimensional perception the

$$X^{RP}_{kXN} = R_{kXd} \tag{1}$$

$X_{d\times N}$ is the projection of data onto a lower k-dimensional subspace.

3.4 Clustering

Clustering or cluster analysis is the assignment of collecting bunch of items in a manner, that items in a similar gathering (called a Cluster) are more comparable to each other than those in different gathering(Cluster).

3.4.1 *k-means Clustering Definition*

This is a technique for vector quantification plans to parcel n- perceptions into k-groups in which every perception has a place with the cluster, that appears to be the closest means model in the group. In this manner k-means clustering is processed using WEKA tool.

4. Proposed Work

4.1 Basic Definitions

4.1.1 *TSA – Time Series Data Anonymization*

Time series Data Anonymization believes that each and every time series tuples in a Time series Database TDB constitutes below mentioned three things,

- An identity attribute Tid;
- A collection of quasi identity attributes that represent n time series values derived at each time instants (i.e. each values are different and consecutive) meant by TQI=($At_{1,}At_{2,}At_3$….At_n)
- A collection of sensitive attributes that are represented by A_{st} as a whole ensures, that all the TQI attributes values of each tuple in the published ATDB (Anonymized Times series Data base) namely TR_1 …..TR_n are indistinguishable to at least k-1 other tuples.

4.1.2 *CATs – Clustered k- Anonymization of Time Series*

Clustered Anonymization of time series data issue is to locate a collection of clusters from the given n time series tuples (T_r), that has a place with a TDB to such an extent that every clusters contains at least(k≤n) time series data points and the entirety of all intra cluster distances are consolidated. As needs be, Let TS be the collection of n time series record and k is the said anonymized parameter, at that point, ideal CATs is as follows,

$$\mathbf{TC} = (\mathbf{tc1}, \mathbf{tc2} \ldots . \mathbf{tcm}) \tag{2}$$

Such that

1. ¥ i ≠ j €{c1….cm), $tc_i \cap tc_j = ø$
2. Ui = c_1….c_m , tc_i = TS.
3. ¥ Tc_i € TC , $|tc_i| \geq k$

Hence, $|tc_i|$ is the size of the each cluster tc . CATs time series record in a database TDB contains the set of clusters TC ($tc_{1,}tc_2$….tc_m),CATs ensures that all the TQI attribute value of each records in each clusters ATC_i=Atc_1….Atc_m (Anonymized time series Cluster) are identical to at least k-1 other records and ATDB (Anonymized Time series Database is formed by below mentioned equation 3.

$$\mathbf{ATDB} = \sum_{i=1}^{m} \mathbf{Atci} \tag{3}$$

4.1.3 *Information Loss*

Non-Uniform Entropy: This metric calculates the loss of information by accounting the loss of entropy (i.e.) Information content.

Let TDB = {Tr_1,Tr_2,……,Tr_m} be a time series database having its quasi identifiers At_1,At_2,…,At_n and V_j symbolizes the random variable that generates the value of the jth quasi identifier Atj, 1≤j≤n,in a arbitrarily selected time series record from TDB. Then, if A (TDB) = {$\overline{Tr1,}\ \overline{Tr2,}$ ……$\overline{Trn}$} is a anonymization of TDB then

$$\mathbf{TDB} \prod \square\, \mathbf{e} \left(\mathbf{TDB, A\ (TDB)}\right) = \sum_{i=1}^{m} \sum_{j=1}^{n} H(Vj \mid Tr, (j)) \tag{4}$$

is the entropy measure which denotes the loss of information caused by anonymizing TDB to ATDB.

4.1.4 *Utility Measures*

SUM: This metric calculates the utility by accounting the utilities of all quasi identifiers At_n.

Utility measures of time series database TDB is calculated based on the individual information loss of each quasi identifiers At_n (called multi-dimensional). SUM, is the aggregate function, which define how the individual measures for each quasi identifier will be gathered into a comprehensive measure for the whole dataset.

4.1.5 *Re-Identification Risk*

The uncertainty that may exist in identifying a time series record or its definite attributes by linking the exposed relation with publicly available time series dataset with the assistance of having some background knowledge called as Re-Identification Risks.

Re-Identification Risks are measured based on the unique records that reside in the Time series DataBase (TDB). Following metric is derived to calculate the uniqueness of the records which termed to be Re-Identification Risk.

Let TDB_{TA} be the publicly available time series database, TA is the uncovered sample. The Re-Identification parameter RI_{TDB} is estimated by dividing the number equivalence class that are distinctive in both (TDB_{TA} and TA) by the number of records in TA.

$$RI_{TDB} = {}^{1}\!\Big/_{(n\ )} \left(\Sigma_i [\![C\, (\vartheta i = 1,]\!]\ \omega i = 1)\right) \tag{5}$$

where, $\supseteq\ \omega i$, ϑi , ωi are mentioned as equivalence class of TDB and TA respectively and C( ) meant uniqueness function.

4.2 Objective

Our motive is to prevent privacy breach of time series membership disclosure and attributes disclosure with minimal information loss and optimal Re-Identification risk factor.

Avoiding such privacy breaches with minimal information loss and optimal re-identification risk factors, our novelized approach had given a way to incorporate two bench mark tools i) WEKA and ii) ARX here to achieve the objective.

(i) WEKA is employed to perform a pre-processing phase that comprises dimensionality reduction and a mining phase that implicates clustering process over the given time series data. Thus, WEKA is one among righteous tool to be involved in data mining process because of its versatile role in development and usage of classification, association rule mining and clustering in the data mining community.

(ii) ARX is utilised to forestall security rupture; ARX actualizes protection model that apply syntactic protection criteria on the time series data. This tool obviously gives way to arbitrary combination of well privacy approaches (1) k-anonymization (2) l-diversity (3) t-closeness (4) k-map and (5) (e,d)-differential privacy . Thus, the focus of our work is to apply anonymity with novel perspective.

WEKA supports well-known mechanisms: (i) Random Projection and (ii) Principal Components towards Dimensionality Reduction and (i) Hierarchical Clustering (ii) EM Clustering (iii) Simple k-means (iv) Make Density Based Clustering towards Clustering. According to the feasibility and essential perception Random Projection and Simple k-means were employed to yield the better result that provisions our motive.

4.3 System Architecture

Our methodology CATs consists of three significant procedures

(i) Dimensionality Reduction (ii) Clustering and (iii) Anonymization.

Figure 1 depicts CATs state-of-art performing the anonymization process over time series data. Accordingly, the data provider supplies the raw time series data, due the dimensionality curse the classic problem always associated with the time series data, Dimensionality reduction is being performed, consequently acquired low dimensional data is given to the clustering process to form clusters of time series data that are really close to each other in terms of distance of the values. Then these formed clusters are anonymized individually and appended together to form the ATDB (Anonymized Time Series Database).

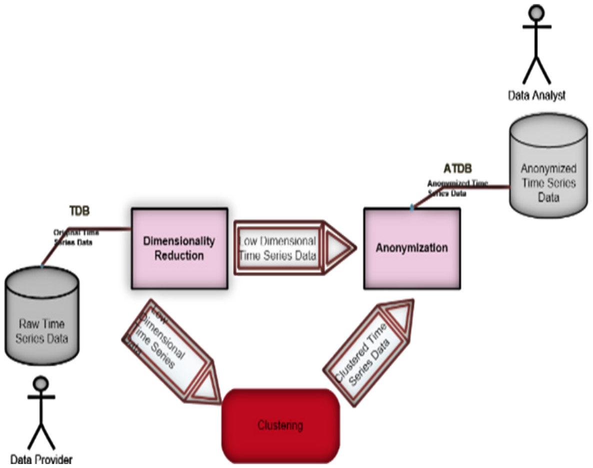
Data Analyst
Anonymized Time Series Data
ATDB
Anonymized Time Series Data
TDB
Original Time Series Data
Dimensionality Reduction
Low Dimensional Time Series Data
Anonymization
Raw Time Series Data
Low Dimensional Time Series Data
Clustered Time Series Data
Clustering
Data Provider

Figure 1. System architecture of CATs.

4.4 Coding Model and Solution Space

A novel combination of WEKA and ARX was utilized to frame the coding model and solution space for our proposed state of work that stood well against the privacy threats remains with time series data. By having minimal information loss when compared with previous approach at expense of admissible increase in the re-identification risk factor. We would like to explain our process in terms of two aspects, on part of WEKA tool; we utilize to perform dimensionality reduction .WEKA offers two types of dimensionality reduction method (i) Principal Component Analysis (ii) Random Projection .According[21,22] authors conveyed, usually principal component analysis gain importance in terms of accuracy than random projection subject to the classification, whereas recently random projection is gaining popularity due to its feasibility towards the computational complexity and the accuracy factor is terming only to classification not to clustering. Thus in our work we are going to perform clustering, which appealed us in employing random projection towards our procedure.

Second, on part of anonymization we utilized ARX anonymization tools. ARX anonymization tools offer seven different privacy models to support privacy preserving data mining. To reinforce our claim, we experiment to apply k- anonymization, by tuning different values for k and analysing the optimal k value for k-anonymization of our claimed attempt.

However, as part of our work, we carry out the procedures as follows:

Step 1: Reduce High dimensional time series data (Dimensionality Reduction)
Step 2: Apply clustering to the low – dimensional time series data (Clustering)
Step 3: Generalize each clustered time series data individually and append the individual
anonymized Clustered time series to form ATDB (Anonymized Time Series Database) (ARX anonymization)
Step 4: Publish anonymized time series database (Privacy Preserving time series data publishing)

Step 1: Dimensionality Reduction

Dimensionality reduction is widely applied to trim down the problem of controlling and operating high dimensional time series data set. Dimensionality reduction adopts linear transformation in deciding basic dimensionality of the stat and as well deriving its principal directions. Underlying this standard, different methods have been developed, including principal component analysis, canonical analysis, linear discriminant analysis, discrete cosine transfer and random projection etc.

Random projection has increased late significance because of its effortless and proficient computability, this strategy trims down the dimensionality of time series data by offering restrained measure of error for speedier processing and modelling data in smaller size. On account of its simplicity, error controlling rate and computational efficiency, we incorporate random projection process to diminish the dimensionality of our large time series data to be utilized for our state-of-art.

Random projection with WEKA: Here WEKA a powerful data mining tool offers a filter random project which term to diminish the dimensionality of the large time series data set.

Procedure to apply Random projection with WEKA:

i) Launch WEKA, Choose Explorer
ii) Load time series data set – Do open file (CSV, XLS format)

(i) Choose filter option, select unsupervised, opt for attribute and choose random projection method.

- Set distribution according to data set type, in this regard of time series data we have choosen Gaussian distribution.
- Set number of attributes required to reproduced and also set random seed value.

(ii) Apply Random Projection filter adopting mentionedparametersandsavethefileinCSV,ArfforXLS format.

Step 2: Clustering

Time series clustering have its diversion in all aspects, when composed to normal clustering process. Like static data clustering process, time series clustering also in need of a clustering algorithm or procedure to form clusters on a given set of unlabelled data points. Adoption of clustering algorithm mainly resides on the type of data available and purpose to which it will be applied. Up to this extent, time series data are exercised, featuring whether the data are discrete/ real valued uniform/non uniform, uni variate/multi variate and finally data length equal / unequal. Non –uniform time series data should be transformed into uniform time series before applying the clustering procedure.

A variety of algorithms like Relocation clustering, agglomerative clustering, hierarchical clustering, k-means clustering and fuzzy c means clustering have been proposed to perform clustering. Out of which we attempt to employ k-means clustering on account of its feasibility, simplicity and computationally faster. k-means can able to produce tightly coupled clusters, best suited for numeric high dimensional data set, fast robust and relatively efficient but sensible to outliers that arise.

Algorithmic steps for k-means clustering on time series data.

Let TDB = (t1, t2…..tn) be the set of time series data point and VT (VT1, VT2,……VTn) be the set of centres.

i. Randomly choose time series cluster centres.
ii. Compute the distance between every time series data point and time series clusters.
iii. Assign the time series data point to the time series cluster centre whose distance from the time series clusters centre is least of the whole time series cluster centres.
iv. Re-compute the new time series cluster centre by VT = (1/C)

$$\sum_{j=1}^{c_i} tj \qquad (6)$$

$V_T = (1/c)$

where ‘tc_i’ represent quantity of time series data points in i^th time series cluster.

(v) Re-compute the distance between each time series data point and newly acquired time series cluster centre.

(vi) On the off chance that no time series data point was reassigned then stop otherwise rehash from step iii.

K-means clustering of time series data with WEKA.

WEKA is the most powerful data mining tool which were being used by most of the data mining researchers’ community, which offers a wide range of clustering algorithms like EM, filtered clusters, hierarchical clusters, density based clusters, and simple k-means clustering. On behalf of vigorous advantages of k-mean clustering as said before, we adopt simple k-means algorithm and procedure to apply simple k-means algorithm to time series as follows.

Procedure to apply simple k-means clustering with WEKA

(i) Choose the cluster tab, under list of clusters, select simple k-means.
(ii) Set cluster node as use training set.
(iii) Set all require parametrical value that closely suits simple k-means algorithm
(iv) Check store clusters for visualization to export the clustered time series data
(v) Apply simple k-means process by having mentioned parametrical setup on low dimensional time series data.
(vi) Export the Clustered Assignment time series data.

Step 3: ARX Anonymization

The basic steps involved in anonymizing time series data with respect to ARX

anonymizing tool is as follows

(i) Configure transformation of time series data

a. Define transformation model, Import Input time series data (CSV, XLS…etc.)
b. Generalization hierarchies of all quasi identifying attributes are created / imported.
c. Define privacy model and criterion, Set type of privacy model to k-anonymization.
d. Define coding model, set suppression limit, Define utility measure, Set information loss measure to non-uniform entropy and Set aggregate function to SUM, Define Attribute Weight, Set equal attribute weight for all involved quasi identifiers.
e. Run anonymize.

ii) Exploring the solution space of generalized time series data.

- Tabulate the viewed Information loss (i.e. calculated using equation 4) arisen on account of executing k-anonymization.

iii) Analyse input/output time series data

- Compare transformed time series data set to the original data set.
- Direct analyse risk phase, calculate re-identification risk (i.e. calculated using equation 5) involved in the transformed data set, Average risk is taken into validation.
- Repeat from Step (i) – e, formulate the privacy model for distinct requisite values of k.

Step 4: Privacy preserving time series data publishing

Append anonymized version of each cluster's time series records into sole relation and publish as whole Anonymized Time series DataBase (ATDB).

5. Experimentation and Results

Here we explain how experimental setup have been set and evaluated to support and calculate our approach with existing methods by having two bench mark time series data sets from UCR (23) i) ECG and ii) Two Patterns, accommodating different parametric privacy criterions.

5.1 Data sets

Necessary parametrical setup and expected outcomes of anonymization process widely depend on the input time series data we use in our experiment. Underlying a broad spectrum of data analysis, we used two real-world time series data that are publicly available in UCR dataset23. The UCR dataset were extensively used in time series data mining and we adopted two datasets ECG 200 (106-time series records of 97 length) and Two patterns (1000-time series records of 128 length), that are comfortable for evaluating clustering and classification algorithm on time series datasets. Mostly already been utilized to compare and evaluate existing approaches on time series data anonymization. An outlier on basic properties of the pre-processed and mined (Dimensionality Reduction and Clustering) datasets are appeared in Table 1.

Table 1. Pre-processed and mined data set properties

Dataset	Quasi Identifier	Sensitive Attribute	Identifying
ECG 200	K1 to K10	att 97 (least attribute value)	Cluster
Two patterns	att 2 to att 128	att 129	Cluster

This dataset consists of 100 to 1000-time series data, which is of 97 to 128 length. When selecting the quasi identifiers and transformation process we apply the same model to all of the quasi identifier except the last attribute (i.e.) sensitive attribute (att 97 and att 129). Generalization hierarchies consist of 2 to 6 generalizations levels were created for each quasi identifier using create hierarchy functionality of ARX anonymization tool.

5.2 Privacy Models and Parameters

Existing evaluation and analysis of anonymization algorithm focused on the whole time series data with varying k-parameters values. These parametric evaluations with different value of k in turn influences with varying information loss.

Proposed framework's evaluation and analysis is carried over on each clustered time series data separately with varying k-parameter values. The summation on the information loss of each cluster is performed and total information loss is calculated and compared with the existing solution which termed to have a better outcome subject to reduced information loss. As part of dimensionality reduction we opted random projection method in WEKA for feasibility and robustness, applied its parametrical setup as distribution = sparse, number of attributes = 10, present: 0.0, random seed 42, replace missing values to false. Consequently, random projection reduced the dimensionality of the time series data to 10 attributes (i.e. k1 to k10). Then the low dimensional datasets are given towards anonymization process, as a parameter for k-anonymity towards this approach k, the evaluation relies on three factors:
(i) k-value ii) Information loss iii) Re-identification risk

For varying values of k i.e. 2,4,6,8,10,12,14,16,18,20 the information loss (Non -Uniform measure) and re-identification risk is calculated and depicted in below mentioned figures. The Information Loss and Average risk analysed toward the figure clearly pictures the Information Loss is due directly proportional to k value. The re-identification risk is due inversely proportional to k-value.

CATs parametrical setup and execution

- As part of CATs Dimensionality reduction (Random Projection) method have been performed with WEKA by adopting with same parametrical setup as mentioned.
- Then the Low dimensionality time series dataset is given toward clustering process in WEKA.

We opted for k- means clustering in WEKA for its popularity in handling numeric data points. Thus we employed simple k-means clustering over the UCR ECG and Two Patterns dataset by having its parametrical setup as follows i) distance function: Euclidean distance ii) max iteration 500 and iii)numclusters = 3 iii)opted for use training set under cluster mode, iv) checked store cluster for visualization , after applying the clustering process in WEKA by agreeing the mentioned parametrical setup ,we exported three clusters namely cluster0, cluster1 and cluster2 time series data respectively. These cluster0, cluster1 and cluster2 time series data need to undergo anonymization exclusively, then the low dimensional clustered time series data cluster0, cluster1 and cluster2 have been employed in ARX k-anonymization process. As a part of parametrical setup for k-anonymization towards CATs approach the evaluation again relies on three factors of each individual clusters (cluster0, cluster1 and cluster2).

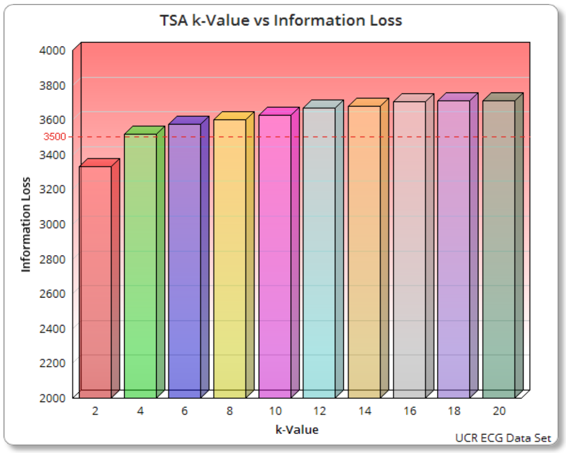

Figure 2. TSA approach: - k-value vs. information loss for ECG data set.

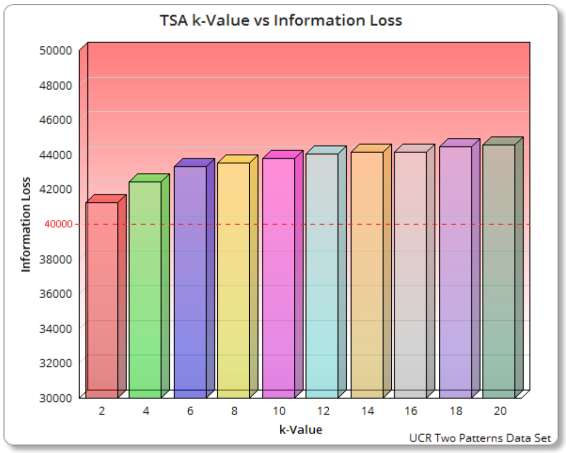

Figure 3. TSA approach- k-value vs. information loss for two pattern data set.

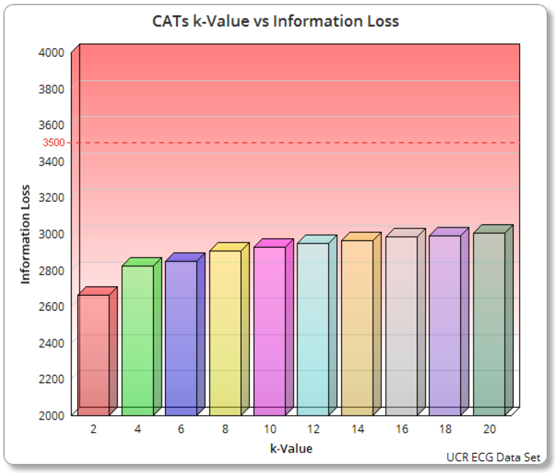

Figure 4. CATs approach:- k-value vs. information loss for ECG data set.

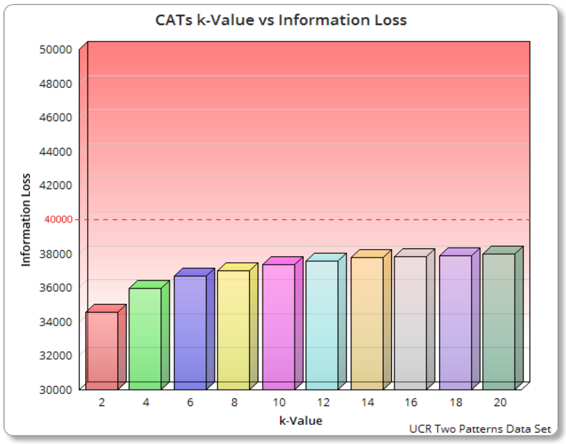

Figure 5 . CATs approach: k-value vs. information loss for two patterns data set.

i. k-value
ii. Information loss
iii. Re-identification risk

According to TSA and CATs approach, for varying values of k i.e. 2,4,6,8,10,12,14,16,18,20 the Information loss (Non- uniform Entropy measure) and Re-Identification Risk % are calculated for two benchmark dataset (ECG and Two Patterns) and depicted in Figures 2–9. Analyses over these graphs clearly reveals the two certainties,
i) Information loss α k-value ii) Re-identification risk α 1/k value

Figures 4, 5 definitely states CATs Approach leads minimal information loss when compared to existing approach.

By analysing the comparative results of ECG time series data portrayed in Figure 10. Our CATs approach gives the optimal result in terms of Information loss (i.e. 24 % reduction rate when compared with the TSA Approach). By analysing comparative results of Two Patterns time series data portrayed in Figure 11, our CATs approach excels to deliver optimal results in terms of Information loss (i.e. 18 % reduction rate when compared with the TSA Approach). Through our experimental result, we strive to state that our novel execution approach CATs confirms to deliver optimal information loss subject to 18 % to 24 % reduction rate.

Though by analysing the comparative results of Re-identification Risk % of CATs and TSA shown in Figures 12 and 13, for smaller values of k our CATs approach miscarries to give optimal Re-identification

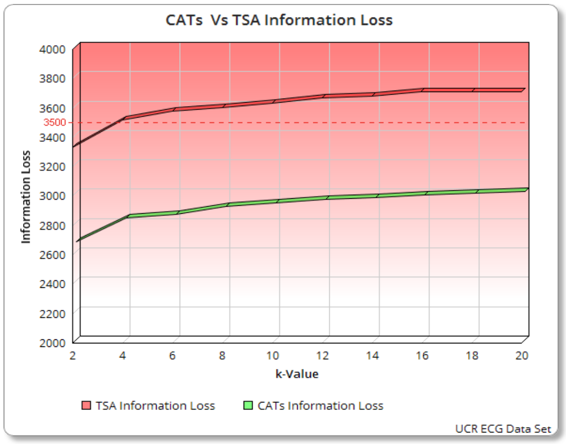

Figure 6. Comparative results of CATs and TSA approach of ECG Data Set w.r.t. information loss and k-value.

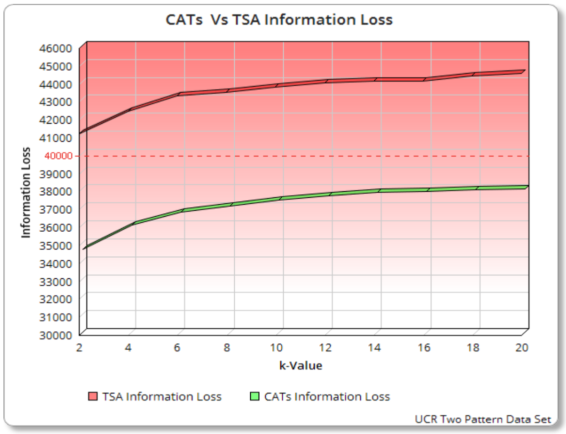

Figure 7. Comparative results of CATs and TSA approach of two pattern data set w.r.t. information loss and k-value

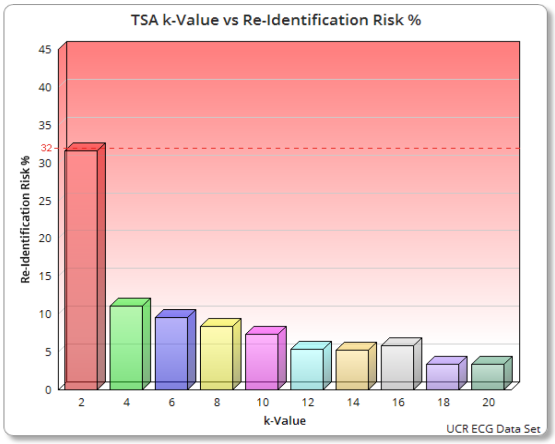

Figure 8. TSA approach: - k-value vs. identification risk% for ECG data set.

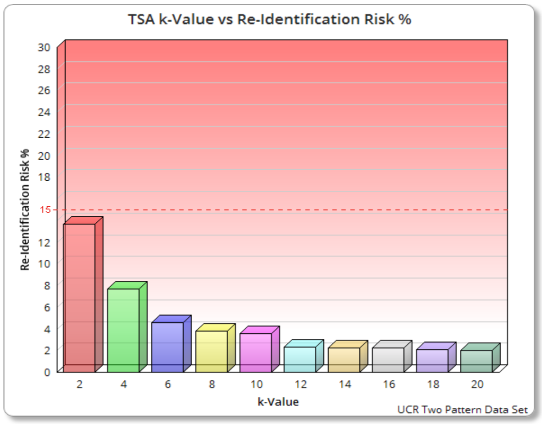

Figure 9. TSA approach: k-Value vs. re-identification risk for two patterns data set.

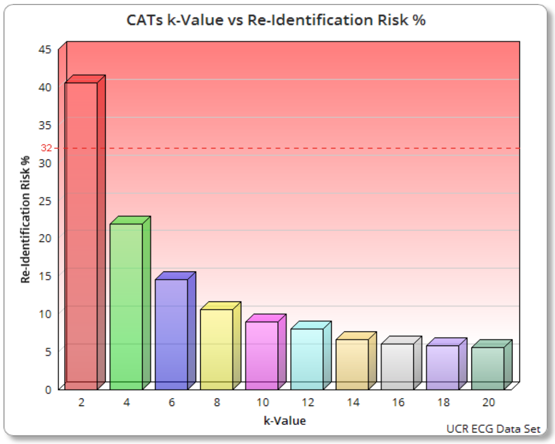

Figure 10. CATs approach: k-value vs. re-identification risk % for ECG data set.

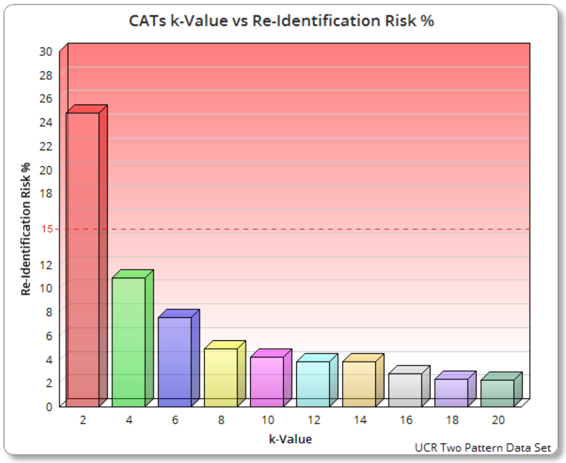

Figure 11. CATs approach: k-value vs., re-identification risk % for two pattern data set.

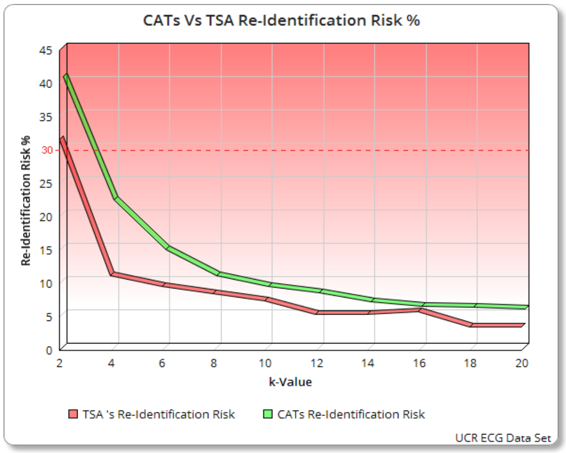

Figure 12. Comparative results of CATs and TSA approach of ECG data set w.r.t. re-identification risk and k-value.

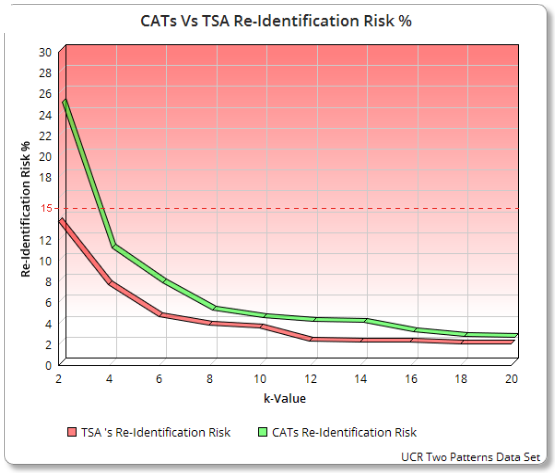

Figure 13. Comparative results of CATs and TSA approach of two pattern data set w.r.t. re-identification risk and k-value.

Risk % factor less values of k i.e. <8, for k value ≥ 8 optimizes with the existing TSA approach. From this fact, CATs is advisable for k value ≥ 8 to outperform the TSA approach with respect to the Re-identification Risk% factor. Thus time series data are usually sparse in nature, k-Anonymization with k< 8 do not gain proper reflection in anonymization and CATs with optimal k value > 8 outperforms all existing approaches. Experimenting with our approach had given insight that CATS is purely applicable to anonymize the time series database with k-anonymizing strategy with k value >8.

6. Conclusion

In this paper, we proposed competent approach for time series data by renovating k –anonymity problem of time series to the CATs problem. Our novel approach CATs experimentation results conformed to have minimal Information loss (i.e. 18 to 24 %) reduction rate and is advisable for k value ≥ 8 perform well than all existing TSA approach with respect to the Re-identification Risk% factor. Our approach is the primary deliberate experimentation portraying on the most proficient method to execute clustering and generalization, when time series data is to be anonymized with minimum information loss, using WEKA and ARX tools. CATs may prompt to a couple of intriguing directions for future review as follows,

i. Exercising with different clustering methods and loss functions.
ii. Applying on different data models

We will put forth our effort to work on above-mentioned directions.

7. References

1. Sweeney L. k-Anonymity: Privacy protection using generalization and suppression. International Journal Uncertainty Fuzziness and Knowledge-based Systems. 2002; 10(5): 571–88.
2. Pensa RG, Monreale A, Pinelli F, Pedreschi D. Pattern-preserving k-Anonymization of sequences and its application to mobility data mining. International Workshop Privacy in Location-Based Applications (PiLBA); 2008.
3. Abul O, Atzori M, Bonchi F, Giannotti F. Hiding sequences. Proceeding IEEE 23rd International Conference. Data Engineering (ICDE) Workshops, India; 2007. p. 147–56.
4. Mohammed N, Fung BCM, Debbabi M. Walking in the crowd: Anonymizing trajectory data for pattern analysis. Proceeding 18th ACM Conference Information and Knowledge Management (CIKM), China; 2009. p. 1441–4.
5. Nergiz ME, Atzori M, Saygin Y. Perturbation-driven anonymization of trajectories. Technical Report 2007-TR-017, ISTI-CNR; 2007.
6. Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M. l-diversity: Privacy beyond k-Anonymity. 22nd International Conference on Data Engineering (ICDE); 2006. p. 24.
7. Li N, Li T, Venkatasubramanian S. t-closeness: Privacy beyond k-Anonymity and l-Diversity. IEEE 23rd International Conference on Data Engineering (ICDE); 2007. p. 106–15.
8. Papadimitriou S, Li F, Kollios G, Yu PS. Time series compressibility and privacy. 33rd International Conference on Very Large Data Bases (VLDB); 2007. p. 459–70.
9. Singh L, Sayal M. Privacy preserving burst detection of distributed time series data using linear transforms. IEEE Symposium Computational Intelligence and Data Mining (CIDM); 2007. p. 646–53.
10. Malaisamy A, Nawaz GMK. Data privacy using k-Anonymization with clustering technique. International Journal of Innovations in Engineering and Technology. 2016 Feb; 6(3).
11. Byun J-W, Kamra A, Bertino E, Li N. Efficient k-anonymization using clustering techniques. Advances in Databases: Concepts, Systems and Applications. 2007; 4443:188–200.
12. Fung CMB, Wang K, Wang L, Debbabi M. A framework for privacy preserving cluster analysis. ISI 2008 Jun IEEE, Taipei, Taiwan; 2008.
13. Mandapati S, Bhogapathi RB, Rao MVPCS. Classification via clustering for anonymization data. International Journal of Computer Network and Information Security. 2014; 3:52–8.
14. Rajalakshmi V, Mala GSA. Anonymization by data relocation using sub-clustering for privacy preserving data mining. Indian Journal of Science and Technology. 2014 Jul; 7(7):975–80.
15. Hariharan R, Mahesh C, Prasenna P, Kumar RV. Enhancing privacy preservation in data mining using cluster based greedy method in hierarchical approach. Indian Journal of Science and Technology. 2016 Jan; 9(3).
16. Kohlmayer F, Prasser F, Kuhn KA. The cost of quality: Implementing generalization and suppression for anonymizing biomedical data with minimal information loss. Journal of Biomedical Informatics. 2015; 58:37–48.
17. Taneja H, Kapil, Singh AK. Preserving Privacy of Patients based on Re-identification risk. Procedia Computer Science. 2015; 70:448–54.
18. Aggarwal CC, Yu PS. A condensation approach to privacy preserving data mining. Ninth International Conference on Extending Database Technology (EDBT); 2004. p. 183–99.
19. Keogh EJ, Chakrabarti K, Mehrotra S, Pazzani MJ. Locally adaptive dimensionality reduction for indexing large time series databases. Proceedings ACM SIGMOD Conference; 2001. p. 151–62.
20. Nin J, Torra V. Towards the evaluation of time series protection methods. Information Sciences. 2009; 179(11):1663–77.
21. Keogh EJ, Chakrabarti K, Pazzani MJ, Mehrotra S. Dimensionality reduction for fast similarity search in large time series databases. Knowledge Information Systems. 2001; 3(3):263–86.
22. Deegalla S, Bostrom H. Reducing high-dimensional data by principal component analysis vs. random projection for nearest neighbor classification. 3rd International Conference on Intelligent Computational Systems (ICICS'2013) 2013 Apr 29–30, Singapore; 2013.
23. Keogh E, Xi, Wei L, Ratnamahatana CA. The UCR time series for classification/clustering [Internet]. Available from: http://www.cs.ucr.edu/~eamonn/time_series_data.