I J C T A, 9(28) 2016, pp. 325-341

(k, P^d) ANONYMIZATION OF TIME SERIES DATA AND PATTERN REPRESENTATION

Adeline Johnsana. J.S* Rajesh. A and Kishore Verma. S*****

Abstract: Time Series Data Mining (TSDM) is latest consideration of researchers in Data Mining community because of extensive accessibility of data with temporal dependency. The apparent use of these data on the internet has fostered the most pioneering applications encompassing from financial analysis to social community, tracking and partner matching. However, such an appealing data usage over the applications as well suggests incredible measure of privacy to be considered, which if not properly ensured, might get to be demoralised as a hotspot for abuse and violations.

A variety of schemes have been proposed to protect privacy on time series, out of which the traditional k-anonymization gained the prominence, however it arises up in providing restricted fortification to the patterns of the time series data as it might suffer extreme pattern loss. Recently (*k*,P) anonymity method have been proposed to protect the value and as well as the patterns of the time series. This method models the pattern representation using the SAX method which accounts only average, through which it can compute the PRs of coarser granularity, often fails to have convincing pattern loss, which being overwhelmed by our previous proposal Dμ-SAX and knob ripping. This method accounts the trend values and produces the pattern representation (PRs) of finer granularity and strives to progress in offering limited pattern loss than (*k*,P) anonymity model. Now we propose (*k*,P^d) anonymity model , which is a novel exertion of mishmash k-anonymization ,R-SYMPOL,AMSS distance function and Pattern Anonymizer proficient in generating fine granular pattern representation than other existing two schemes and restricted pattern loss comparatively in almost all circumstances than (*k*,P) anonymity model and our own previous proposal Dμ-SAX Knob ripping methods. We meticulously experimented our framework on two bench marks datasets GDP,ECG and on synthetic data set all our experimental results confirms to reveal the excellence of our approach in terms of granularity of pattern representation and pattern loss.

Keywords: Privacy, Symbolic polynomials, k-Anonymity, pattern preservation, and time-series

1. INTRODUCTION

Now a days, mostly every scientific researches and observations are performed in collecting a structured data over time instances, these data's are denoted as time series data. The process of collecting those structured data over time instance and analysing to extract some meaningful knowledge over them are called as time series data mining. Subsequently, time series is the merely mode to quantise significant and typical non-reproducible temporal information, it focuses time series data to be vitally ever-present phenomenon in both scientific area and day-to-day life. For instance, it usually have communal to realize time series had grounded its impact in denoting financial data, weather forecasting, social trends and medical observation like blood pressure, heart beat rate or body temperature etc.

Time series data mining focuses to mishmash various statistical and pattern recognition technique to extract useful information hidden in time-ordering of samples. The publishing of these information have given way to create privacy threats to the individual or the organisation involved in the event. Thus there

* Research Scholar, Department of Computer Science Engineering ,St.Peter's University, **Email:** adeline.j.s@gmail.com
** Professor and Head, Department of Computer Science Engineering ,C.Abdul Hakeem College of Engineering & Technology, Email: amrajesh73@gmail.com
*** Research Scholar, Department of Computer Science Engineering, SCSVMV University, **Email:** kishore.saj3@gmail.com

being an implicit need to protect the time series data without affecting its utility. These reasons grasped the attention of the researchers to incorporate privacy preserving schemes into time series data mining. Though a variety of privacy preserving data mining approaches is being existing, privacy protection in the publication of the time series data is a mysterious task, typically due it's composite in nature and the manner they are used.

The traditional solution to prevent linkage attack is to impose k-anonymity [1],[2]on the original database, which results that where every record's quasi identifiers(QI) attributes appears to be alike at least k-1 other records. Thus k-anonymization [1], [2] suffers from uniformity and background knowledge attacks. [3] Condensation method are similar to [1],[2] which splits the data set into multiple sets of predefined size adhering k-anonymization strategy. Certain privacy schemes l-diversity [4], t-closeness [5] evolved to prevent the k-anonymization [1], [2] from background and homogeneity attacks. [4] l-diversity strives to resist the homogeneousness and background knowledge attacks by promising that at least l sensitive attributes present in each k group. [4] t-closeness strives to resist semantic issues that arise on account of applying l-diversity scheme. [6] proposed methods to perform anonymization based on utility using local recoding method of generalization.

Focussing on privacy preserving time series, the existing schemes may fall under two broad categorization as follows

(i) Perturbation Method

(ii) Partition Method

Perturbation method agrees to distort the data by adding noise to the original micro data, where added noise should have their properties nearly equal to the original data, which may complicate the attacker to discern the perturbed data from the original data.

Partition methods divides the records of the dataset into separate sets and reveals some generalised information of each set. [7], [8], [9] proposed schemes that are applicable to sequences and trajectory database, suffers from enormous pattern loss. [10] have been proposed to anonymize the finite item set sequences using prefix tree, not applicable to general time series data. However the schemes conversed above are not possibly capable to addressing the privacy issues existing with time series data and as well as the patterns generated on those data simultaneously. Recently [11] (k,P) anonymity and our own previous work [12] Dμ-SAX with knob ripping models have been proposed for preserving the data and patterns of time series data. [11] (k,P) anonymity model employs modified k-anonymization [1], [2] strategy to anonymize the data, a well-known pattern representation (PR) method [12] SAX to represent the patterns of the time series and KAPRA algorithm to anonymize the patterns. Nevertheless the model fails to produce patterns at fine granular patterns due to constraint imposed on pattern representation, which in turn suffers from huge pattern loss for higher values of P. Our previous work [13] attempts to replace the SAX by Dμ-SAX, which is the modified version of [14] 1-D SAX to crack the PR constraint of [11] while generating the fine granular PRs and also employed knob ripping a top down approach to anonymize patterns with minimal loss than the [11] (k,P) anonymity model.

Now, we effort fully propose (k,P^d) anonymity model which incorporates three innovates (i) R-Sympol, a revised method of [15], which applicable to repetitive sliding window time series data, to compute PRs (ii) [16] ASSM(Angular Metric for Shape Similarity) metric to calculate the distance between original and anonymized pattern and (iii) A PA(Pattern Anonymizer) to reduce the granularity of the generated patterns.

Thus by incorporating the R-SymPol and ASSM into our (k,P^d) anonymity model effectively produces fine granular patterns and reduced pattern loss arise on account of pattern anonymization when compared

with existing approaches specifically (k,P) anonymity model and our own previous work Dμ-SAX with knob ripping.

(k, P^d) anonymity model originates in anonymizing the time series data values by adopting necessaries from traditional [1], [2] k-anonymization strategies conforming k-requisite and proceed towards computing information rich PRs and generalizing it conforming P^d - requisite using R-SymPol and PA. In our (k, P^d) anonymity model k denotes the factor used to anonymize the time series data values, P^d denotes the factor anonymizing the patterns. Since our method R-SymPol does the symbol representation based polynomial computation, d denotes the degree of the polynomial co-efficient derived as a controlling factor in producing generalised patterns. Our innovative mishmash of all three schemes comfortably excels in generating fine granular pattern by cracking the PR constraint in minimal pattern loss.

The rest of the paper is organised as follows Section 2 Summarizes the existing approaches and related works, Section 3 discusses the basic necessitates that framed our proposed work, Section 4 Introduces our novel (k, P^d) model, Section 5 evaluates the effectiveness and efficiency of our (k, P^d) model with existing approaches and Section 6 conveys the conclusion and possible future directions reside through our work.

2. LITERATURE SURVEY

In this section, we precise the existing approaches of privacy preservation in time series data mining. We would like to categorize the existing works into two broad classification

(i) Perturbation Methods

(ii) Partition Methods

Perturbation methods distorts the data by generating noise to the original data under a constraint that generated noise should fulfil certain conditions or adapt some kind of distribution on the distorted data to have numerous characteristics that are compatible with the original data. Perturbation does not focus on preventing the data from linkage attacks which is being deviated from our objective.

Partition methods divides the records of the given dataset into separate sets and reveals some general information representing each set. K-Anonymity [1], [2], Condensation [3] ,Micro Aggregation[17] and Slicing[18] are well-known methods of this classification.

k-Anonymization [1], [2] is one of the necessitate methods used in privacy preserving data publishing(PPDP), this methodology efforts to reduce the granularity in representing the quasi identifies, which acting as core grounds substantial in executing the linkage attacks. Two main techniques Generalization and Suppression were employed in k-Anonymization [1], [2] approaches, out which Generalization had gained its reputation due its wide range of practise. Though k-anonymity is a notable technique for its simplicity, nevertheless it is vulnerable to two kinds of privacy disputes (i) Background Knowledge attack and (ii) Homogeneity attack. Obviously k-anonymization is effective in preventing identity disclosure not to the above said disputes. l-diversity [4] was proposed to sustain the minimum set size to k and sensitive attribute's diversity, t-closeness [5] was proposed to maintain the semantic relationships between attributes values and records, which got disturbed due to partitioning. Condensation [3] method splits the micro data into multiple sets of pre-defined size, both k-anonymity [1], [2] and condensation efforts to divide the micro data into sets, the factor that differentiates them is, Condensation [3] is applied on pseudo data not on the original data, whereas k-anonymity is applied on the original micro data. Condensation [3] fails in preserving the correlation between the attributes of the individual records, not suitable for time series data. [7] proposed perturbation driven k-anonymization method to anonymize the trajectories, which strives to apply k-anonymization requirement for the trajectory datasets. This method resolve the privacy issues concerning the individuals involved during trajectories data sharing, this

scheme addresses the privacy disputes in two aspects (i) applying *k*-anonymity to *k* users or trajectories and (ii) recreating arbitrarily an original data from the anonymized data set, it is disposed to significant pattern loss. [8] conveyed thriving scheme to preserve the structure of the time series data to withstand the privacy breaches that stood against. This method concentrates on uni-variate time series, by introducing compressible distortion that regulates the distortion of time series data subjecting distortion's magnitude and data prospectus, this method too flops to provide tolerable pattern loss due to anonymization. [9], [10] schemes focuses on imposing *k*-anonymization on sequences and trajectories, [9] fails to proceed if the adversary have the unconstrained background knowledge. [10] is applicable to finite set of sequences and it employs prefix tree to realize the anonymization, however the pattern similarity loom involved in this scheme restricted to exact string match and not applicable to time series data. [17], [18] are well known methods of partitioning , on resultant of anonymization process these schemes alters the micro data through partitioning by employing their unique procedures. [17] Micro aggregation utilized to avert the linkage attacks that are meant with micro data, it proceeds in clustering the micro data into clusters and substitutes the original values with the centroids of the formed clusters. However this method will noticeably change the size of the data set involved in the process, which leads to be a cumbersome task for quantity sensitive applications, collectively anonymizing through micro aggregation suffers uncontrolled pattern loss. [18] Slicing mishmashes both vertical partitioning and horizontal partitioning to preserve the dataset from membership disclosure as well as gaining the utility. Vertical partitioning divides attributes into columns considering the associations between the attributes. Horizontal partitioning divides records into buckets, lastly, inside every bucket, values in every column are arbitrarily permutated to disrupt the linkage among different columns. Vertical partitioning employed in slicing is not appropriate for time series data since trends and seasonality properties get affected. [19] attempts to apply *k*-anonymization for N-grams of the time series data, it protects the rare N-grams from other data, by confirming their frequencies to at least k, this scheme is not applicable to the entire crowd of time series data. [11] (*k*, P) anonymity model were proposed to protect the privacy breaches that exist in values and patterns representations of time series data, this method put-up *k*-Anonymization[1], [2], SAX[12] and KAPRA to anonymize values conforming k requirement, to represent patterns and to anonymize pattern representation conforming p requirement. Thus the employed SAX [12] method generates the PRs by accounting the average values not the trend values, which in turn imposed constraint on generating fine granular patterns with goodness of fit ,that leads to significant pattern loss to higher values of P. Our own previous work [13] Dμ-SAX with knob ripping attempts the resolve the PRs constraint in extracting the fine granular PRs of [11] and yields quality PRs with fine granularity with reduced pattern loss than [11] revealed owing to generalizing the pattern representation confirming P- condition. In [13] we incorporated Dμ-SAX a variant of 1-D SAX [14] , which computes the PRs by containing the trend values associated with the time series data and knob ripping to generalize the computed pattern representation confirming P-condition with minimal pattern loss than [11] (*k*,P) anonymity model.

3. BACKGROUNDS

3.1 k-Anonymization

k-anonymization model assumes an individual's private data are kept in a relation (i.e.) table of records and attributes. The vital goal of *k*-anonymization is to transform the relation as that any record in a relation is indistinguishable from at least (*k*-1) other records.

3.2 k-Anonymization of time series data

Each time series record (T_r) in a time series database TDB contains following three parts of data:

- An identifier Tid;

- A collection of quasi identifier (TQI) attributes at n different but typically consecutive time instants, denoted by TQI = ($At_1, At_2, At_3 \ldots At_n$)
- A collection of sensitive attributes which are denoted by At_s as a entirety ensures that all the TQI attributes values of each record in the published TDB* (Anonymized Times series Data base) namely TR_1 …..TR_n are identical to at least k-1 other records.

3.3 Time series Patterns

Time series patterns are defined as, x is a feature vector of y correlation function of respective time series data and derived as follows

$$p(x) = <cf1, cf2, \ldots, cfy> \quad (1)$$

Here y is the system factor. Every two time series pattern seems to be similar, if and only if their corresponding feature vectors are similar.

3.4 Linkage Threats on Time series Data

Accordingly, the possible linkage threats imposed on time series data if the adversary have a background knowledge on time series quasi identifiers TQI and patterns P generated on those time series data

(i) Threats based on linkage of data values of the TQI termed as T_V

(ii) Threats based on linkage of patterns of entire TQI or a portion of them, termed as T_P or

(iii) Threats based on linkage of both values and patterns, termed as $T_V \cup T_P$

4. PROPOSED WORK

4.1 Basic Definitions

4.1.1. (k, P^d) Anonymity

Let TDB be a Time series DataBase, and At_1, At2, ….., At_m being its TQI (Time series Quasi Identifiers). A Published TDB* is formed, if TDB* meets the following two requisites,

- k- requisite: Each Anonymity cluster AC = (tr_1, tr_2,…, tr_n) views to be at least k times in the entire database TDB*.
- P^d – requisite: Every k-set S of time series pertaining the similar anonymization cluster, if any time series record Tr € S, there would be at least P^d – 1 other time series in S pertaining the same TQI pattern representation as PR [Tr].

  * The granularity of pattern representation in our novel method is based on d value, Which termed to be the degree of the polynomial.

4.1.2 Utility Indicators

Here, we notify the utility metrics used in our *(k,P^d)* Anonymity framework, (i) Information Loss and ii) Pattern Loss which are derived to justify the utility of our novel *(k,P^d)* Anonymity framework.

Information Loss: In order improve the ultimate interest of the published database, it is essential to reduce the information loss occurs due to anonymization, we measure the information loss based on the anonymization cluster of each set.

Let time series Q belonging TQ of Set S, the anonymization cluster has a lower bound (tr_1^-, tr_2^-, tr_3^-…..

tr_n^-) and an upper bound (tr_1^+, tr_2^+, tr_3^+…..tr_n^+). The information loss of Q is derived by the following equation 2

$$IL(Q) = \sqrt{\sum_{i=1}^{n}(tr^+ - tr^-)^{2/n}} \tag{2}$$

The computation of information loss for the anonymized time series database TDB* is attained by summing up the information loss of each time series records Tr and derived using the following equation 3

$$IL(TDB^*) = \sum_{i=1}^{n} IL(TQI_i) \tag{3}$$

Pattern Loss: Let P*(x)* be the original feature vector that represents the patterns generated on original micro data and $P^{\#}(x)$ be the patterns generated conforming the P^d-requisite on P*(x)*. Thus the pattern loss is measured by evaluating the distance between P*(x)* and $P^{\#}(x)$ by integrating [12] as mentioned in the below equation 4,

$$PL(TQI) = ASSM(P(TQI), P^{\#}(TQI)) \tag{4}$$

The computation of pattern loss for the anonymized time series TDB* is attained by summing up the pattern loss incurred on each time series record Tr and derived using the equation 5 mentioned below,

$$PL(TDB^*) = \sum_{i=1}^{n}(PL(TQI_i) \tag{5}$$

4.2 Objective

(i) Anonymizing time series data values in compliance with *k*-requisite.

(ii) Computing virtuous pattern representation that accompanies fine granularity in pattern representation and as well as in compliance with P^d-requisite.

(iii) Improve the quality in extracting the distance between the original P*(TQI)* and anonymized $P^{\#}(TQI)$ using [12] ASSM(Angular metric for Shape Similarity) distance function.

(iv) Preventing the threats, T_V reside with data values of TQI and T_P reside with generated patterns of TQI.

(v) Examining the Information Loss $IL(TDB^*)$ and Pattern Loss $PL(TDB^*)$ associated in implementing our (k,P^d) anonymity model.

(vi) Publishing the anonymized Time Series DataBase (TDB^*), which intensely withstands the privacy attacks with respect to time series data values and patterns generated on those data values ,accompanying reduced pattern loss comparatively with the existing approaches.

4.3 System Architecture

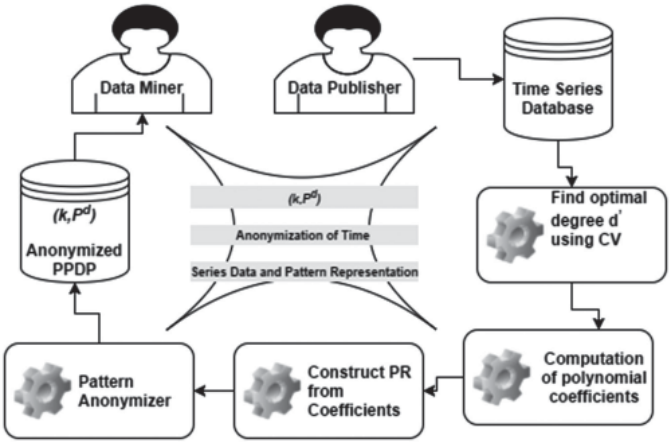

Figure 1. System Architecture of (k,P^d) Anonymity Model

The system architecture of novel framework *(k,P^d)* anonymity model encompasses five important components

(a) K-anonymization

(b) Cross validation

(c) Polynomial co-efficient computation and Generating pattern symbols

(d) Pattern Anonymizer (PA)

(e) (k,P^d) anonymized PPDP(Privacy Preserving Data Publishing)

The necessitate of above mentioned components are discussed below

(a) k-anonymization: This component is entirely performs

(i) Anonymize the data values i.e. Creating k-sets comprising anonymization clusters (AC).

(ii) Prevents the published datasets from T_V threats.

(b) Cross Validation: This component attempts to find out the optimal d' through which the patterns of original time series are derived. Aids in deriving patterns of finer quality and in the process evaluating the pattern loss associated generalized patterns.

(c) Polynomial co-efficient computation and Generating pattern symbols: This component performs

(i) Creation of quality patterns with fine granularity

(ii) Fine granular pattern progresses in reducing the pattern loss incurred due to pattern anonymization process.

(d) Pattern Anonymizer: This component pull-outs the pattern representation i.e. PRs confirming the P^d-requisite.

(i) Creates anonymized patterns i.e. P-Set

(ii) Prevents the published datasets from T_P threats.

(e) (k,P^d) anonymized PPDP(Privacy Preserving Data Publishing)

This component publishes the time series database in anonymized form TDB^* conforming *(k,P^d)* anonymity model.

4.4 Methodologies to Enforce (k,P^d)Anonymity Model

Here, we discuss the methodologies that we carried out to enforce *(k,P^d)* Anonymity model, our objective is to mininmize the information loss and pattern loss which confirming the constraints principled by *k* and P^d requisites. Our novel methodology have been named as (*k,P^d)* Anonymity model , *k*-stands to denote k- anonymization adoption in anonymizing the data values. P^d stands to denote Pattern Anonymizer methodology proposed and used in anonymizing the finer patterns, especially *d*- denotes the degree of the polynomial co-efficient used to generalize the pattern confirming P^d that computed through polynomial method.

The solution to enforce (k,P^d) anonymity model is exercised in top-down approach as follows

- Apply k-anonymization strategy and create k-Sets anonymized table TDB^* that adheres *k*-requisite.
- Pull out PRs from the anonymized table, by adhering on the selected PR forms for all k-Set individually. The pulled PRs, should minimize the pattern loss while confirming the P^d- requisite with own k-Sets.

- For all k-set individually make P^d subset based on the extracted PRs.

4.4.1. K-anonymization of time series

Here we, apply k-anonymization to original time series micro data shown in Table 1, which initiates from de-identifying all the time series record and the TQI are generalized to prevent linkage attack. Explicitly each TQI are generalized to the value that fall between [].The traditional k-anonymization confirms that each records in the published table is identical to k-1 other records of the table. Through which our anonymization process creates anonymization clusters AC on the time series quasi identifiers (TQI) attributes as shown in Table 2. Each cluster contains at least k original records that are bounded by sequence of [] lower bound and sequence of [] upper bound, time series records belongs to same cluster called anonymity cluster (AC).

Table 1.
Original Micro Data

Id	*Y2008*	*Y2009*	*Y2010*	*Y2011*	*Y2012*	*Y2013*	*Y2014*
1	170	175	188	197	213	221	200
2	145	157	165	177	204	196	180
3	176	181	147	134	125	112	160
4	98	120	125	132	151	161	110
5	117	107	87	74	51	56	85
6	32	54	59	67	96	101	90
7	88	93	56	43	20	25	55
8	71	63	47	38	43	20	46
9	500	600	300	100	500	1000	600
10	200	800	600	500	200	600	400

Table 2.
Anonymized Original Micro Data

Id	*Y2008*	*Y2009*	*Y2010*	*Y2011*	*Y2012*	*Y2013*	*Y2014*	*Anonymity Clusters*
1	145-176	157-181	147-188	134-197	125-213	112-221	200	1
2	145-176	157-181	147-188	134-197	125-213	112-221	180	1
3	145-176	157-181	147-188	134-197	125-213	112-221	160	1
4	32-117	54-120	59-125	67-132	51-151	56-161	110	2
5	32-117	54-120	59-125	67-132	51-151	56-161	85	2
6	32-117	54-120	59-125	67-132	51-151	56-161	90	2
7	71-500	63-600	47-300	38-100	20-500	20-1000	55	3
8	71-500	63-600	47-300	38-100	20-500	20-1000	46	3
9	17-500	63-600	47-300	38-100	20-500	20-1000	600	3
10	200-200	800-800	600-600	500-500	200-200	600-600	400	4

4.4.2. Computing Pattern representation (PRs)

Principles: Our work strives to discover the patterns in the time series data by calculating the polynomial .The polynomials remains a remarkable methodology in discovering the patterns when compared with constant and linear models, because they observe the information about the curvature of the sub- time series data too. The polynomials resolves the over fitting problem arises during approximation process.

After deriving the polynomials from the given dataset, we utilize the polynomial co-efficient in calculating the frequencies of the patterns. Then these polynomial co-efficient are transformed into symbolic words.

Cross Validating: In order to quantify the effects of bias and variance and construct the best possible estimator, we will split our training data into three parts: a *training set*, a *cross-validation set*, and a *test set*. As a general rule, the training set should be about 60% of the samples, and the cross-validation and test sets should be about 20% each. The general idea is as follows. The model parameters (the coefficients of the polynomials) are learned using the training set as above. The error is evaluated on the cross-validation set, and the Meta parameters (the degree of the polynomial) are adjusted so that this cross-validation error is minimized. Finally, the labels are predicted for the test set.

Algorithm1: To find the optimal degree of Polynomial (Cross Validation Approach)

Precondition: Data set of size n, Polynomial degree d and polynomial coefficients Φ

Post condition: Optimal polynomial degree *d*'

Split the data set of size n into three parts training set *m* (60%), cross validation set *mcv*(20%), test set *m*test (20%)

{Compute Training Error TRE (Φ)}

Step 1: for $j \square \{1, \ldots\ldots, d\}$ do

Step 2: $h_{\Phi}(x^{d}) = \Phi_{0} +$

Step 3: for $i \square \{1,\ldots, m\}$ do

Step 4: $\mathrm{TRE}(\Phi) = (x^{(i)}) - y^{(i)}))^2$

Step 5: end for

Step 6: {Compute Cross validation Error CVE (Φ)}

Step 7: for k $\square$ {1, ……., *mcv*} do

Step 8: $\mathrm{CVE}(\Phi) = (x^{(k)})) - y^{(k)}))^2$

Step 9: end for

{Compute Test Error TE (Φ)}

Step 10: for l $\square$ {1,.., mtest} do

Step 11: $\mathrm{TE}(\Phi) = (x^{(l)})) - y^{(l)}))^2$

Step 12: end for

Step 13: end for

Step 14: Plot (X,Y) = (Degree of Polynomial *d*, {Cross validation error CVE(Φ), Training Error TRE(Φ)})

Step 15: d' → Choose corresponding degree which has minimum CVE(Φ)

Step 16: return *d*' {optimal value of polynomial degree}

Polynomial fitting: Our work operates throughout the normalized time series data and calculates the polynomial co-efficient by accounting all-time series data. Algorithm 2 is used to compute co-efficient of a polynomial, as shown in the table 3, regression is exercised by reducing the least square errors that arises among the polynomial estimates and real values of the time series database. The polynomial is realized using the below mentioned equation 6

$$\mathrm{Pl}(\mathrm{TY},\widehat{\mathrm{TY}}) = \left\|\mathrm{TY} - tZ\beta\right\|^2 \tag{6}$$

$$\mathrm{TY}: = [\mathrm{At}_1, \mathrm{At}_2, \ldots\ldots\ldots, \mathrm{Atm}]$$

The time indexes predictors are transformed to linear regression form by apply vandermonde matrix $tZ \in R^{MxN}$ as shown in equation (7)

$$tZ = \begin{pmatrix} 0^0 & 01 & \dots\ \dots\ \dots\ \dots & 0^{d'} \\ 1^0 & 11 & \dots\ \dots\ \dots & 1^{d'}. \\ (m-2)^0 & (m-2)^1 & \dots\ \dots & (m-2)^{d'} \\ (m-1)^0 & (m-1)1 & \dots\ \dots & (m-1)^{d'} \end{pmatrix} \quad (7)$$

The least square system is manipualted by determining the initial derivative with respect to polynomial co- efficient β as shown in the equation 8 below

$$\frac{\eta Pl(TY, \widehat{TY})}{\eta\beta} = 0 \text{ Leads to } \beta = (tZ^T, tZ)^{-1} tZ^T Pl \quad (8)$$

Algorithm 2: Computation of Polynomial Co-effcient

Precondition: Time series Dataset TDB $\in R^{mxn,}$ optimal polynomial degree *d*'

Post Condition: Polynomial coefficients set $\Phi \in R^{mxnx\,(d'+1)}$.

Step 1: Derive the projection matrix Q

$$Q = (tZ^T, tZ)^{-1}\, tZ^T$$

Step 2: Intialize the polynomial coefficient set Φ to empty

$$\Phi^{(i)} = \varnothing$$

Step 3: Obtain the polynomial coefficient for all-time series attributes as

$$\text{for } At_i = (1,\dots..,m)$$
$$TY = [At_1, At_2,\dots\dots\dots,At_m$$
$$\beta = QTY$$

Step 4: Add the derived coefficient to the polynomial set Φ

$$\Phi^i = \Phi^i \cup \{\beta\}$$

end for

end for

Step 5: return all the computed coefficients

Transforming polynomial coefficients to symbolic words: Here we strive to transform the calculated polynomials coefficient derived by applying algorithm 2 into symbolic words. The intension of the transformation keens in mapping the each derived coefficients to a symbols, which frames a words of length d'+1 symbols often called as original pattern representation i.e. PRs of exact granularity by applying the below mentioned algorithm 3. Each β values of the polynomials coefficients will undergo an equi-area discretization process to equally represent the area of similar time series promptly. The inceptions among the diverse areas are named as primary points. A histogram is formed based on the polynomials coefficents of time series data and divided into regions of equivalent areas by adopting the alphabet size (δ). In order to divide the histogram in δ regions, sorting of the coefficients are done and primary points are chosen by considering the values fall under multiples of, sorting of coefficients is performed by applying below mentioned equation (9)

$$B: = \text{sort}(\{\beta \in \Phi^i, i = 1,\dots\dots, m\}) \quad (9)$$

and S = |B| denotes the sorted coefficient list's size.

The transformation of coefficients to symbols i.e. words consists of two phases,

(i) Primary point computation

$$\eta_k = B[S^k/_\delta], \text{ where } \square_k \in (1\ldots \delta - 1) \text{ and } = \infty$$

(ii) Symbol Conversion

(a) Computes the primary points δ to separate the distribution of each adopting the equi-model

(b) The second phase exercise all the β coefficients of time series data and transforms each specific coefficient to a character C, based on the β values position with respect to the primary points.

Algorithm 3: Symbolic Words Generation (SWG)

Preconditions: Polynomial coefficients, Size of the alphabet δ.

Post conditions: $W \in R^{m*n(d'+1)}$

Step 1: Sort all the derived polynomial coefficients from applying algorithm 1

$$B: = \text{sort}(\{\beta \in \Phi^i, i = 1,\ldots \ldots, m\})$$

Step 2: Calculate the size of the sorted polynomial coefficient

$$S = |B|$$

Step 3: Assign primary points to be infinity in case utilizing entire alphabet size

$$\delta = \infty$$

Step 4: Calculate the size of the primary points underlying alphabet size (i.e.) number of characters to Symbolize,

$$\text{for } k = 1,\ldots \ldots, \delta{-}1) \text{ do}$$

$$\delta_k = B[S^k/_\delta]$$

end for.

Step 5. Symbol conversion

Formulate all alphabets, real values and obtained matrix values.

$$\Sigma = \{A, B, C\ldots TY, tZ)$$

Step 6. For all attribute values of time series, by concerning the derived polynomial coefficient, map them to symbolize words with respect to the primary points.

We formulate our novel process of pattern representation as R-SymPol .

Advantages of R-Sympol - Revised Symbolic Polynomial:

- ✓ Classification accuracy
- ✓ Reduced error rate
- ✓ Optimal Running time
- ✓ R-SymPol approximates polynomials of arbitrary degrees, instead of simple average
- ✓ SAX words are built from locally constant approximation, which are less expressive than the polynomials

- ✓ SAX method averages the content and loses information about their curvature thus results representing different time series curvature in same SAX word.

Pattern Anonymizer

We proposed a novel approach to exercise (*k*, P*d*) anonymity in two phases (i) applying k-anonymity on original dataset to obtain k- set or Anonymity Cluster (ii) successively, a generate tree procedure is applied on each k-set or Anonymity cluster to frame each anonymity cluster as tree and by applying the pattern anonymizer algorithm 4, we divide the nodes representing the anonymity cluster or k-set into Pd- subset conforming P-requisite. Degree d is considered as a generalizing parameter, for each time series increase in the degree but not beyond d' decreases the pattern loss. Since d' is the optimal degree derived through cross validation algorithm 1, to compute the patterns of exact representation exact PRs from the original time series data.

Generate Tree Configure Settings: Set the entire k-set S to root, which will utilized by the pattern anonymized to generate the patterns conforming p-requisite, Set R-SymPol's degree to 1, which represents that all-time series data have same symbolic representation i.e. termed to be coarser granularity. We notify that the highest granularity appropriate for pattern anonymization is MAX_D which is always < d'. The tree representation have five attributes to be mentioned

- Degree : denotes the current degree of R-SymPol of N
- PR: denotes the R-SymPol's PR of the node with degree d
- Elements : denotes the time series data contained in N, all have similar R-SymPol PR as N.PR
- Num_Node: denotes the number of time series data existing in the Node N.
- Tag: denotes the tag entitled on the node based on the requisite, there are 3 tags used in the process, False_leaf: indicates the one with its Num_Node <d, True_leaf indicates one with its Num_Node ≥ d and Moderate_leaf indicates that N is not leaf node.

Anonymization of Patterns: The pattern anonymization is done by dividing nodes, which attempts to scrutinize the PRs of the records in a node. All elements of the tree node should have similar PRs N.PR and tier N.tier. Accordingly, an increment in the tier, will trim its members to have different PR under N.tier+1.

Initializing from the root, Pattern anonymization handle node N in recursive procedure as follows

- ✓ If Pt_n, Num_Node < d, then the node will be tagged as False_leaf.
- ✓ If Pt_n.tier = MAX_D, then the node will be tagged as True_leaf and the recursion terminates
- ✓ Else if d ≤ Pt_n.Num_Node < 2*d, then try to maximize the tier of the node as long as all-time series records have similar PRs.

Algorithm 4. Pattern Anonymizer

Pre-condition: P_tree node Pt_n, d degree of the polynomial, MAX_D

Post condition: P subset PRs conforming the P-requisite

Step 1: Begin check with the degree d,

Step 2: If Pt_n.Num_Node < d then

Step 3: Pt_n.Tag = False_leaf

Step 4: Check the node level with the maximum attainable level

Step 5: If Pt_n.tier = = MAX_D then

Step 6: Pt_n.Tag = True_leaf

Step 7: Recognising direct True_leaf to proceed without node division

Step 8: If $d \leq Pt_n$.Num_Node < 2 * d then

Step 9: Pt_n.Tag = True_leaf

Step 10: Maximize the Ptn.tier without pattern anonymization

Step 11: Else

Step 12: Recycle the False_leaf to True_leafs

Step 13: If Ptn can be false splitter then

Step 14: If entire number of all TL_nodes ≥ d then

Step 15: Form child $_{unify}$.tier = Pt_n.tier

Step 16: Tier of all FL.nodes is Pt_n.tier + 1

Step 17: Else

Step 18: Tier of the entire child node is

Step 19: Pt_n.tier + 1

Step 20: Else

Step 21: Pt_n.Tag = True_leaf.

Step 22: End

5. EXPERIMENTAL RESULTS

To evaluate efficiency and competence of proposed framework *(k,P^d)* Anonymity model, we exercised wide experimental analysis on both real and generated dataset. The algorithms comprised under our framework have been developed and executed in Dot (.)Net platform, all our experiments were exercised on PC containing Intel Core i5 processor, 8GB Ram running on windows 8.1 operating system.

Experimental Configuration Settings

Data sets: We employed our framework over two publicly available UCR[20] benchmark time series datasets, namely (i) ECG 200 (ii) GDP and one generated dataset named as (iii) SynIncome

Table 3.
Data Set Properties

Dataset	*Number of Time series Records*	*Length*
ECG 200	106	97
GDP	25	71
SynIncome	125	50

These datasets were comfortably utilized in many experimental time series anonymization research works, all of these datasets last attribute will be considered as sensitive attributes At_s and we apply the anonymization process with respect values and patterns to remaining TQI i.e. Time series quasi identifiers.

The sensitive attribute consideration with respect to all involved data sets as follows

- ECG data set 97th attribute have been locked as sensitive attribute Ats.
- GDP dataset 71st attribute have been locked as sensitive attribute Ats.
- SynIncome data set 50^{th} attribute have been locked as sensitive attribute At_s.

We conducted the experimentation in four perspectives

(i) Regulating *k* value

(ii) Regulating P^d value.

(iii) Regulating *k* and P^d to analyse the optimized information loss and pattern loss

(iv) Utility comparison of the published data with our proposed model and existing two approaches

Regulating k value: Here we experimented *(k,P^d)* anonymity model on the mentioned three datasets to show how Information Loss IL (TDB*) varies with respect to the k value. Figure no.2, reveals that increasing values of k increases the information loss, which seems to be similar for our proposed method *(k,P^d)* anonymity model, our previous work Dμ-SAX with Knop Ripping and (k,P) anonymity model., since all these three proposals follow same strategy in applying k-anonymization.

Through this evaluation we contented to adhere the traditional proclamation of k-anonymization (i.e.) *k* is directly proportional to Information Loss,

$$\blacktriangleright\ k \ \alpha\ \text{IL}\ (\text{TDB}^*)$$

From the figure 2, clearly pictures that ECG 200 dataset accounts for higher information loss comparatively than the other two dataset. From analysing the results and the dataset properties we state huge number of time series with long sequence, having big variation between upper bound and lower bound values accounts higher information loss.

By varying the values of k as 2,4,6,8,10,12,14,16,18, and 20, we calculated the information loss applying equation 2 and equation 3 , moreover we exercised the *k* – anonymization process sticking on the constraint k<<*n*, *n*-denotes the number of records in original table.

Regulating P^d value: We executed Pattern Anonymizer algorithm on each k set that are obtained for different values of k to show how pattern loss varies with respect to P^d value. Figure 3, figure 4, and figure 5 reveals that increase in P^d will lead increase in pattern loss. Again through this evaluation we state that P^d is directly proportional to Pattern Loss.

$$\blacktriangleright\ P^d\ \alpha\ \text{PL}\ (\text{TDB}^*).$$

From Figure 3, 4, and 5 , shows that by applying our *(k,P^d)* anonymity model, GDP data had gained more pattern loss than the other two datasets, By varying the values of P^d as 2,4,6,8,10,12,14,16,18 and 20, we generated the pattern representation PRs using R-SymPol and Pattern Anonymizer PA conforming the P-requisite constraint. Then we computed the pattern loss happened due to pattern representation anonymization by incorporating AMSS distance function using the equation 4 and equation 5.While varying the P^d value = 10, 12 and 14 according to *(k, P^d)* anonymity model, shows in that particular instance, our model doesn't adapt to the statement $P^d\ \alpha$ PL (TDB*).

Regulating k value and P^d value: Here we attempt to analyse the optimal k and P^d value that are advisable to support our objective and claim. Thus we computed the pattern loss occurred in due to the P^d value for each k- set shown in Figure 3, 4 and 5.

From the figure 3, 4 and 5, gives us a clear picture that for k and P^d >14 progresses according to the statements mentioned below,

- $k \alpha$ information loss
- $P^d \alpha$ Pattern Loss

Whereas for k and P^d values 8, 10, and 12 shown the results which contradicts the above mentioned proclamation.

Utility Comparisons of the published data of our proposed approach and the existing two models: Here for anonymizing the data values of the time series data we adopted the strategy followed in (k, P) anonymity model and Dμ-SAX with Knob Ripping. Thus our proposed scheme reveals the same utility measure in terms of data values anonymization (i.e.) for k values 2 – 20 on all the three mentioned data set with respect to the previous schemes. Since our claim and objective concentrates on reducing pattern loss. The generated PRs i.e. Pattern representation using R-SymPol are of fine granular than the patterns generated through other two mentioned schemes. The Pattern Anonymizer introduced in our framework is capable of generalizing the PRs obtained through R-SymPol conforming P- requisite ,Figure 6 reveals our *(k,P^d)* Anonymity model framework can able to anonymize data values and pattern representations with reduced pattern loss than the other two existing approaches. However our experiments exposes, the varying k value factor doesn't have impact on the pattern loss, the pattern loss is only determined by the P^d value not by the k value. In some extent, our pattern anonymization process our approach give way to have distinct time series, that lead us to perform recycling the false leaves as declared. Certainly while applying our approach towards the ECG200 time series data set we were not able the recycle the false leaves obtained distinct time series data on account of pattern anonymization conforming P- requisite , such distinct time series are suppressed in order to maintain the pattern anonymization conforming P-requisite which in turn increases the pattern loss. However these are we considered to be negligible circumstances that happens to be in every proposed experimentations. Comparative results figure no of all the three approaches over three datasets articulates our proposed *(k,P^d)* Anonymity model outperforms than Dμ-SAX with Knob Ripping and (k, P) anonymity model. Dμ-SAX with Knob Ripping shows better reduced pattern loss than (k, P) anonymity model. Specifically it appears (k, P) anonymity model delivers reduced pattern loss for values P^d = 6, 10 and 12 than Dμ-SAX with Knob Ripping for SynIncome dataset. For GDP dataset, all the three approaches strives to give reduced pattern loss as per their order of significance i.e. *(k,P^d)* Anonymity model, Dμ-SAX with Knob Ripping and (k, P) anonymity model with slight variations. Thus through this vigorous experimentation of our proposed *(k,P^d)* Anonymity model on two bench mark datasets and one generated dataset, we had shown through our results that *(k,P^d)* Anonymity model is accomplished method of anonymizing data values and pattern representation PRs beneath Privacy Preserving Time series Data mining and proved its efficiency through generating patterns of finer granularity than the other approaches and generalizing the PRs with reduced rate of pattern loss.

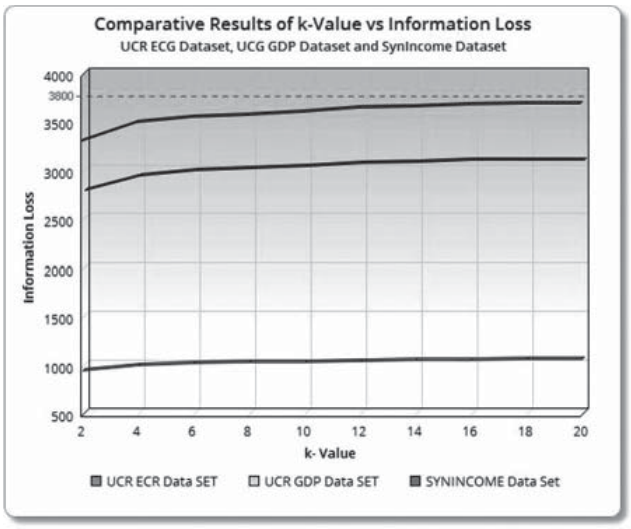

Figure 2. Comparative Results of k-Value vs Information Loss – UCR ECG 200, UCR GDP and SynIncome Dataset

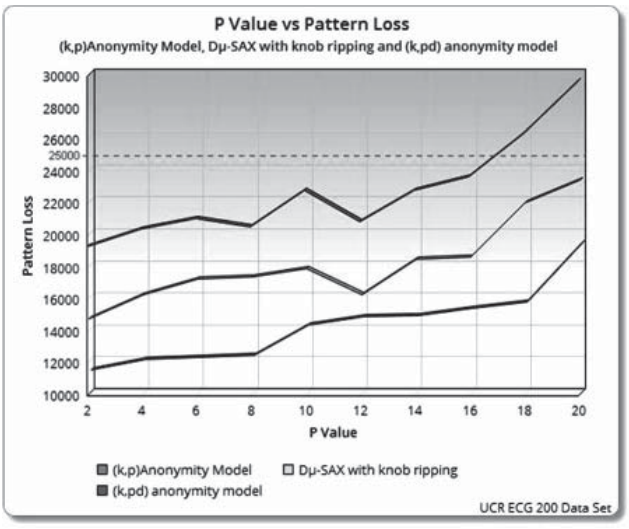

Figure 3. Comparative Results of P-Value vs Pattern Loss -(k,p)Anonymity Model, Dμ-SAX with knob ripping and (k,pd) anonymity model for UCR ECG 200 Data Set

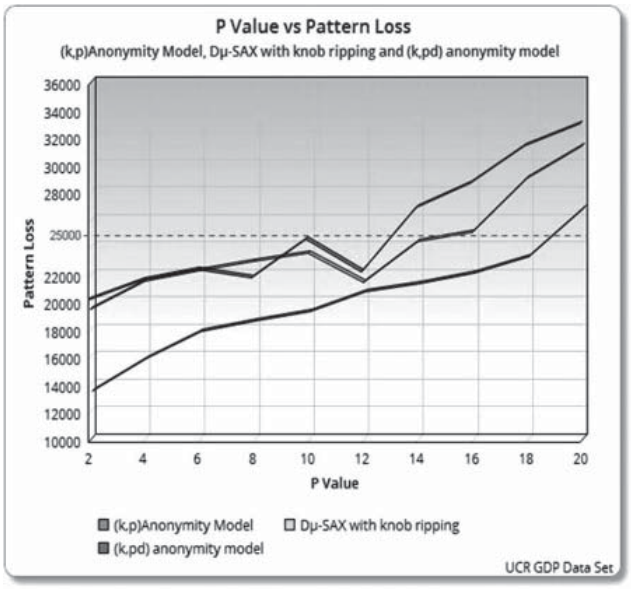

Figure 4 Comparative Results of P-Value vs Pattern Loss -(k,p)Anonymity Model, Dμ-SAX with knob ripping and (k,pd) anonymity model for SynIncome Data Set

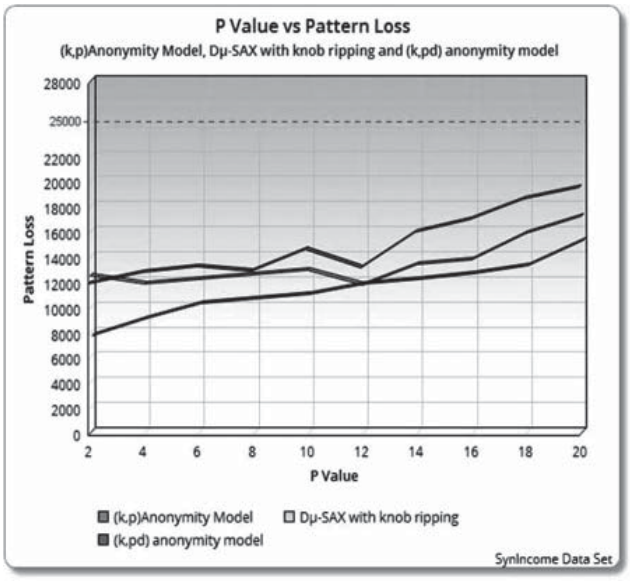

Figure 5 Comparative Results of P-Value vs Pattern Loss -(k,p)Anonymity Model, Dμ-SAX with knob ripping and (k,pd) anonymity model for UCR GDP Data Set

6. CONCLUSION

We introduced and implemented an innovative framework to anonymize i) time series data values and ii) symbolic patterns generated on those data. Our framework confirms to generate fine granular patterns (goodness of fit) and anonymize patterns with minimal pattern loss, our model could produce the anonymized time series databases that withstands value and pattern linkage attacks. The intense experiments determines the effectiveness of our (k,P^d) Anonymity model by generating unrestricted pattern representation and acquiring minimal pattern loss overwhelmingly with i) (k,P) anonymity model and reasonably than Dμ-SAX with Knob Ripping.

Thus this proposed solution (k,P^d) Anonymity to have few future directions a) To analyse the compatibility of our model with data mining tasks like classification and clustering, b) Try with a different mishmash of k-anonymity, representation methodology and pattern anonymization that yields better pattern representation and minimalistic in terms of information loss and pattern loss.

References

1. P. Samarati, "Protecting Respondents' Identities in Microdata Release," IEEE Trans. Knowledge Data Eng., vol. 13, no. 6, pp. 1010-1027, Nov./Dec. 2001.
2. L. Sweeney, "k-Anonymity: Privacy Protection Using Generalization and Suppression," Int'l J. Uncertainty Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 571-588, 2002.
3. C.C. Aggarwal and P.S. Yu, "A Condensation Approach to Privacy Preserving Data Mining," Proc. Ninth Int'l Conf. Extending Database Technology (EDBT), pp. 183-199, 2004.
4. A.Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam,"l-Diversity: Privacy Beyond k-Anonymity," Proc. 22nd Int'lConf. Data Eng. (ICDE), p. 24, 2006.
5. N. Li, T. Li, and S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), pp. 106-115, 2007.
6. J.Xu et al., "Utility-Based Anonymization for Privacy Preservation with Less Information Loss," SIGKDD Explorations, vol. 8, no. 2,pp. 21-30, 2006.
7. M.E. Nergiz, M. Atzori, and Y. Saygin, "Perturbation-Driven Anonymization of Trajectories," Technical Report 2007-TR-017, ISTI-CNR, 2007.
8. S.Papadimitriou, F. Li, G. Kollios, and P.S. Yu, "Time Series Compressibility and Privacy," Proc. 33rd Int'l Conf. Very Large DataBases (VLDB), pp. 459-470, 2007.
9. N. Mohammed, B.C.M. Fung, and M. Debbabi, "Walking in the Crowd: Anonymizing Trajectory Data for Pattern Analysis," Proc.18th ACM Conf. Information and Knowledge Management (CIKM),pp. 1441-1444, 2009.
10. R.G.Pensa, A. Monreale, F. Pinelli, and D. Pedreschi, "Pattern-Preserving k-Anonymization of Sequences and its Application to Mobility Data Mining," Proc. Int'l Workshop Privacy in Location-Based Applications (PiLBA), 2008.
11. Lidan shou, Xuan Shang, Ke Chen, Gang Chen an Chao zhang"Supporting pattern preserving Anonymization for time series data", IEEE Transactions On Knowledge and Data Engineering, Vol. 25, No. 4, April 2013.
12. J. Lin, E.J. Keogh, S. Lonardi, and B.Y. chi Chiu, "A Symbolic Representation of Time Series, with Implications for Streaming Algorithms," Proc. Eighth ACM SIGMOD workshop Research Issues in Data Mining and Knowledge Discovery (DMKD), pp. 2-11, 2003.
13. J.S.AdelineJohnsana, A.Rajesh, S.KishoreVerma, "An Enhanced PPDM framework for preserving patterns in time series data", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 17 (2015).
14. Simon Malinowski, Thomas Guyet, Rene Quiniou and Romain Tavenard "1d-SAX: a Novel Symbolic Representation for Time Series", 2013
15. Josif Grabocka, Martin Wistuba, and Lars Schmidt- Thieme "Scalable Classification of Repetitive Time Series through Frequencies of Local Polynomials ", IEEE Transactions on Knowledge and Data Engineering, Vol. 27, no. 6, June 2015
16. Tetsuya Nakamura, Keishi Taki ,Hiroki Nomiya , Kazuhiro Seki and Kuniaki Uehara, "A Shape-based Similarity Measure for Time Series Data with Ensemble Learning"@ Springer- Veilag London limited 2012.
17. J. Nin and V. Torra, "Towards the Evaluation of Time Series Protection Methods," Information Sciences, vol. 179, no. 11, pp. 1663-1677, 2009.
18. Tiancheng Li, Ninghui Li, "Slicing: A New Approach for Privacy Preserving Data Publishing" IEEE Transactions on Knowledge and Data Engineering, Vol. 24, No. 3, March 2012.
19. Mohammad-Reza Zare-Mirakabad, Fatemeh Kaveh-Yazdy, Mohammad Tahmasebi," Privacy Preservation by *k*-anonymizing Ngrams of Time series" 2013, **IEEE Conference Publications.**
20. E.Keogh and T. Folias, "UCR Time Series Data Mining Archive,"http://www.cs.ucr.edu/ eamonn/TSDMA/, 2012.