IJREAT International Journal of Research in Engineering & Advanced Technology, Volume 2, Issue 1, Feb-Mar, 2014
ISSN: 2320 - 8791
www.ijreat.org

Tri Patternization On Generic Visualized Time Series Data

Beenu Ann Oommen[1], R.Aroul Canessane[2]

[1]*M.E Student, Computer Science & Engineering, Sathyabama University, Chennai.*

[2]*Associate Professor, Faculty of Computer Science & Engineering, Sathayabama University, Chennai.*

***Abstract* – Time series is a set of observed data that are measured at regular intervals of time.The privacy of the time series data can be preserved through generalization and anonymity techniques. It has been useful in various fields of financial, medical and weather analysis. The data whose privacy is to be preserved is first identified and then generalization is made. The data will be generalized based on grouping algorithm and then the data will be displayed in an approximation format. The data will then be displayed in graphical format, providing no clue about the original data. This can be implemented using K-P anonymity algorithm for time series data with rich patterns.For ensuring privacy preservation and to eliminate pattern loss, a detectable noise will be incorporated with the data along with all the necessary features. The noise can be incorporated using data fly algorithm.**

Index Terms: Privacy, time series, pattern, noise

I.INTRODUCTION

Time series data can be either discrete or continuous observations. If the observations are at spaced intervals of time, then it is called discrete and if the observations are made at every instant of time, it is called continuous time series data. If the observations are made continuous then the data should be displayed in the order they arose. Time series data includes identifier, quasi identifier and sensitive attributes. During anonymization, the identifier and quasi identifiers are de identified. The quasi identifiers will be generalized into a range of values. It thus forms a group of anonymization envelop. The records that belong to the same envelope are called anonymity group. The sensitive attributes will be retained in their real form and will be kept secret.

Anonymization of data is a process of removing or replacing the identity information from a record. While anonymizing data, values and patterns have to be retained in data base for supporting different queries and linkage attacks should be prevented. This can be achieved through the information hiding process. Information hiding approaches include perturbation based approach and partition based approach. Perturbation based approach protects data by adding noises.The noise that is being added should satisfy certain constraints and conditions. But it does not aim at preventing linkage attacks. In partition based approach, the whole data set will be divided into different groups and then generalization of the data will be made. The approaches in K anonymity implementation use this approach of partition. Though it aims at protecting linkage attacks, it cannot preserve patterns and will lead to information loss. In this proposed system, these two approaches are combined thus removing the draw backs of both approaches. The whole data set will be segregated and then generalized in a range of values. The whole data set will be shown in approximation formats, then will be displayed graphically leading to anonymization. Then a

dectatable or deformable noise will be added to the data, thereby avoiding linkage attacks and pattern loss and information loss.

II.RELATED WORK

Several tactics have been proposed in literature for ensuring privacy in the sensitive data.

Charu C Aggrwal and Philip S Yu[2] have proposed for the condensation approach where the whole data will be condensed to multiple groups of predefined size. For each group, a level of statistical information about different records will be maintained. This statistical information preserves the correlations among different dimensions. Earlier, the privacy of data was ensured by adopting a perturbation based approach and each dimension in the data would be treated independently. This ignored the correlations between the dimensions. In this system, instead of using such an algorithm, a method is used where the original data set will be mapped with anew anonymized data set. This anonymized data set will be closely related and matching to the characteristics of the original data.Once the condensed groups have been generated, any known data mining algorithm can be directly applied to retrieve the original data.

Adam Meyerson and Ryan Williams[3] have proposed two versions of K-anonymity for ensuring data privacy and integrity of information. One of the version is NP hard or suppression version where the data entries will be deleted from viewing and will be viewed in '*' form. The other version is the polynomial time where approximation ratio will be used. The value in the table will be made available in a range of values. , the concept of privacy preserving has been formulated to preserve the data without any clue even when multiple generalized techniques based on grouping algorithms have been introduced to filter the data which was revealed in many fields. Hence the strategy of approximation was formulated. Hence for anonymizing time series data conventional K-anonymity was enforced. In this visualization of data was done in statistical, graphical and hierarchical representations.

Spiros Papadimitriou, Feifei Li, George Kollios and Philip S Yu[4] proposed a method of making the perturbation similar to the original data for preserving the structure of data. If the perturbation does not have the same compressibility properties as the original data, it can be detected and filtered out.

Keogh, Chakrabarthi, Mehrotra and Pazzani[5] proposed for a dimensionality reduction mechanism on data by indexing the reduced data with a multidimensional index structure. This is done, otherwise similarity search is a difficult task in high dimensional data. This proposes for a APCA(Adaptive Piecewise Constant Approximation) where each time series data is approximated separately so that reconstruction errors are made minimum. Then the reduced data is indexed with a multidimensional index structure. It explains the superiority of APCA.

Lidan Shou, Xuan Shang, Ke Chen, Gang Chen and Chao Zhang[1] have proposed for a novel anonymization model called the K-P anonymity model for pattern rich time series data. Either of the two algorithms i.e., Naïve algorithm and KAPRA algorithms is implemented for this model. It supports for customized data publishing, in which the attribute values and patterns are published separately. It will thus prevent linkage attacks and will prevent pattern loss. But it is not possible to retrieve the original data from the anonymized data and it imposes a very restraint on pattern representation equality and this may lead to pattern loss.

Latanya Sweeney [11] has proposed for using preferred minimal generalization algorithm which combines the methods of suppression and generalization and brings minimum distortion.

Li Xiong[12] describes many anonymization strategies like local suppression, global attribute generalization, k-anonymization with suppression. Cost metric is used for finding the optimal anonymization. It also suggested out for different taxonomies of generalization algorithms.

Section I of this paper is a background introduction. Section II discusses a brief literature review. Section III describes the existing work. Section IV (1) Time series input, IV (2) Individual group clustering, IV (3) Group re weightage, IV (4) Graphical mapping integration, IV (5) Hierarchial time series classification, IV (6) Security evaluation metrics. Concluding remarks and scope for further research are given in section V.

III. EXISTING WORK

For anonymizing the time series data, conventional K-anonymity algorithm is used. It visualized the data in statistical, graphical or hierarchial formats. Multiple generalized techniques are used to filter the data. In this, strategy of approximation is formulated. It does not provide the end user to frame a clear idea about the data It leads to pattern loss and query accuracy is low.

IV.PROPOSED SYSTEM

The instantaneous values and inclusive patterns of time series data progression have to be retained in the available data as a great deal as probable impact over the various patterns of representation to support an assortment of queries. On the other hand, the association attacks based on acquaintance of values, patterns, or both have to be appropriate..

In general, privacy risks are critical quandary associated with confidential data of every organization, hence care has to be given to preserve the confidential data. Visualization of data by statistical, graphical and hierarchical representations is evolved to represent the data as in the existing system. In this, K-P anonymity algorithm is implemented on the time series data where a privacy constraint 'p' was proposed. It can be implemented as either Naïve algorithm, which is the extension of K-anonymity algorithm and uses a top-down approach or as KAPRA (K and P Reinforced Anonymity) algorithm, which uses a bottom-up approach of partitioning algorithm to split the whole data base. Along with this, by using the range values, interpretation of disturbance called noise are added. The algorithm called Data fly is to be used to add noisy data with range values in the time series, thus in turn allows the end-user to predict the actual time from the data visualization.

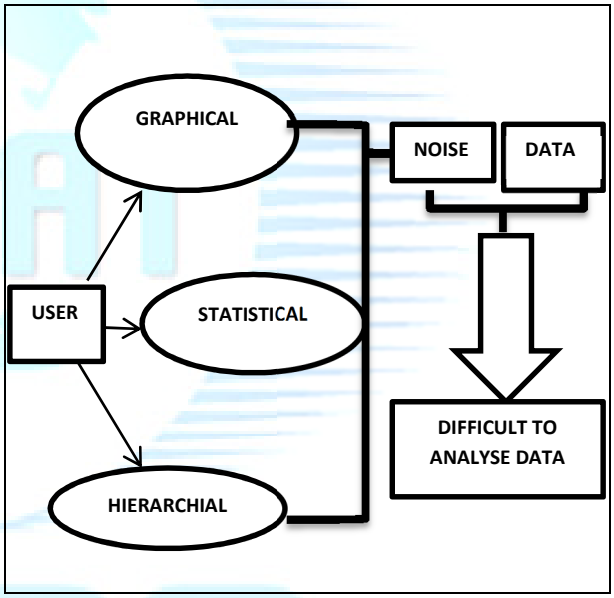

Fig 1 Functional Architecture

IV (1) TIME SERIES INPUT

Time series is a sequence of observations which are ordered in time. They are best displayed in a scatter plot. If the observations are made on some phenomenon, it is sensible to display the data in the order in which they arose, particularly since successive observations are probably be dependent. The input from the user is gathered from the user in the form of time series, in order to proceed with further proceedings.

IV (2) INDIVIDUAL GROUP CLUSTERING

Clustering is a technique used to identify different segments in a market. Data clustering is a technique in which the information that is logically similar is physically stored together. Clustering strategy can be used effectively when trying to memorize long lists of information. Similar set of input datas are grouped up using the clustering method individually.

IV (3) GROUP RE WEIGHTAGE

Grouped data from previous modules are revamped in the series as well as similarity measures between them are checked individually. Revamp are the methods to patch up or renovate; repair or reconstruct.

IV (4) GRAPHICAL MAPPING INTEGRATION

Mapping is a powerful data integration tool. The data will be visualized graphically. Added tuples are visualized graphically that makes the data unpredictable.

IV (5) HIERARCHIAL TIME SERIES CLASSIFICATION

Time series classification is to build a classification model based on labeled time series and then use the model to predict the label of unlabeled time series. The time series are represented as hierarchial with noise. The hierarchical partition is made by considering good and bad leaves segregation.

Data fly algorithm

Input : Private table PT, quasi identifier set QI, anonymity parameter K.

Output : MT= a K anonymization of PT.

1. Construct a frequency list of data containing unique values across QI.
2. With the most no of unique values, make the solution by generalizing the attribute.
3. Recalculate frequency list.
4. Suppress rows in MT occurring less than k times.
5. Enforce k requirement on suppressed rows in MT.

IV (6) SECURITY EVALUATION METRICS

The security evaluation, testing, protection profiling and risk assessment of information systems are processes in which evidence of assurance is analyzed against criteria for security functionality and assurance level. Security metrics are important indicators of how well security services are present in the information systems and can be used to measure the organization's security maturity level.

V. CONCLUSION

In this paper, we have described and proposed a work of fiction anonymity representation called (K P)- anonymity for time-series data. Relying on a nonspecific description to pattern representations, our reproduction may perhaps put off three types of connection attacks and in point of fact support the the majority extensively used queries on the anonymized data. We proposed a naive explanation and a more highly residential method called KAPRA to enforce K,P-anonymity on timeseries data. It provides for personalized data publishing and provided estimation methods to support queries on such data. The extensive experiments demonstrated the efficiency of K P-anonymity in resisting linkage attacks while preserving the outline information of time series.

REFERENCES

[1] Lidan Shau, Xuan Shang, Ke Chen, Gang Chen,Chao Zhang,"Supporting Pattern-Preserving Anonymization For Time Series Data",2013.

[2] C.C. Aggarwal and P.S. Yu, "A Condensation Approach to Privacy Preserving Data Mining," 2004.

[3] Adam Meyrson, Ryan Williams, "On the complexity Of Optimal K-Anonymity",2005.

[4] Spiros Papadimitriou, Feifei Li, George Kollios, Philip S Yu,"Time Series Compressibility and Privacy",2007.

[5] E.J. Keogh, K. Chakrabarti, M.J. Pazzani, and S. Mehrotra,"Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases," 2001.

[6] D. Gunopulos and G. Das, "Time Series Similarity Measures," 2000.

[7] R. Dewri, I. Ray, and D. Whitley, "On the Optimal Selection of k in the k-Anonymity Problem,"2008.

[8] O. Abul, M. Atzori, F. Bonchi, and F. Giannotti, "Hiding Sequences,"2007.

[9] CMU Graphics Lab Motion Capture Database, http://mocap.cs.cmu.edu/, 2012.

[10] E. Keogh and T. Folias, "UCR Time Series Data Mining Archive",2012.

[11] Latanya Sweeney, "Achieving K-Anonymity Privacy Protection Using Generalization And Suppression",2002.
[12]LiXiong,"DataAnonymization and Generalization Algorithms",2012.

[13] L. Sweeney ,"Datafly:A System For providing Anonymity In Medical Data",1998.