International Journal of Natural and Engineering Sciences 10 (3): 15-17, 2016
ISSN: 1307-1149, E-ISSN: 2146-0086, www.nobel.gen.tr

Hybrid Methodology of K-Annomity and Randomization for Privacy Preserving in Data Mining

Shaista CHIRAGH* Nayyar IQBAL
Department of Computer Science, Faculty of Sciences, University of Agriculture, Faisalabad, Pakistan

*Corresponding Author:
E-mail:shaistachiragh@gmail.com

Received: September 14, 2016
Accepted: December 06, 2016

Abstract

The purpose of this research is to transform the data in original form that is in unsecured form while doing the mining procedure and protect different attacks that involve in privacy theft and work on preserving privacy in a better way. In recent years, quick developments in the field of privacy preserving monitors because of the progress in the skill to store data. Privacy preserving has become increasingly more general because it allows sharing of privacy sensitive data for investigation purpose. People today have become aware of the privacy conflicts of their sensitive data and are very doubtful to share data. Current techniques are different among each other with respect to number of conditions such as, data quality, privacy level and performance. These techniques face different types of difficulties like, homogeneity and background knowledge concerns. The main problem is the misuse of data and cause of data misuse is mining procedure. In this research work a hybrid methodology of k-anonymity and randomization was used for different privacy protective difficulties. The recommended approach secures individual information with no loss of data which makes ease of use of information.

Keywords: Privacy Preserving; Data Mining; Randomization; K-Anonymity

INTRODUCTION

Various organizations, for example, banks, saving money organizations and health related facilities provider store massive size of information concerning about points of interest. The data is more utilized by the information excavator for inspection reason which helps the associations for accomplishment of important data. This data may hold secret information of any people, for instance, associations, banks contain account information or record of the individual [1].

In this specific strategy, the information or data extractor might have the capacity to get sensitive material and in this method an outside database may attempt to accomplish individual information so secrecy is turned into an essential issue for any organization or individual that contain secret information or data, through data mining procedure control this issue and a resourceful way has been seemed known as privacy preserving [2].

Data mining is a method which is useful to mine information in knowledge detection procedure from large set of record. Several areas have one of the most important data set, where this procedure is use for extracting convenient data. There are several methodologies which have been adopted in privacy preserving data mining but Privacy maintaining in different field is one of the interesting and evolving task [3].

Privacy alludes for extraction of sensitive information finished through mining process. Then again, the pointless handling impact of brilliant frameworks put the delicate and individual information that exist in vast and scattered information stores at danger. Late enhancements in information growth have kept up the social event and treatment of enormous volume of individual data, for instance, remedial record, information, shopping inclinations, credit and restorative history, and driving records [4].

RESEARCH OBJECTIVES

The objective of this research is to use hybrid methodology for secrecy maintaining in data extraction process and protect or preserve individual data with low information loss. Hybrid methodology is preserve sensitive or non-sensitive data of individual or any organization in more secure way and this approach protect different attacks that involve in privacy theft.

PROBLEM STATEMENT

Different procedures for privacy stabilizing are used for privacy preservation but there is still some confines these are given below [5].

- Similarity threats
- Heterogeneity threats

MATERIALS and METHODS

In this consolidated procedure two techniques are utilized to acquire any protection saving strategy. In the given figure consolidate procedure is this procedure indicate view of the information change and information encryption systems [6]. This approach utilizes the consolidated strategies of randomization and k-anonymization and it is containing three primary preferences:

- It secures private information with low data misfortune.
- Effectiveness of information is expanded.
- Data can likewise be recreated.

There is a harmony between the security and data misfortune. To keep up this trade, an effective protection conservation technique is used.

In this technique, first we apply randomization on unique information and afterward after randomization sort the delicate quality qualities into high touchy and low delicate class. Besides apply k-anonymization on those tuples that exist in a place with high touchy class and those tuples who relate with low place delicate stay as it seems to be. So, it decreases the data misfortune and enhances the ease of use of information. The mix of anonymization with randomization method is made troublesome for the aggressor to assault on database. Security preservation is an exceptionally immense field. Numerous calculations, for example, k-anonymity,

information irritation, l-diversity calculation and so forth have been proposed with a specific end goal to secure the information.

Essentially hybrid approach is joined into two algorithms. In Ist process randomization is performed on dataset utilizing characteristic transitional likelihood grid and in process II k-anonymity is performed on randomized dataset which is aftereffect of 1 process. Calculation I is demonstrated as follows.

In segment 1, after picking virtual identifiers, classified qualities and imperative focuses; the medium of probability ought to be planned. Medium of probability is the odds of presence of every event under various circumstances. There are 3 virtual identifiers: stage, sex, and postal division. Here target is to decide the possibility of occasion of every case. Through this every case ought to be considered. Study the first quality of age, the first virtual identifier. Every esteem for age i.e. 33, 29, 21, 31, 60, 25 is reflected with each column. To compute the likelihood of event of every sexual orientation or dieses underneath circumstances such when the sex is "so thus", when the manifestation 1 is "so thus", when the side effect 2 is "so thus", when the side effect 3 is "so thus". The chances of presence of age 33 under conditions such when the sexual orientation is "Male", when the side effect 1 is "Hack", when the indication 2 is "Mid-section torment", when the manifestation 3 is "breath issue" is arranged first. After that next column is considered at this same age.

After completion of segment 1, the consequence of segment 1 is given to next fragment as reaction. In this procedure randomization strategy is helpful for entire information which may bring data about misfortune due to over anonymization. So here it considers just exceedingly delicate data and applies randomization just to this area. On the off chance that Age and possibility of presence is 21 and odds of occasion event is 1/3 correspondingly if age is 35 and odds of event is 2/3 likelihood then likelihood of age is p(J) p(n) p(q) and likelihood (event of age 21 under conditions when gender='Male' and Disease='HIV+') * likelihood (event of age 35 under conditions when sexual orientation ='F' and Disease='Cancer').

P(J)= 1/3* 2/3* 2/3' 21
P(n) = 2\3 * 1\3 * 1\3 " 35
P(q) = 2\3 * 2\3 * 2\3 " 35

RESULTS and DISCUSSION

Apply transitional probability matrix of above given data set of patient and calculate probability of p1 these values show in below figure first in which apply mapping on data and match probability values with quasi identifier and then randomly generate probability (p1) according to column size and measure chance of success and failure through probability matrix and draw graph in which along x-axis show number of rows and y-axis show probability of data set and graph show random probability values of different rows or tuples and show the percentage of similar data in our data set through transitional probability matrix chances of similarity attacks are minimized. Probability matrix apply randomly these random values are not fixed and by this bundle of different values is produce and when a data set values are tested then different values for the same data set are generate in the form of output through this benefits and loss of data is reduced and data utility is improved in a better way and provide privacy for data set or individual records. Below probability of 1st data set given.

Table 1. Probability of 1st Data set

0.7	0.8	0.24	0.16	0.1	0.78
0.16	0.45	0.9	0.31	0.56	0.66
0.55	0.4	0.39	0.52	0.25	0.34
0.75	0.4	0.37	0.75	0.58	0.47
0.61	0.69	0.71	0.68	0.8	0.38
0.81	0.72	0.3	0.8	0.35	0.36
0.6	0.54	0.1	0.3	0.1	0.8

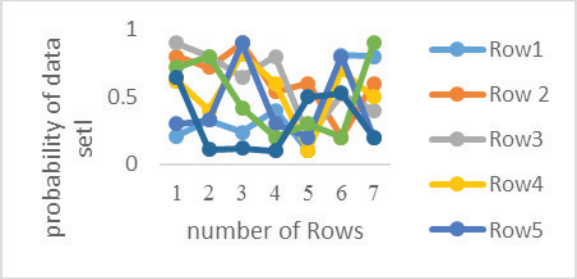

Figure 1. Graph of Random Values of 1st Data set

Above graph show random values of first data set it contain tuples and probability.

These values show probability p2 of same data set when again apply probabilities.

Table 2. Probability of 2nd Data set

0.21	0.33	0.24	0.4	0.81	0.8
0.8	0.72	0.9	0.84	0.2	0.6
0.9	0.8	0.65	0.8	0.52	0.4
0.62	0.4	0.82	0.6	0.7	0.5
0.3	0.33	0.9	0.3	0.8	0.2
0.72	0.8	0.42	0.2	0.2	0.9
0.65	0.11	0.12	0.1	0.53	0.2

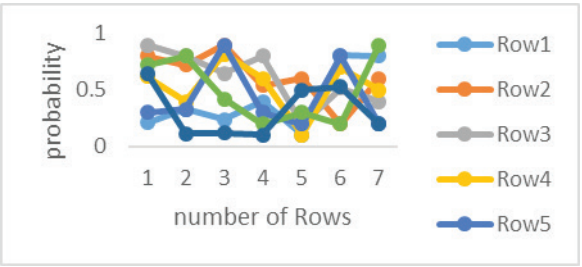

Figure 2. Graph of Random Values of 2nd Data set

This graph show random result of probability of p(2) for same data set.

Table 3. Probability of Combined Technique

0.1	0.9	0.7	0.7	0.5	0.9	0.7
0.4	0.8	0.8	0.5	0.6	0.2	0.8
0.1	0.3	0.2	0.1	0.2	0.6	0.3
0.2	0.2	0.3	0.9	0.7	0.5	0.4
0.8	0.5	0.8	0.8	0.6	0.3	0.1
0.6	0.7	0.9	0.9	0.9	0.4	0.2
0.5	0.3	0.3	0.1	0.8	0.7	0.3

Above values show combine random probability of p1 and p2=p3

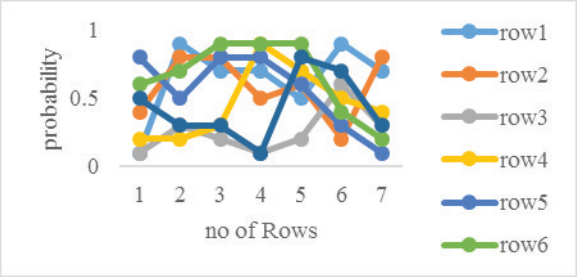

Figure 3: Graph of combine random results of p1 and p2

This graph show combine random results of p1 and p2 Output of Matrix of Probability for 1st and 2nd for same data set. The objective of this research is to handle the data integration, analytics for students of UAF by using big data. To bring data together will be from multiple data sources and providing relevant information a unified view to achieve the goal of user will be the main purpose of the study.

CONCLUSION

Main purpose of this investigation is improvement in data security for this purpose a combine approach of k-anonymity and randomization used both techniques partially removes the problems of different privacy preserving techniques. Through this data or information value increase, and data defeat ratio is decreased and it also provide privacy at the same time. But this hybrid technique give a better solution to this problem. In future to sort out these problems, hybrid approach of different techniques is introducing we combine three or more privacy preserving techniques for better data protection purpose through this data threats decreased and data of individual is secure in more effective way.

SUMMARY

Data mining has risen as a remarkable innovation for picking up information from endless amounts of business information, monetary information, arranged information and therapeutic information. The objective of data mining methodologies is to create summed up learning as opposed to distinguish data against particular person. There has been growing threat that utilization of this innovation is disregarding singular protection. This is opening new difficulties in the range of Privacy preserving in Data Mining. Security directions and other protection concerns may keep information proprietors from sharing data for performing information examination. To accomplish an answer for this issue, information proprietors must outline a methodology that meets protection necessities and certifications substantial information extraction outcomes. In the suggested system, we solidified K-obscurity with randomization. It makes troublesome for the attacker to recognize establishment and homogeneity ambush. A side that it secures private data with better precision and gives no loss of information which extends data utility.

REFERENCES

[1] Agrawal, R. and R. Shrikant., “Privacy preserving data mining”, ACM SIGMOD reco New York, 2(1): 439-450(2000).

[2] Malik, M. B., M. A. Chazi and R. Ali., “Privacy Preserving Data mining techniques current scenario and future Prospects”, International Conference on computer and communication Technology on IEEE, 1(1): 1740-1745(2012).

[3] Wang, J., Y. Luo, Y. Zhao and J. Le., “A survey on privacy preserving”, *Internationa Workshop on Database Technology an Application*, 1(1): 114-147(2009).

[4] Li, T and N. Li. 2009., “On the tradeoff between privacy and utility in data publishing”, *In Proceedings of ACM SIGKDD international conference on knowledge discovery and data publishing*, 1(1): 517–26(2009).

[5] Vassilios, S., E. Verykios, I. N. Bertino, L. P. Fovino, Y. Provenza, Y. Saygin and Y. Theodoridis., “State of the art in privacy preserving data mining”, *In the proceeding of SIGMOD Record*, 33(1): 50-57(2004).

[6] Liu, K., H. Kargupta and J. Ryan. 2006. “Random Projection Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining”, IEEE transactions on knowledge and data engineering, 18(1): 92-106(2006).