Turkish Journal of Computer and Mathematics Education Vol.12 No. 5 (2021), 1450-1456

Research Article

A General Framework Information Loss of Utility-Based Anonymization in Data Publishing

Waleed M.Ead[a*], Emad El-Abd[b], Mohamed M.Abbassy[c]

[a*]Faculty of Computers and Artificial Intelligence, Beni-Suef University, Egypt.
E-mail: waleedead@hotmail.com
[b]Faculty of Computers and Information, Menoufia University, Shebin El Kom, Egypt.
Department of Computer, Deanship of Educational Services, Qassim University, Buraydah, Saudi Arabia.
E-mail: Emadqap@gmail.com
[c]Faculty of Computers and Artificial Intelligence, Beni-Suef University, Egypt.
E-mail: Moh.abbassy@gmail.com

Article History: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstract: To build anonymization, the data anonymizer must determine the following three issues: Firstly, which data to be preserved? Secondly, which adversary background knowledge used to disclosure the anonymized data? Thirdly, The usage of the anonymized data? We have different anonymization techniques from the previous three-question according to different adversary background knowledge and information usage (information utility). In other words, different anonymization techniques lead to different information loss. In this paper, we propose a general framework for the utility-based anonymization to minimize the information loss in data published with a trade-off grantee of achieving the required privacy level.

Keywords: Anonymization, Privacy, Utility, Information Loss, Data Publishing.

1. Introduction

Publishing data in its original raw material may lead to privacy-breached issues. Moreover, the data owner's different sources of data may contain a piece of sensitive information that needed to be preserved and protected from any attacking performed by anyone have any piece of background knowledge.

Besides, to publish such valuable data for market-based analysis purposes, such data may be needed to be anonymized to keep the data privacy from and breaches issues may occur. Nevertheless, there many different anonymization techniques subject to the different data requirements of analysis. So, the different anonymization techniques lead to information loss of the data utility.

To build anonymization, the data anonymizer must determine the following three issues: 1- which data to be preserved? 2. Which adversary background knowledge used to disclosure the anonymized data? 3. The usage of the anonymized data? We have different anonymization techniques from the previous three-question according to different adversary background knowledge and information usage (information utility). In other words, different anonymization techniques lead to different information loss.

This paper aims to propose a general framework for utility-based anonymization to keep the data's privacy from one side and minimize the information loss of the data utility from the other side. Also, a literature review of the different data utilities is presented.

The paper is organized into two main sections. The first section reviews the related backgrounds works in published data with different data utilities. The other section presented the proposed general framework for the utility-based anonymization. Finally, the conclusion of the paper is presented.

2. Related Backgrounds

Some previous works [1-5,30-31] study the adjustment of parameters in the anonymization process for the trade-off between privacy and utility. In the following, we summarize some related works in secure published the original data for different data utilities.

2.1. Query Answering

Hussien et al. [1], proposed a utility-based privacy-preserving technique. Their approach takes into account the attributes of sensitive values represented in the queries. They allowed data owners to assign weights to the attributes and anonymized the queries using generalization boundaries only if the sensitive attribute's values exceed the threshold. The value assigned to each attribute depends on the utility of such attribute. The attributes whose total weights exceed the threshold values is anonymized using generalization boundaries, and the other queries can be directly published without any modifications. They measured the data utility by calculating the distributions' differences between the original and anonymized data by Cluster Analysis Measure and Empirical CDF Measures. Their work preserved the privacy and increased the utility by reducing the information loss because the only sensitive attributes in queries are generalized at all and generalized using generalization boundaries.

A.W. Fu et al. [2] proposed a framework, called SPLU allows publishing data with a high data utility in case of large sum aggregate queries because the returned results are with high accuracy. In contrast, SPLU offering high inaccuracy for small sum aggregate queries to ensure privacy. The framework applies randomized perturbation on the sensitive values. The small count privacy and large count utility in this framework use the number of records involved to differentiate the reconstruction for privacy concern and utility reconstruction. They measure relative error to demonstrate the privacy and utility levels, where a higher relative error corresponds to more privacy and less utility.

Mohan et al.[3] mention GUPT to increase accessibility and precision of data analysis on a particular privacy framework. The' privacy budget' term used to determine privacy through current variable privacy frameworks means better privacy with a smaller data security budget. Nevertheless, this Privacy Measurement Unit does not easily translate into the application's usefulness and is therefore not easy to interpret for data analysts who are not privacy experts. Also, analysts can effectively allocate this privacy budget through several queries on a data set to prevent incorrect analysis and reduce the number of queries that can be carried out on the dataset safely. The GUPT overcomes the constraints of traditional differential anonymity by encouraging organizations to evaluate their datasets differently. Any improvements are allowed to incorporate the new Procedure. In addition to current differential privacy systems, the platform allows data analysts to determine the target performance's exactness.

Zhu et al. [4] provide an approach for answering private queries with many differentially private results in a continued released dataset. They identified and used the coupled information in the continual datasets and presented the notion of coupled sensitivity that satisfied the requirement of differential privacy with less noise in the query output. Their experimental results showed that their technique is robust, effective, and preserving privacy with little loss of utility.

Gkoulalas et al.[5] presented a survey on privacy-preserving of published electronic health records. Their survey analyzed more than 45 algorithms in publishing structured electronic health recorded data under privacy constraints. To maintain privacy and good usage, the algorithms chosen for analysis are effective. In addition to data privacy, three methods are being explored to protect the data utility: 1) calculate data loss using a method of optimization, which they try to minimize. 2) to define the data analysis process to be carried out with the published data and to e its accuracy, and 3) to take inconsiderefulness requirements set out by the data owners. The data that meets such specifications should be given.

2.2. Statistical Sensitive Inference

Sathiyapriya and Sadasivam[6] proposed a survey on privacy-resistant group rules mining. Since introducing algorithms to the privacy protection of association rules, various association rules can still be extracted from the anonymized results. The privacy-preserving algorithms of association rule mining can be categorized into 3 categories: 1) heuristic techniques, 2) reconstruction-based association rules and 3) cryptographic techniques. They proved that forbidden query related algorithms are restricted to binary data, which can be applied to quantitative data to cover sensitive associated laws such as privacy-conserving rule mining utilizing genetic algorithms.

Abdulkader et al. [7] have proposed a valuable association rule method that hides data from online social network (OSN) participants, even OSN apps, favouring privacy-saving user profiles. They suggested a rule-hiding algorithm to mask the confidential information contained in the publishing of a user profile. The algorithm is based on avoiding violations of user privacy of the user's most important feature. This technique covered critical attribute sets of the user's profiles.

Wang et al. [8] presented an approach to solving the effect of Non-independent reasoning (NIR) in privacy without a big effect in utility. The NIR allows the information about one record in the data to be learned from the

information of other records in the data. They proposed a data perturbation approach that allows learning statistical relationships and prevents sensitive Non-independent reasoning (NIR) about an individual.

Wang et al. [9] presented an anonymization approach for publishing sensitive data based on grouping records into small buckets to prevent reasoning using sensitive information to identify information. Unlike traditional packaging approaches that use defined privacy options, the solution uses dynamic values with each sensor feature's privacy setting. Buckets are thus shaped in various sizes. They proposed an effective solution based on two separate types of buckets. There are two aspects to this approach. The first section discusses whether a bucket system has a correct bucket record assignment while the privacy requirement is followed. The second part searches for an optimum bucket design. The efficiency criteria they considered involve answers to numbers of important components for many analysis activities, such as contingent tables, analysis of correlation often mined objects, building decision tree or naive Bayes classifications. The technique protects exactly the counts for a count query with only public characteristics, information loss only exists for a count request with public characteristics and confidential attributes.

Calmon et al. [10] propose a general statistical inference framework to preserve the privacy attacks acquired by a passive adversary given utility constraints. This adversary attempts to infer the user's private information from the user's public (released) data. In this model, the user releases a set of measurements to an analysis while keeping data correlated with these measurements private. The analyst is an authorized receiver for these measurements, from which he expects to derive some utility. Their privacy model attempts to balance the privacy criteria with the analyst's usefulness needs by mitigating the user's privacy threat while meeting the analyst's utility limitations. The framework enables the elimination of a portion of the public user's data, modifying the meaning of other items of a public profile, among certain forms of records distortion. For certain distributions of the input on databases, an adversary may deduce the reference source from a differentially private query with arbitrarily high accuracy [11].

2.3. Association Rule Mining

Jisha et al.[12] provided a model of anonymization adaptive utility-based (AUA), which uses help and trust in related mining. There may be specific needs and criteria among different data holders. Consequently, the details exchanged can often differ depending on these issues and requirements. According to the usability of the top k association [13, 14], the structure is not appropriate. It may lead to fewer usage items [7] if associated objects are identified. Data anonymization without affecting data mining results [15] can also be done.

A multilevel innovative data mining methodology was developed by Yapping et al.[16]. These avoid malicious data miners from correlating incorrectly disrupted versions of the same data and infringing on the privacy. They also checked the same efficiency of the disrupted copy of their model as the independent noise version when their confidence level is the same.

2.4. Data Classification

Zaman et al. [17] proposed a literature review on maximizing the anonymized data utility from the data classification perspective. They utilize differentiated privacy in the use of decision-making arrangements to anonymize and publicate data that show significant quality changes—also, Jaiswal et al.[18] implemented a comprehensive study; the system proposed showed significant privacy and efficiency enhancements centred on the relationships across various datasets' attributes. They utilize entropy and the advantage of information to identify data distributions to protect their privacy. The proposed research can classify multiple relationships based on data context and use. Zaman et al. (19), after the inclusion of several noises in the data released, include an analysis to release data to classification purposes. Anonymous data generated by the current systems are, therefore, less useful [20]. Identity cannot be adequately covered, and the confidential release of knowledge cannot be accurately presented.

Lee et at. [21] proposed a k-anonymity approach for preserving health care data publishing. Their proposed algorithm is based on full-domain generalization [22]. To overcome the less information utility due to full domain generalization, they proposed an h-ceiling concept to limit the overgeneralization levels [23].

Furthermore, they also suggested the insertion of fraudulent records in the corresponding classes to fulfil the K-anonymity but the h-ceiling. Fake records have the same quasi identity details as similar class data, and sensitive data is randomly selected within the relevant attribute domain. Moreover, all fake records are registered

in a fraudulent records database. Adversaries receive an anonymous table, index tables and try to recover sensitive details with falsified records. They should remove a record that has a false one.

Sheikhalishahi et al.[24] suggested a shared data classification data sharing system for privacy knowledge. Their privacy mechanism approach is focused on the collection of privacy utilities which exclude the most irrelevant data accuracy and privacy features. Kayem et al.[25] suggested a k-anonymization method to cover tuples with highly sensitive values in each equal category of extreme weighing and packing. They used a random sampling method to balance privacy with utility. It only considers crime data since it includes a huge volume of sensitive information. Therefore, information efficiency matrices have not been used. Many methods were documented utilizing techniques of generalization. Mehmet et al.[26] suggested hybrid generalizations that comprise generalizations and the framework for data relocation. Cell migration involves changing some cell cells to complement tiny, inseparable groups of tuples. Such a methodology seeks to counteract the consequences of overallocation and outliers.

2.5. Mining Datastream

A hot topic in the area of data privacy protection is data mining, which quickly grows at an exponential pace. In these scenarios, the privacy issue is quite complicated, as the details are gradually revealed. It is noted that data streams and data mining are relatively new, and the synthesis of these two topics did not take much time. Krempl et al. [27] identified two big data stream privacy challenges.

The incompleteness of information is the first challenge. The data were modified online in sections and models. Therefore, the layout is never definitive, and privacy protection is difficult to judge before you see all evidence. Suppose, for example, that traces of people are gathered for traffic modelling. Suppose individual A is moving from the campus to the airport at present. If there are no such trips of other individuals very shortly, a person's privacy will be compromised. But the current time, when the pattern must be revised, is uncertain shortly. On the other hand, data mining algorithms may have some intrinsic privacy protection capabilities because not all modelling data need to be accessed all at once, and parts of data can be slowly modified. Another important guide for potential work is analyzing the privacy protection properties of current data stream algorithms.

The definition of data flow is the second major problem for the protection of privacy. When data changes over time, there may no longer be fixed rules for privacy protection. Suppose winter is coming, for example, with snowcaps and fewer cyclists cycling. By recognizing that a person comes by bike to work and has a range of GPS signs, this individual may not be detected in summer only when there are more cyclists, but in winter. Creating flexible data protection strategies to identify such a condition and change to maintain privacy in new circumstances is thus an essential roadmap for future research.

Few research concern data streams and the protection of privacy. To identify the data streams, Chhinkaniwala et al.[29] suggested a system of data protection. Proposed algorithms can disrupt responsive attributes only with number values. More comments can be found in Nyati et al.[29]'s appraisal report.

3. Utility-Based Anonymization (UBA) General Framework

Given that each analyst/researcher may have different needs and data requirements, it may not be efficient to provide the same anonymized data to different analysts, even though information loss could be well balanced due to information utility. The following figure 1 depicted a general framework for the utility-based anonymization (UBA). The generalized framework has two types of data analysis requests, e.g. request for query answering data or knowledge discovery through data mining techniques.

3.1. Sanitizing Query Answering

It is assumed that the data itself is kept secret, but that the data owner wants to allow some query access to it while at the same time preventing private information from being revealed. For example, a hospital may want to allow analysts studying prescribing practices to query the patients' records for information about medicines dispensed in the hospital. Still, they want to ensure that no information is revealed about the medical conditions of individual patients. To concentrate, the hospital wants to check whether answering specified legal queries could increase knowledge (from whatever source) that an attacker may have about the answer to a query for patient names and their medical conditions; sensitive query. Considering that an attacker may have previous

knowledge about the system is of crucial importance, as such knowledge may connect the answers to legal and sensitive queries, and lead to the (partial) revelation of the latter.

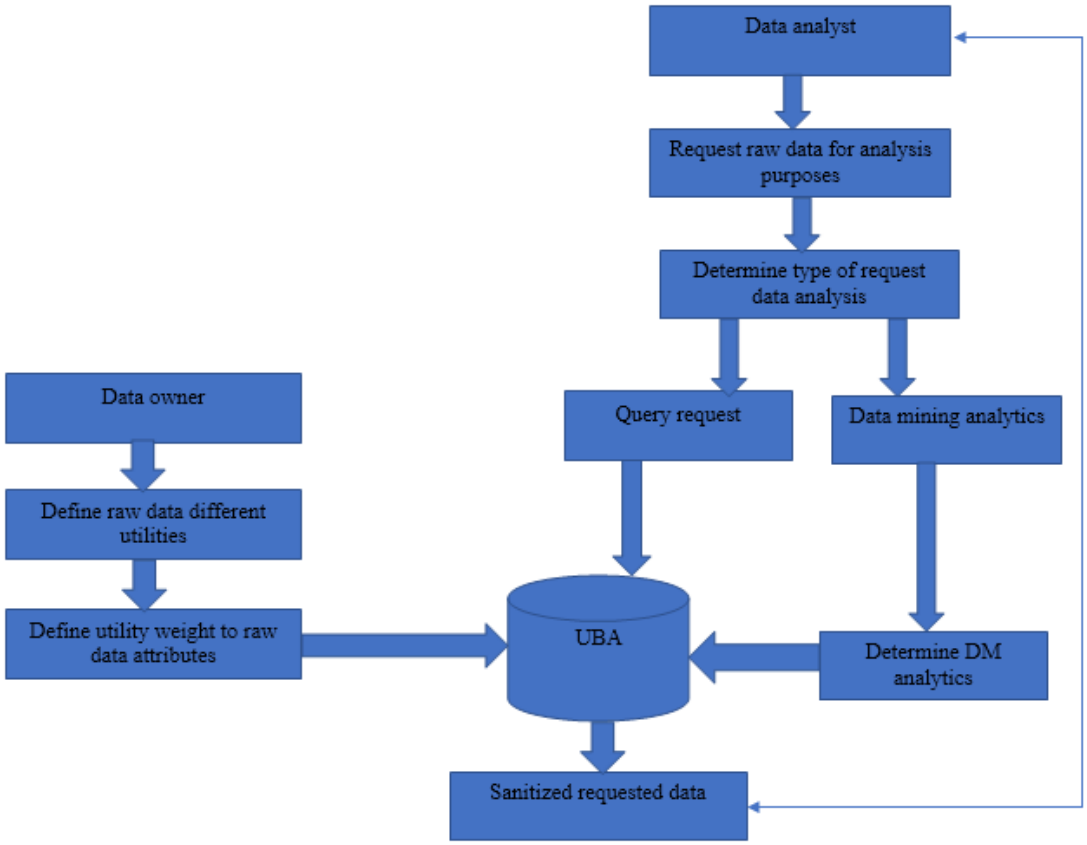

Figure 1. Utility-Based Anonymization General Framework

The submitted query contains different attributes with different information utilities that may be sensitive or not. Furthermore, such queries come from different sources with different concerns. The correlation between such attribute together may disclose the privacy of the individual.

The main idea of our proposed general framework for preserving privacy under query answering is that:

The data owner knows the sensitivity degree to each attribute in the database records. So, the proposed framework will be based on such sensitivity degree. Also, we can record the degree of sensitivity for all attributes in the submitted query for each data request. Furthermore, the data owner may make an authentication channel for each receipt's query. To summarize the basic idea in the following steps:

- According to the data available in each attribute, the data owner defines the privacy sensitivity degree for each owner.
- The data owner defines the total sensitivity threshold to be discovered.
- For each request submitted, compute the total sensitivity for attributes submitted in the query. Also, compute the sensitivity correlation degree between submitted query attributes. Unlike previous works, they don't see the correlation between attributes submitted in the query request. Without study, such correlation between query attributes may be used in the similarity attack between data tables.
- If the total sensitivity for all query attribute and sensitivity correlation degree is greater than the predefined thresholds defined by the data owner.
- Anonymize sensitive data attributes
- Return query results

3.2. Sanitizing Knowledge Discovery

Data mining methods are an effective use of data to derive useful information from a large data set. Data mined information is based on it and may, therefore, enable inferences on original data to be removed, even if not expressly having original data, thus jeopardizing privacy constraints imposed upon the original data. K-

anonymity often refers to this finding. Consequently, the desire to ensure the data gathered's privacy may allow the future performance of the data mining operation to be restricted. Specific privacy risks emerging from mining a collection of personal data stored under confidentiality restrictions in a private data table.

4. Conclusion

The trade-off and adjustment of the parameters between the anonymization techniques and the different data analysis requirements will maximize the benefits of sharing the data. We have proposed a brief literature survey of the different data utilities that lead to different data anonymization techniques. Moreover, the paper proposed a general framework for the utility-based anonymization (UBA) techniques. It is the first general framework that adjusts the anonymization levels based on the different utilities of the data requirements for our best of knowledge. As future work, the authors will provide an experimental study on different applications using their proposed general framework of UBA. Besides, they will propose a modified version in case of streaming the data.

References

1. Darwish, N.R., & Hefny, H.A. (2015). Utility-based anonymization using generalization boundaries to protect sensitive attributes. Journal of Information Security, 6(03), 179.
2. Fu, A.W.C., Wang, K., Wong, R.C.W., Wang, J., & Jiang, M. (2014). Small sum privacy and large sum utility in data publishing. Journal of biomedical informatics, 50, 20-31.
3. Mohan, P., Thakurta, A., Shi, E., Song, D., & Culler, D. (2012). GUPT: privacy preserving data analysis made easy. In Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, 349-360.
4. Zhu, T., Li, G., Xiong, P., & Zhou, W. (2018). Answering differentially private queries for continual datasets release. Future Generation Computer Systems, 87, 816-827.
5. Gkoulalas-Divanis, A., Loukides, G., & Sun, J. (2014). Publishing data from electronic health records while preserving privacy: A survey of algorithms. Journal of biomedical informatics, 50, 4-19.
6. Sathiyapriya, K., & Sadasivam, G.S. (2013). A survey on privacy preserving association rule mining. International Journal of Data Mining & Knowledge Management Process, 3(2), 119.
7. H. AbdulKader, E. ElAbd, and W. Ead, "Protecting online social networks profiles by hiding sensitive data attributes," Procedia Computer Science, vol. 82, pp. 20-27, 2016.
8. AbdulKader, H., ElAbd, E., & Ead, W. (2016). Protecting online social networks profiles by hiding sensitive data attributes. Procedia Computer Science, 82, 20-27.
9. Wang, K., Han, C., Fu, A.W., WONG, R.C., & Yu, P.S. (2015). Reconstruction privacy: Enabling statistical learning.
10. Wang, K., Wang, P., Fu, A.W., & Wong, R.C.W. (2016). Generalized bucketization scheme for flexible privacy settings. Information Sciences, 348, 377-393.
11. du Pin Calmon, F., & Fawaz, N. (2012). Privacy against statistical inference. In 2012 50th annual Allerton conference on communication, control, and computing (Allerton), 1401-1408.
12. S. Salamatian, A. Zhang, F. du Pin Calmon, S. Bhamidipati, N. Fawaz, B. Kveton, et al., "How to hide the elephant-or the donkey-in the room: Practical privacy against statistical inference for large data," in GlobalSIP, 2013, pp. 269-272.
13. Salamatian, S., Zhang, A., du Pin Calmon, F., Bhamidipati, S., Fawaz, N., Kveton, B., & Taft, N. (2013, December). How to hide the elephant-or the donkey-in the room: Practical privacy against statistical inference for large data. In 2013 IEEE Global Conference on Signal and Information Processing, 269-272.
14. Panackal, J.J., & Pillai, A.S. (2015). Adaptive utility-based anonymization model: Performance evaluation on big data sets. Procedia Computer Science, 50, 347-352.
15. Ryang, H., & Yun, U. (2015). Top-k high utility pattern mining with effective threshold raising strategies. Knowledge-Based Systems, 76, 109-126.
16. Y. C. Lin, C.W. Wu, and V. S. Tseng, "Mining high utility itemsets in big data," in Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2015, pp. 649-661.
17. Lin, Y.C., Wu, C.W., & Tseng, V.S. (2015, May). Mining high utility itemsets in big data. In Pacific-asia conference on knowledge discovery and data mining, Springer, Cham. 649-661.
18. Aldeen, Y.A.A.S., & Salleh, M. (2016). A Hybrid K-anonymity Data Relocation Technique for Privacy Preserved Data Mining in Cloud Computing. Journal of Internet Computing and Services (JICS), 5, 51-58.
19. Li, Y., Chen, M., Li, Q., & Zhang, W. (2011). Enabling multilevel trust in privacy preserving data mining. IEEE Transactions on Knowledge and Data Engineering, 24(9), 1598-1612.

20. Zaman, A.N.K., Obimbo, C., & Dara, R.A. (2016). A novel differential privacy approach that enhances classification accuracy. In Proceedings of the Ninth International C* Conference on Computer Science & Software Engineering, 79-84.
21. Jaiswal, J.K., Samikannu, R., & Paramasivam, I. (2016). Anonymization in PPDM based on data distributions and attribute relations. Indian J. Sci. Technol, 9(37).
22. Zaman, A.N.K., & Obimbo, C. (2014). Privacy preserving data publishing: A classification perspective. International Journal of Advanced Computer Science and Applications; The Science and Information (SAI) Organization Limited: West Yorkshire, England, 5.
23. Majeed, A., Ullah, F., & Lee, S. (2017). Vulnerability-and diversity-aware anonymization of personally identifiable information for improving user privacy and utility of publishing data. Sensors, 17(5), 1059.
24. Lee, H., Kim, S., Kim, J.W., & Chung, Y.D. (2017). Utility-preserving anonymization for health data publishing. BMC medical informatics and decision making, 17(1), 1-12.
25. Kohlmayer, F., Prasser, F., Eckert, C., Kemper, A., & Kuhn, K.A. (2012). Flash: efficient, stable and optimal k-anonymity. In 2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing, 708-717.
26. Nergiz, M.E., & Gök, M.Z. (2014). Hybrid k-anonymity. Computers & security, 44, 51-63.
27. M. Sheikhalishahi and F. Martinelli, "Privacy-utility feature selection as a tool in private data classification," 2018.
28. Sheikhalishahi, M., & Martinelli, F. (2017). Privacy-utility feature selection as a tool in private data classification. In International Symposium on Distributed Computing and Artificial Intelligence, Springer, Cham, 254-261.
29. Kayem, A.V., Vester, C.T., & Meinel, C. (2016). Automated k-anonymization and l-diversity for shared data privacy. In International Conference on Database and Expert Systems Applications, Springer, Cham. 105-120.
30. Nergiz, M.E., Gök, M.Z., & Özkanlı, U. (2013). Preservation of utility through hybrid k-anonymization. In International Conference on Trust, Privacy and Security in Digital Business, Springer, Berlin, Heidelberg, 97-111.
31. Krempl, G., Žliobaite, I., Brzeziński, D., Hüllermeier, E., Last, M., Lemaire, V., & Stefanowski, J. (2014). Open challenges for data stream mining research. ACM SIGKDD explorations newsletter, 16(1), 1-10.
32. H. Chhinkaniwala, K. Patel, and S. Garg, "Privacy preserving data stream classification using data perturbation techniques," in Proceedings of International Conference on Emerging Trends Electrical, Electronics and Communication Technologies, 2012.
33. Chhinkaniwala, H., Patel, K., & Garg, S. (2012). Privacy preserving data stream classification using data perturbation techniques. In Proceedings of International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies, 1-8.
34. Nyati, A., & Bhatnagar, D. (2016). Performance Evaluation of Anonymized Data Stream Classifiers. International Journal of Computer Science and Network-IJCSN, 5(2).
35. Duncan, G.T., Keller-McNulty, S.A., & Stokes, S.L. Disclosure risk vs data utility: the R- U confidentiality map. In: Technical report number 121, National Institute of Statistical Sciences; December 2001.
36. Loukides, G., & Shao, J. (2008). Data utility and privacy protection trade-off in k-anonymisation. In Proceedings of the 2008 international workshop on Privacy and anonymity in information society, 36-45.