Impact Factor (SJIF): 3.632

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

(Special Issue for ITECE 2016)

Privacy Preserving Data Mining: An Active Research Area

Priyanka Garach[1], Darshana Patel[2], Radhika Kotecha[3]

[1]*PG Student, Information Technology Department, V.V.P. Engineering College*
[2]*Assistant Professor, Information Technology Department, V.V.P. Engineering College*
[3]*Assistant Professor, Information Technology Department, V.V.P. Engineering College*

Abstract — Data Mining has plentiful benefits and covers a wide range of applications like modern business, e-commerce, government sectors, health environments, etc. But it is apparent that the gathering and analysis of sensitive personal as well as collective data causes a serious menace to privacy, confidentiality and freedom. This crucial issue has led to emergence of the field called Privacy-Preserving Data Mining and is a state-of-the-art research trend. Privacy-Preserving Data Mining deals with efficient conduction and application of data mining without scarifying the privacy of the data. This paper highlights some of the substantial applications and key methods of privacy-preserving data mining that are target of research in this field.

Keywords – Data Mining, Privacy, Privacy-Preserving Data Mining; Pattern Recognition; Privacy-Preserving Data Mining Applications

I. INTRODUCTION

Data mining, also known as knowledge discovery from data, is the process of analyzing data from different perspectives and summarizing it into useful information that can be used to increase revenue, cuts costs, or both [1, 2]. That is, data mining is used for pattern generation from huge datasets. For such pattern finding, proper methods of data mining must be applied according to the requirement. Data mining mainly includes techniques such as classification, association, clustering, etc.

Privacy-preserving data mining (PPDM) deals with hiding sensitive information of individuals like names, addresses, etc. without compromising the usability of data [3, 4]. Privacy-preserving data mining should occur such that while providing access to the published data the adversary should not learn anything extra about any target victim even in presence of any background knowledge obtained from additional sources [5, 6, 7]. An efficient privacy-preserving data mining technique must ensure that any information disclosed should not: 1) be traced to a specific individual, and, 2) form an intrusion.

Figure 1 shows a schematic view of privacy-preserving data mining. In traditional knowledge discovery from data, the data mining techniques are applied on the data sets to yield patterns that violate privacy. On the other hand, in privacy-preserving data mining, sanitized patterns are obtained by first applying privacy-preserving techniques on data sets followed by application of data mining techniques.

Figure 1. A schematic view for privacy-preservation

II. TYPES OF PPDM

In most cases, privacy preservation occurs in two chief dimensions: 1) preservation of individual's personal information and 2) preservation of information concerning their collective activity [8].

2.1. Individual privacy-preservation

The principal objective of data privacy is the safety of personally identifiable information. Generally, information is considered personally identifiable if it can be linked, directly or indirectly, to an individual person. Hence, when individual's data are subjected to mining, the attribute values containing person's secret information must be kept secure and should not be disclosed in any situation. By this, miners are not able to learn from personal information and the goal of overall global pattern mining can be achieved.

2.2. Collective privacy preservation

Protecting personal data may not be enough. Sometimes, we may need to protect against learning sensitive knowledge representing the activities of a group. The protection of sensitive knowledge is referred to as collective privacy preservation. As similar to statistical databases, the aim here is to prevent disclosure of confidential information, even when the aggregate information about groups is available, provided by security control mechanism. Another aim of collective privacy preservation is to protect personally identifiable information and at the same time it should also protect some patterns and trends that are not supposed to be discovered.

III. APPLICATIONS OF PRIVACY-PRESERVING DATA MINING

Privacy-preserving data mining covers a wide range of applications that include: epidemics detection in medical field, attack prediction in bioterrorism, identity theft, watchlist problem in homeland security [3], identity protection in DNA databases for genomic privacy, link prediction in social network analysis, protecting identity disclosure while releasing public information. This section describes some of the major applications in details.

3.1. Credit card fraud detection

Major risky frauds are occurring in the banking sector. Banks have massive amount of databases. From these databases, the precious business information can be identified. Credit card fraud detection is the process of identifying those transactions that are fraudulent into two classes of legitimate (genuine) and fraudulent transactions [9]. Credit card frauds can be broadly classified into three categories, that is, traditional card related frauds (application, stolen, account takeover, fake and counterfeit), merchant related frauds (merchant collusion and triangulation) and Internet frauds (site cloning, credit card generators and false merchant sites) [9].

As shown in figure 2, the dataset of credit card transaction contains valuable information like, IP address, proxy servers, email id, shipping address, average transaction amount, income of an individual. To detect the fraud in credit card transaction, firstly privacy-preserving techniques must be applied to the above dataset. The result obtained by this process is the sanitized data. Further, data mining techniques like classification are applied on this sanitized data to get the category of transaction, i.e., whether it is legitimate transaction (authenticate transaction) or it is fraudulent transaction (unauthenticated transaction).

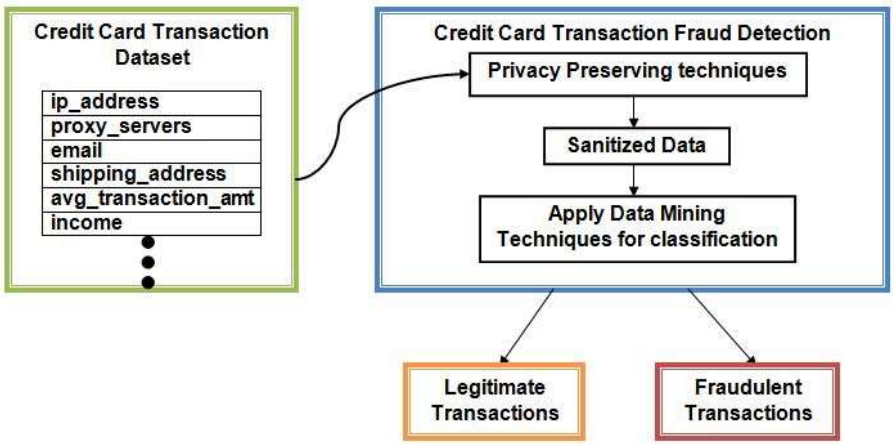

Figure 2. Privacy-preserving in credit card fraud detection

3.2. Video-surveillance

In the context of sharing video-surveillance data, the major threat is the usage of facial recognition software. Let's take an example of driving license database in which the facial images in videos can be matched to the facial images in a driver license database [3], as shown in figure 3. The simple solution to this problem is to totally black out each face. But as a result of this action, all the facial information will been removed. So a new optimal approach is discovered, which contains the use of selective downgraded image of the facial information so that it confines the ability of facial recognition software to consistently spot the exact face and at the same time maintain facial details in images [10]. The algorithm is referred to as *k*-same, and the key is to identify faces which are somewhat similar, and then construct new faces which construct combinations of features from these similar faces [3]. Thus, the identification of an individual remains hidden to a definite amount, while the usefulness of video is preserved, as shown in figure 4. This approach is somewhat similar to k-anonymity approach.

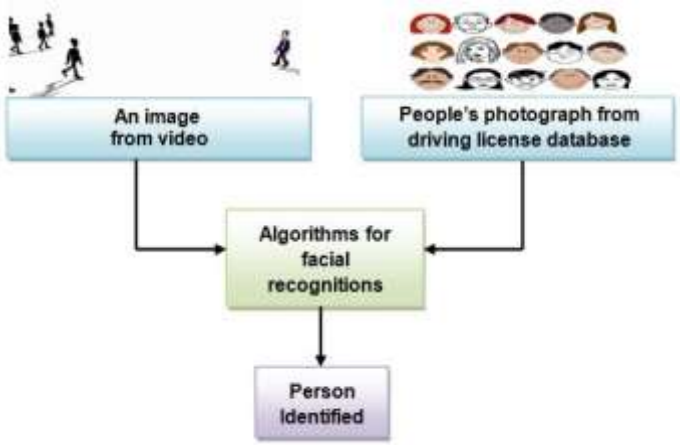

Figure 3. Privacy violated in video surveillance

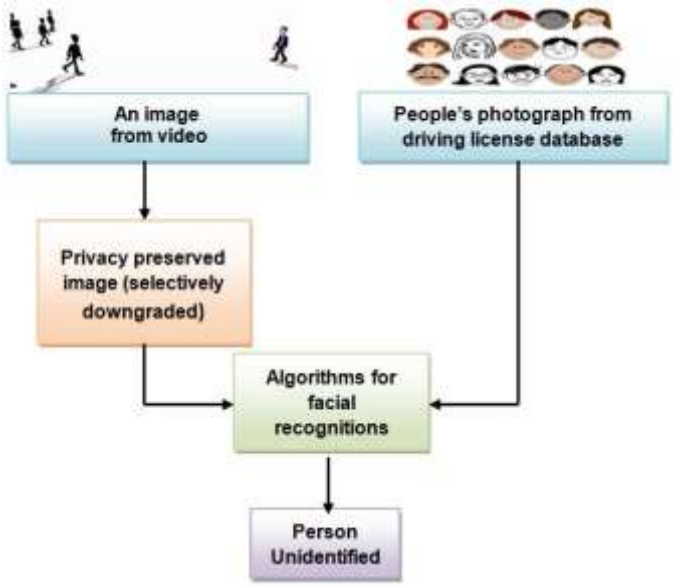

Figure 4. Privacy-preserved in video surveillance

IV. PRIVACY-PRESERVING DATA MINING METHODS

The methodologies used in privacy-preserving data mining protect the data from the theft and disclosure and in parallel maintains the originality of mining results. Taxonomy of different PPDM techniques is shown in figure 5, and their brief information are as follows [11, 12, 13, 14, 15, 16, 17, 18, 19, 20]:

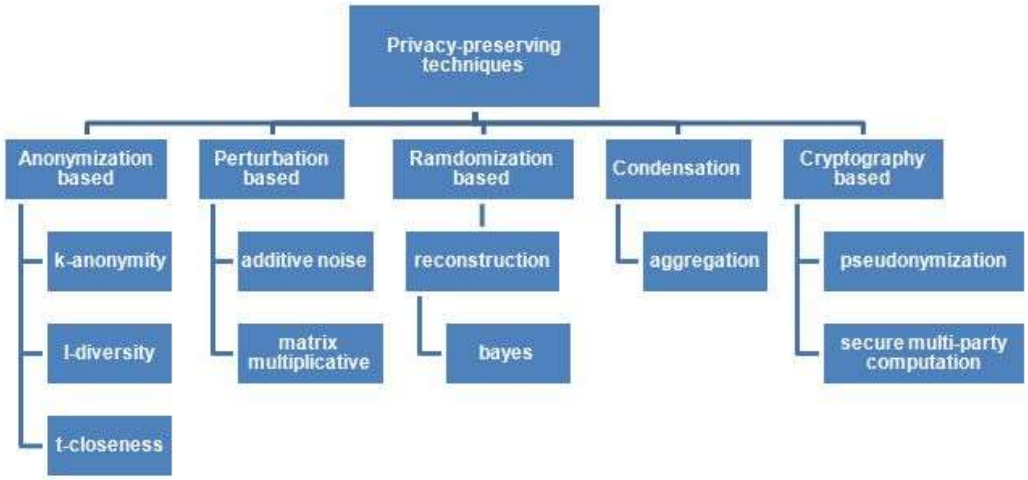

Figure 5. Taxonomy of PPDM

4.1. Anonymization
Anonymization is an approach where identity or sensitive data about record owners are to be kept unidentified, parallely assuming that sensitive data should be conserved for analysis.

4.2. Perturbation
In this technique, the original values of data set are changed with some unreal data values so that the result obtained from the perturbed data does not differ from the statistical information computed from the original dataset.

4.3. Randomization
In this technique, the data of available dataset is twisted in a manner that it proves better than a pre-defined threshold, whether the data from users includes accurate information or inaccurate one. The information acknowledged by every single user is twisted and on the other side, if there exits huge number of users, the collective information of these users contains a good quantity of accuracy.

4.4. Condensation
This approach forms, unnatural group of data in the data set, called clusters and then generates pseudo-data from the statistics of these clusters. The constraints on the cluster are defined in terms of size.

4.5. Cryptography
This method is based on a special encryption protocol known as Secure Multiparty Computation (SMC) technology. Encryption is performed on original data values. Although cryptographic techniques ensure that the transformed data is precise and secure but this approach fails to deliver when more than a few parties are involved.

V. RESEARCH DIRECTIONS
Due to its large number of applications, privacy-preserving data mining is done through several methods as described in above sections. This set of applications and methods gives way to several research opportunities. Some of which are: privacy-preserving classification and association of distributed data, optimization of existing privacy-preserving data mining using soft computing methods, privacy-preserving mining of high dimensional data, privacy-preserving big data mining, privacy-preserving data stream mining, optimizing techniques of privacy-preserving data publishing, changing the results of data mining applications to preserve privacy, etc.

VI. CONCLUSION
Privacy-preserving data mining is one of the growing fields of research these days. The main objective of privacy-preserving data mining is to protect the privacy of certain sensitive information from third parties or intruders while maintaining the usefulness of data for performing various mining tasks. This paper covers the fundamentals and basic methods of privacy-preserving data mining and provides details as well as diagrammatic representation of the prominent applications of privacy-preserving data mining. A number of key research directions in this field are also highlighted. Detailed study of the mentioned privacy-preserving data mining methods, analyzing which method is suitable for specific data mining applications and exploring the research directions will be the target for future work.

REFERENCES
[1] P.Tan, M. Steinbach, V. Kumar, Introduction to Data Mining, 1st ed., Addison-Wesley Longman Publishing, Co., Inc., 2005.
[2] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers Inc., 2005
[3] C. Aggarwal, P. Yu, "A General Survey of Privacy-Preserving Data Mining Models and Algorithms", Springer, pp. 11-52, 2008.
[4] S. Taneja, S. Khanna, S. Tiwalia, "A Review on Privacy Preserving Data mining: Techniques and research challenges", International Journal of Computer Science and Information Technologies, vol. 5, 2014.
[5] B. Fung, K. Wang, R. Chen, P. Yu, "Privacy-Preserving Data Publishing: A Survey Of Recent Developments", ACM Computing Surveys, June 2010.
[6] J. Vaidya, Christopher, W. Clifton, Y. Zhu, "Privacy Preserving Data Mining", Springer, 2006
[7] R. Aggrawal, "Privacy Preserving Data Mining", ACM SIGMOD Record, vol 29, no.2, pp. 439-450, June 2000.
[8] S. Oliveira, "Data Transformation for Privacy-Preserving Data Mining", 2005
[9] F. Ogwueleka, "Data Mining Application in credit card fraud detection system", Journal of Engineering Science and Technology, vol. 6, no.3, 2011.
[10] E. Newton, L. Sweeney, B Malin, "Preserving Privacy by De-identifying Facial Images", IEEE Transactions on Knowledge and Data Engineering, February 2005.
[11] G. Nayak, S. Devi, "A Survey On Privacy Preserving Data Mining: Approaches And Techniques", in International Journal Of Engineering Science And Technology, March-2011.
[12] H. Vaghashia, A. Ganatra, "A Survey: Privacy Preservation Techniques in Data Mining", International Journal of Computer Applications, vol. 119 , no.4, June 2015.
[13] K. Saranya, K. Premalatha, S. Rajasekar, "A Survey son Privacy Preserving Data Mining", 2nd IEEE International conference on electronics and communication system, 2015.

[14] V. Verykios, E. Bertino, I. Favino, L. Provenza, Y. Saygin and Y. Theodoridis, "State-of-the-Art in Privacy Preserving Data Mining", ACM SIGMOD Record, vol. 33, no.1, pp. 50-57, March 2004.

[15] X. Ge, J. Zhu, K. Funatsu "Privacy Preserving Data Mining", Intech, 2011.

[16] C. Aggarwal, P. Yu, "On static and dynamic methods for condensation-based privacy preserving data minig", ACM Transactions on Database Systems, vol. 33, No. 1, March 2008.

[17] J. Panackal, A. Pillai, "Privacy Preserving Data Mining: An Extensive Survey", ACEEE, Procedings of International Conference on Multimedia Processing, Communication and Information Technology, 2013.

[18] C. Aggarwal, P. Yu, "A Survey of Randomization Methods for Privacy-Preserving Data Mining", Springer, pp. 137-156, 2008.

[19] M. Keyvanpour, S. Moradi, "Classification and Evaluation the Privacy Preserving Data Mining Techniques by using a Data Modification–based Framework", International Journal on Computer Science and Engineering, vol. 3, pp. 43-48, February 2011.

[20] A. Evfimievski, "Randomization in privacy preserving data mining", ACM SIGKDD Explorations Newsletter, vol. 4, no. 2, December 2002.

[21] H. Kargupta, A. Joshi, K. Sivakumar, Y. Yesha, Data Mining Next Generation Challenges and Future Directions, Prentice-Hall of India, 2005.