Chapter

Privacy Preserving Data Mining

Esma Ergüner Özkoç

Abstract

Data mining techniques provide benefits in many areas such as medicine, sports, marketing, signal processing as well as data and network security. However, although data mining techniques used in security subjects such as intrusion detection, biometric authentication, fraud and malware classification, "privacy" has become a serious problem, especially in data mining applications that involve the collection and sharing of personal data. For these reasons, the problem of protecting privacy in the context of data mining differs from traditional data privacy protection, as data mining can act as both a friend and foe. Chapter covers the previously developed privacy preserving data mining techniques in two parts: (i) techniques proposed for input data that will be subject to data mining and (ii) techniques suggested for processed data (output of the data mining algorithms). Also presents attacks against the privacy of data mining applications. The chapter conclude with a discussion of next-generation privacy-preserving data mining applications at both the individual and organizational levels.

Keywords: privacy preserving data mining, data privacy, PPDM methods, privacy attacks, Anonymization

1. Introduction

Especially with the 2019 pandemic, in today's world where business and education life is done electronically over the internet, fast and voluminous data sharing is made with the undeniable effect of social media and unfortunately technology works against privacy. The rapid widespread use of data mining techniques in areas such as medicine, sports, marketing, signal processing has also increased the interest in privacy. The important point here is to define the boundaries of the concept of privacy and to provide a clear definition. Individuals define privacy with the phrase "keep information about me from being available to others". However, when it comes to using these personal data in a study that is considered to be well intentioned, individuals are not disturbed by this situation and do not think that their privacy is violated [1]. What is missed here is the difficulty of preventing abuse once the information is released.

Personal data is information that relates to an identified or identifiable individual. This concept consists of the components that the data pertain to a person and that this person can also be identified. Personal data is a concept that belongs to the "ego" and is handled in a wide range from names to preferences, feelings and thoughts. An identifiable person is someone who can be identified directly or indirectly, in particular by reference to an identification number or one or more factors specific to their physical, physiological, mental, economic, cultural or social identity. For this reason, the loss of the individual's control authority over these data brings about the loss of the individual's freedom, autonomy, privacy, in short, the property of being me. The main way to ensure the use of these data without harming the privacy of individuals is to remove the identifiability of the person.

Data analysis methods, including data mining, commodify data and turn it into economic value. Apart from the ethical debates about this, it is an undeniable fact that the digital environment increases the risk of losing control of all information about one's own intellectual, emotional and situational, in short, losing its autonomy and violating the informational privacy area. The main dilemma here is; the freedom in the flow of information provided by technology, the interest relationships it provides and the benefit provided by the information source is the control power required by the concept of being an individual [2].

In addition, legal regulations aiming to protect personal data are made by governments, including for what purpose (historical, statistical, commercial, scientific) data is used, how it is collected and how it should be stored. For example, the US HIPAA rules aim to protect individually identifiable health information. These are information that is a subset of health information, including demographic information collected from an individual [3]. In the EC95/46 [4] directive, the European parliament and of the council allow the use of personal data in the case of (i) if the data subject has explicitly given his permission, or (ii) the need for a result requested by the individual. This also applies to corporate privacy issues. Privacy concerns bring corporate privacy concerns with them. However, corporate privacy and individual privacy issues are not much different from each other. The disclosure of information about an organization can be considered a potential privacy breach. In this case, it involves both views to generalize to disclosure of information about a subset of data.

The point to note here is that while focusing on the disclosure of data subjects, the secrets of the data providers' organization should also be taken into account. For example, considering that data mining studies were carried out on student data of more than one university in an academic study. Although the methods used protect the privacy of the student, certain information that is specific to the university and they want to keep may be revealed. Although the personal data owned by the organizations are secured by contracts and legal regulations, information about a subset of the combined data set may reveal the identity of the data subject. The organization that owns the data set must be involved in a distributed data mining process as long as it can prevent the disclosure of the data subjects it provides and its own trade secrets.

In the literature, solutions that take data privacy into account have been proposed in data mining. A solution that ensures that no individual data is exposed can still publish information that describes the collection as a whole. This type of corporate information is often the purpose of data mining, but some results can be identified, various data hiding and suppression techniques have been developed to ensure that the data are not individually identified.

The concept of privacy can be examined under three headings as "physical–physical, mental-communicative and data privacy [5]. The main subject in this study is data privacy.

1.1 Data privacy

Data privacy can be defined as the protection of real persons, institutions and organizations (Data Subject) that need to be protected in accordance with the law and ethical rules during the life cycle of data (collecting data, processing and analyzing data, publishing and sharing data, preserving data, re-use data) [6]. In this process, for what purpose the data will be processed, with whom it will be shared, where it will be transferred, and being able to be controlled by the data subject at a transparent and controllable level are important requirements of data privacy. On the other hand, there is no exact definition of privacy, the definition can be made specific to the application.

Data controllers who need to take privacy precautions in order to prevent data breaches are assumed to be reliable and have legal obligations; stores and uses the data collected with digital applications using appropriate methods, and shares them by anonymizing when necessary. Collected data are classified into four groups [7];

- Identifiers (ID): It contains information that uniquely and directly identifies individuals such as full name and social security number.

- Quasi-identifiers (QID): Identifiers that, combined with external data, lead to the indirect identification of an individual. These attributes are non-unique data such as gender, age, and postal code.

- Sensitive attributes (SA): It contains data that is private and sensitive to individuals, such as sickness and salary.

- Insensitive attributes: It contains general and non-risky data that are not covered by other attributes.

1.2 Privacy metrics

It is not sufficient to measure privacy with a single metric because different definitions can be made for different applications and multiple parameters must be evaluated for this purpose. It is possible to examine the proposed metrics for PPDMs [8, 9] as privacy level metric and data quality metric, depending on which aspect of privacy is measured. While evaluating these metrics, they can be measured in two subgroups to evaluate the level of privacy/data quality on the input data (data criteria) and data mining results (result criteria). How secure the data is in terms of disclosure is measured by the level of privacy metrics [10]:

Bounded knowledge: The purpose here is to restrict the data with certain rules and prevent the disclosure of the information that should remain confidential. It can be transformed into limited data by adding noise to the data or by generalizing the data.

Need to know: With this metric, keeping unnecessary data away from the system prevents privacy data that will arise. It also ensures that access control (access reason and access authorization) to data.

Protected from disclosure: In order to keep the confidential data that may come out as a result of data mining, some operations (such as checking the queries) can be done on the results to provide privacy. Using the classification method to prevent the disclosure of data, which is one of the criteria for ensuring privacy, is one of the effective methods [11].

Data quality metrics: It quantifies the loss of information/benefit, and the complexity criteria that measure the efficiency and scalability of different techniques are evaluated within this scope.

2. Data mining with privacy

Privacy Protected Data Mining (PPDM) techniques have been developed to allow the extraction of information from data sets while preventing the disclosure of data subjects' identities or sensitive information. In addition, PPDM allows more than one researcher to collaborate on a dataset [11, 12]. Also PPDM can be defined as performing data mining on data sets to be obtained from databases containing sensitive and confidential information in a multilateral environment without disclosing the data of each party to other parties [13].

In order to protect privacy in data mining, statistical and cryptographic based approaches have been proposed. The vast majority of these approaches operate on original data to protect privacy. This is referred to as the natural trade-off between data quality and privacy level.

PPDM methods are being studied on to perform effective data mining by guaranteeing a certain level of privacy. Several different taxonomies have been proposed for these methods. In the literature, based on data life cycle stages (data collection, data publishing, data distribution and output of data mining) [10] or they are classified based on the method used (Anonymization based, Perturbation based, Randomization based, Condensation based and Cryptography based) [14].

In this study, PPDM approaches are examined with a simple taxonomy as methods applied to input data and processed data (output information) that is subject to data mining.

2.1 Methods applied to input Data

This section includes the methods suggested for collecting, cleaning, integration, selection and transformation phases of input data that will be subject to data mining.

Although it varies according to the application used or the state of trust to the institution collecting the data, it is recommended that the original values not be stored and used only in the conversion process in order to prevent disclosure of privacy. For example, the data collected with sensors, which are now widely used with internet of things, can be transformed at the stage it collects, randomizing the obtained values and transforming the raw data before being used in data mining.

In this section, data perturbation, randomization, suppression, data swapping, anonymity, cryptography and differential privacy methods are discussed.

2.1.1 Data perturbation

The creation of data resistant to privacy attacks can be done by perturbation significantly preserving the statistical integrity of the data [15, 16]. Randomization of the original data is widely used in data perturbation [17–19]. Another approach is the Microaggregation method [20].

In the randomization method, noise signals are added to the data with a known statistical distribution, so when data mining methods are applied, the original data distribution can be reconstructed without accessing the original data. For this, data providers first randomize their data and then transmit them to the data recipient. Then, receiving this random data, the data receiver calculates the distribution using distribution reconstruction methods.

During the data collection phase, it can be calculated independently for each data, and after the original distribution is reconstructed, the statistical properties of the data are preserved. For example; the result of the randomization of A with B is C (C = A + B) if A be the original data distribution, and B, a publicly known noise distribution independent of A. Then, A may be reconstructed with "A= C− B". However, this reconstruction process may not be successful if B has a large variance and C's sample size is not large enough. As a solution, approaches that implement the Bayes [21], or EM [22] formula can be used. While the randomization method limits data usage to the distribution of C, it requires a lot of noise to hide outliers. Because in this approach, outliers are more vulnerable to attacks when compared to values in denser regions in the data. Although this reduces the use of the data for mining purposes, it may be necessary to add too much noise to all records in the data that would result in loss of information, in order to prevent it [7].

Randomly generated values can be added to the original data with an additive or multiplicative method [23]. The aim is to ensure that noise added to individual records for privacy is non-extractable. Multiplicative Noise is more efficient than the Additive Noise method because it is more difficult to predict the original values.

With Microaggregation method, all records in the data set are first arranged in a meaningful order and then the whole set is divided into a certain number of subsets. Then, by taking the average of the value of each subset of the specified attribute, the value of that attribute of the subset is replaced with the average value. Thus, the average value of that attribute for the entire data set will not change.

Since data perturbation approaches have a negative impact on data utility and are not resistant to attacks, they are often not preferred in utility-based data models.

2.1.2 Suppression

Data Suppression technique is a technique that tries to prevent the disclosure of confidential information by replacing some values with a special value. In some cases, it is the process of deleting cell values or the entire record [24]. In this way, confidential data can be changed, rounded, generalized or mixed and made available in data mining applications [25].

An example of Suppression may be changing the age attribute in records from 28 to 35, city attribute from Glasgow to Edinburgh, or generalizing the age attribute from 28 to 25–30, and Glasgow data as Scotland. Using these methods in big data can reduce data quality and change general statistics, this may result in data becoming unusable [26]. Another problem is that information is deliberately distorted to suppression. Data providers can obtain artificial inferences that are inaccurate and serve a purpose with the reported values [27].

On the other hand, suppression should not be used when data mining requires full access to sensitive values. For sensitive information in a record, the method of limiting the identity link of a record may be preferred instead.

2.1.3 Data swapping

A technique tries to prevent the disclosure of private information by swapping values between different records.

Data swapping can be explained as each data provider scrambling data by exchanging their data with other data providers, especially in cases where there are more than one data provider. The advantage of the technique is that the data does not affect the sub-order sums, thus allowing accurate and complete collective calculations.

With this technique, as the result of data exchanges, private data can be easily exposed in the system, for this reason it is recommended to use only in safe environments. It can be used in conjunction with other methods such as k-anonymity without violating privacy definitions.

2.1.4 Cryptography

Cryptography is a technique that converts plain text to cipher text using various encryption algorithms to encode messages in a way that cannot be read. It is a method of storing and transmitting data in specific form using cryptography techniques so that only intended persons can read and process it.

In data mining applications, cryptography-based techniques are used to protect privacy during data collection and data storage [25, 28], and guarantee a very high level of data privacy [23]. Encryption is generally costly due to time and computational complexity. Hence, as the volume of data increases, the time to process on encrypted data increases and creates a potential barrier to real-time analysis [29].

Secure multiparty computing (SMC) is a special encryption protocol where, when there is more than one participating party, the interested parties learn nothing but results [30, 31]. The SMC calculation must be done carefully so that it does not reveal sensitive data, but the calculated result can enable the parties to estimate the value of sensitive data.

2.1.5 Group-based anonymization

Many privacy conversions are for creating groups between anonymous records that are converted in a group-specific manner. A number of techniques have been proposed for group anonymity in different studies, such as k-anonymity, l-diversity, and t-proximity methods. The comparison of group anonymity methods is given in **Table 1**.

2.1.5.1 k-anonymity

The k-anonymity method proposed by Samarati and Sweeney in the anonymization of data is a method of providing privacy that protects the identity of the data subject most commonly used in the publication of data [32].

The method ensures that after removing the ID attributes from the table, the QID values of at least k records in the table to be published are the same.

Since the QID attributes of each record in the table published by this method are the same as the other k-1 records, it is aimed to prevent identity disclosure.

To reduce the level of detail of the data representation, some attributes can be replaced with more general values (data swapping), some data points can be eliminated, or descriptive data can be deleted (suppression). However, while k-Anonymity provides protection against attacks on the disclosure of identities, it does not protect against attacks on disclosure of attributes. It is also more convenient to use for individual data rather than directly applying it to restrict data mining results that protect privacy. Besides, k-anonymity fully protects the privacy of users when it comes to the homogeneity of sensitive values in the data. Providing optimum k-anonymity is a problem in the NP-Hard class and approximate solutions have been proposed to avoid calculation difficulties [33].

In the literature, different studies such as k-neighbor anonymity, k-degree anonymity, cotomorphism anonymity, k-candidate anonymity and l-grouping derived

Method	Based on	Vulnerability under	Strong against
k-anonymity	Sensitive data disclosure	Homogeneity attack	record linkage only
l-diversity	Semantic similarity of sensitive data	Skewness attack	record linkage and attribute linkage
t-closeness	Distance measures	Attribute linkage attack	probabilistic attack and attribute linkage

Table 1.
Group based anonymity methods.

from the k-anonymity approach have been proposed according to the structural features of the data.

2.1.5.2 l-diversity

The l-diversity approach was proposed by Ashwin Machanavijjhala in 2007 to address the weaknesses (homogeneity attack) of the k-anonymity model [34].

This method aims to prevent the disclosure of confidential information indirectly by ensuring that each QID group has at least l well-represented sensitive value.

L-diversity only guarantees the diversity of sensitive features within each QID group, but the problem that different values may belong to the same category is not solved.

In other words, it is not resistant to attacks based on semantic similarity between values.

2.1.5.3 t-closeness

In order to balance the semantic similarities of SA attributes within each QID group, it has been proposed to solve the limitations of the l-diversity approach by guaranteeing t-closeness to each other [35].

Accordingly, in t-closeness method, the distance of the distribution of sensitive attributes in any equivalence class to the distribution of the attributes in the whole table will not exceed a threshold value (t). While the t-closeness approach provides protection against disclosure of attributes, it cannot protect against disclosure of identities. In addition, it limits the usefulness of the information disclosed however, by setting the t-threshold in applications, it can exchange benefit and privacy.

In the protection of privacy, t-proximity and k-anonymity methods are used together to protect against attacks on identity disclosure and quality [36].

2.2 Methods applied to processed Data

The outputs of data mining algorithms can disclose information without open access to the original data set. Sensitive information can be accessed through studies on the results. For this reason, data mining output must also protect privacy.

2.2.1 Query auditing and inference control

This method is examined as query inference control and query auditing. In the query inference control, the input data or the output of the query is controlled. In t Query auditing, the queries made on the outputs obtained by data mining are audited. If the audited query enables the disclosure of confidential data, the query request is denied. Although it limits data mining, it plays an active role in ensuring privacy. Query auditing can be done online or offline. Since queries and query results are already known in offline control, it is evaluated whether the results violate privacy. In online auditing, since the queries are not known, privacy metrics are carried out simultaneously during the execution of the query. This method is examined within the scope of statistical database security.

2.2.2 Differential privacy

k-anonymity, l-diversity and t-closeness approaches are holistic approaches that try to protect the whole data privacy. In some cases, there is a need to protect the privacy of data at the record level. For this reason, differential privacy approach has been proposed by Dwork to protect the privacy of database query results [37].

With this model, the attacks that may occur between sending database queries and responding to the query are targeted. Failure to distinguish from which database the answer of the same query, made in more than one database, is returned will prevent the disclosure of the existence of a single record between databases.

In addition, when querying output data, it can be ensured that the query results obtain approximate values with the database approach technique. Also, it is recommended to keep the data in the system mixed during the execution of queries, just like the data collection phases to protect data privacy.

2.2.3 Association rule hiding

In data mining, it is one of the most frequently used methods of Association Rules to reveal the nature of interesting associations between binary variables. During data mining, some rules may explicitly disclose private information about the data subject (individual or group).

Unnecessary and information-leaking rules may occur in some relationships. The aim of the Association rule hiding technique first proposed by Atallah [38] is to protect privacy by hiding all sensitive rules. The weakness with this technique is that a significant number of insensitive rules can be hidden incorrectly [39].

3. Attacks against privacy

In this section, the common types of attacks that lead to the development of the methods given above and lead to privacy violations are summarized [6].

3.1 Semantic similarity attacks

Attacks that are made by making use of the intuitive similarity of sensitive attribute values within anonymous groups.

In this case, it is not sufficient for the sensitive attribute values to be different from each other in terms of protecting privacy [40]. This attack can be prevented by calculating the similarities of sensitive attributes in the same anonymous group and by providing solutions to include similar sensitive attribute values in different groups.

3.2 Background knowledge attacks

Background knowledge is non-sensitive information that can be obtained from data published by different organizations, social networks and media even by using social engineering methods. Background knowledge obtained by attacker's causes privacy attacks and breaches.

Data subject's privacy violation occurs as a result of associating background knowledge with other records using data binding methods [41].

In addition, when information obtained from data owners through requests such as promotion, campaign, research, etc. is associated with background information, it is not even possible that it will not cause a violation of privacy.

3.3 Homogeneity attacks

In cases where all or most of the sensitive attributes in the groups included in the anonymous tables are similar, the privacy of data owners is at risk of violation.

In order to prevent homogeneity attacks, it is necessary to prevent similar sensitive attributes within the groups in the anonymous table from being in the same group or to reproduce heterogeneous records by diluting the homogeneous attributes with the record duplication approach [34].

3.4 Skewness attacks

The statistical distribution of sensitive attribute values in published or shared anonymous data sets can lead to the success of skewness attacks against privacy. The distortion in the general distribution of sensitive attributes occurs when these values are too dominant and anonymous data sets become vulnerable [35].

3.5 De-Finetti attacks

It has been shown that with theoretical and experimental methods, interchangeability concepts and inferences about privacy can be made with Definetti's theorem [42]. The fact that the people who want to carry out this attack do not need extensive background knowledge makes this attack attractive. An attacker can perform an attack using machine-learning techniques on non-sensitive attributes in the dataset.

3.6 Minimality attacks

The fact that the information about which data anonymization algorithm is used in the data mining application is public is also considered as a privacy vulnerability [43]. It is based on the principle that changes on data should remain at minimum level in anonymization processes and should not be overly anonymized.

3.7 Temporal attack

Publicly declaring previously published generalized data over time causes this attack. For this reason, previously published tables should be used and new records that may cause data disclosure should not be shared [44].

4. Discussion

The fact that the digitalization process has become mandatory all over the world with Covid-19 pandemic has accelerated the data flow. It has become even more important to collect the necessary data, analyze it correctly and reveal reliable information. This situation has triggered the use of data mining methods to increase productivity and provide high quality products/services in almost all sectors. While applying data mining methods, it is obvious that if privacy is not taken into consideration during the data life cycle, irreversible damages will occur for individuals/institutions and organizations.

In order to increase the access and benefits of data mining technology, before applying PPDM techniques, "privacy" should be defined precisely, measurement metrics should be determined and the results obtained should be evaluated with these metrics. For this reason, this study primarily focused on the definition of privacy. The term privacy is quite extensive and does not have a standard definition. It is quite challenging in measuring privacy, as there is no standard privacy definition. Some measurement metrics are mentioned in this chapter, but metrics are usually determined by application. The lack of a standard privacy measurement metric also make challenging the comparison and evaluation of the developed PPDM techniques.

In the age of digital and online business, privacy protection needs to be done at the individual and organizational levels. Privacy protection at the individual level depends on person who is influenced by religious beliefs, community norms and culture. For this reason, the concept of personalized privacy, which allows individuals to have a certain level of control over their data, has been proposed. However, it has been observed that there are difficulties in implementing personalized privacy, as people think that compromising their privacy for applications they think is well-intentioned will not damage. Therefore, in the context of personalized privacy, new solutions are required for the trade-off between privacy and utility.

To effectively protect organizational level data privacy [7]; Policy makers in organizations should support privacy-enhancing technical architectures/models to securely collect, analyze and share data. Laws, regulations and fundamental principles regarding privacy should be analyzed by organizations. It is necessary for organizations to include the data owners in their assessment of privacy and security practices. Data owners should involve the whole process about what data is collected, how it is analyzed and for what purpose it is used. In addition, they should have the right to correct personal data in order to avoid negative consequences of incorrect data. Organizations should employ data privacy analysts, data security scientists, and data privacy architects who can develop data mining applications securely.

From a technical point of view, methods that protect confidentiality in data analytics are still in their infancy. Although studies continue by different scientific communities such as cryptography, database management and data mining, an interdisciplinary study should be conducted on PPDM. For example, the difficulties encountered in this process should also be addressed from a legal perspective. Thus, a better roadmap for next-generation privacy-preserving data mining design can be developed by academic researchers and industrial practitioners.

5. Conclusion

Businesses and even governments collect data through many digital platforms (social media, e-health, e-commerce, entertainment, e-government etc.) they use to serve their customers/citizens. The data collected can be sensitive data and this data can be stored, analyzed and, in good probability, anonymized and shared with others. In studies where data is used at any stage of the life cycle, regardless of the purpose, it is necessary to explain a privacy permission and the reason why the data should be accessed. Privacy Preserving Data Mining (PPDM) techniques are being developed to allow information to be extracted from data without disclosing sensitive information.

There is no single optimal PPDM technique for any stage of the data lifecycle. The PPDM technique to be applied varies according to the application requirements, such as the desired privacy level, data size and volume, tolerable information loss level, transaction complexity, etc. Because different application areas have different rules, assumptions and requirements regarding privacy.

In this chapter, the previously proposed PPDM techniques are examined in two sections. First section includes the methods suggested for collecting, cleaning, integration, selection and transformation phases of input data that will be subject to data mining and second section covers methods applied to processed data. Finally, attacks against the privacy of data mining applications are given in this chapter.

Author details

Esma Ergüner Özkoç
Başkent University, Ankara, Turkey

*Address all correspondence to: eeozkoc@baskent.edu.tr

References

[1] Clifton C, Kantarcioglu M, Vaidya J, Defining privacy for data mining. In National science foundation workshop on next generation data mining. 2002; Vol. 1, No. 26, p. 1

[2] İzgi M. C, The concept of privacy in the context of personal health data. Türkiye Biyoetik Dergisi, 2014. (S 1), 1

[3] Centers for Disease Control and Prevention. HIPAA privacy rule and public health. Guidance from CDC and the US Department of Health and Human Services. MMWR: Morbidity and mortality weekly report, 200352(Suppl 1), 1-17.

[4] Data P, Directive 95/46/EC of the European parliament and of the council on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Official Journal L, 1995; 281(23/11), 0031-0050.

[5] Belsey A, Chadwick, R. Ethical issues in journalism and the media. Routledge. (Eds.) 2002

[6] Vural Y, Veri Mahremiyeti: Saldırılar, Korunma Ve Yeni Bir Çözüm Önerisi. Uluslararası Bilgi Güvenliği Mühendisliği Dergisi, 4(2), 21-34.

[7] Pramanik M. I, Lau R. Y, Hossain M. S, Rahoman M. M, Debnath S. K, Rashed, M. G., Uddin M. Z.,. Privacy preserving big data analytics: A critical analysis of state-of-the-art. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2021; 11(1), e1387.

[8] Bertino E, Lin D, Jiang W, A survey of quantification of privacy preserving data mining algorithms, in Privacy-Preserving Data Mining. New York, NY, USA: Springer, 2008, pp. 183-205.

[9] Dua S, Du X, Data Mining and Machine Learning in Cybersecurity. Boca Raton, FL, USA: CRC Press, 2011.

[10] Mendes R, Vilela J. P, Privacy-preserving data mining: methods, metrics, and applications. IEEE Access, 2017; 5, 10562-10582.

[11] Vaidya J, Clifton C, Privacy-preserving data mining: Why, how, and when. IEEE Security & Privacy, 2004; 2(6), 19-27.

[12] Nayak G, Devi S, A survey on privacy preserving data mining: approaches and techniques. International Journal of Engineering Science and Technology, 2011; 3(3), 2127-2133.

[13] Lindell Y, Pinkas B, Privacy Preserving Data Mining, In: Proceedings of the 20th Annual International Cryptology Conference, 2000; California, USA, 36- 53

[14] Rathod S, Patel D, Survey on Privacy Preserving Data Mining Techniques. International Journal of Engineering Research & Technology (IJERT) 2020; Vol. 9 Issue 06

[15] Hong T. P, Yang K. T, Lin C. W, Wang S. L, Evolutionary privacy-preserving data mining. In: Proceedings of the World Automation Congress 2010; (pp. 1-7). IEEE.

[16] Qi X, Zong M, An overview of privacy preserving data mining. Procedia Environmental Sciences, 2011; 12, 1341-1347

[17] Muralidhar K, Sarathy R, A theoretical basis for perturbation methods. Statistics and Computing, 2003; 13(4), 329-335

[18] Evfimievski A, Randomization in privacy preserving data mining. ACM

Sigkdd Explorations Newsletter, 2002; 4(2), 43-48

[19] Kargupta H, Datta S, Wang Q, Sivakumar K, On the privacy preserving properties of random data perturbation techniques. In: Proceedings of the Third IEEE international conference on data mining 2003; (pp. 99-106). IEEE.

[20] Domingo-Ferrer J, Torra V, Ordinal, continuous and heterogeneous k-anonymity through microaggregation. Data Mining and Knowledge Discovery, 2005; 11(2), 195-212

[21] Agrawal R, Srikant R, Privacy-preserving data mining. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data 2000; (pp. 439-450).

[22] Agrawal D, Aggarwal C. C, On the design and quantification of privacy preserving data mining algorithms. In Proceedings of the Twen tieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems 2001; pp. 247-255.

[23] Niranjan A, Nitish A, Security in Data Mining-A Comprehensive Survey. Global Journal of Computer Science and Technology 2017

[24] Oliveira S, Zaiane O, Data perturbation by rotation for privacy-preserving clustering, Technical Report 2004.

[25] Verykios V. S, Bertino E, Fovino I. N, Provenza L. P, Saygin Y, Theodoridis Y, State-of-the-art in privacy preserving data min ing. ACM SIGMOD Record, 2004; 33, 50-57.

[26] Aggarwal C. C, On randomization, public information and the curse of dimensionality. In: Proceedings of the IEEE 23rd International Conference on Data Engineering; Istanbul, Turkey, 2007, pp. 136-145.

[27] Zhu D, Li X. B, Wu S, Identity disclosure protection: A data reconstruction approach for privacy-preserving data mining. Decision Support Systems, 2009; 48, 133-140.

[28] Yang Y, Zheng X, Guo W, Liu X, Chang V, Privacy-preserving smart IoT-based healthcare big data storage and self-adaptive access control system. Information Sciences, 2019; 479, 567-592.

[29] Lu, R., Zhu, H., Liu, X., Liu, J. K., & Shao, J. (2014). Toward efficient and privacy-preserving computing in big data era. IEEE Network, 28, 46-50.

[30] Yao A. C, How to generate and exchange secrets. In: Proceedings of the 27th IEEE Symposium on Foundations of Computer Science, 1986; 162-167. IEEE

[31] Goldreich O, Micali S, Wigderson A, How to play any mental game - a completeness theorem for protocols with honest majority. . In: Proceedings of the 19th ACM Symposium on the Theory of Computing, 1987; 218-229.

[32] Sweeney L, k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002; 10(05), 557-570.

[33] Samarati P, Sweeney L, Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression, SRI International, Technical Report, 1998; SRI-CSL-98-04

[34] Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M, ℓ-Diversity: Privacy beyond k-anonymity, In: Proceedings of the The 22nd International Conference on Data Engineering, 2006; Atlanta, USA,

[35] Li N, Li T, Venkatasubramanian S, t-Closeness: Privacy beyond

k-anonymity and ℓ-diversity, In: Proceedings of the International Conference on Data Engineering (ICDE), İstanbul, Turkey, 2007; 106-115

[36] Rubner Y, Tomasi C, Guibas L. J, The earth mover's distance as a metric for image retrieval. International journal of computer vision, 2000; 40(2), 99-121.

[37] Dwork C, Differential privacy: A survey of results. In: Proceedings of the International conference on theory and applications of models of computation Springer, Berlin, Heidelberg. 2008; (pp. 1-19).

[38] Atallah M, Bertino E, Elmagarmid A, Ibrahim M, Verykios V, Disclosure limitation of sensitive rules. In Knowledge and Data Engineering Exchange Workshop (KDEX'99), 1999; 25-32.

[39] Evfimievski A, Srikant R, Agrawal R, Gehrke J, Privacy preserving mining of association rules. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2002; 217-228.

[40] Wang H, Han J, Wang J, Wang L, (l, e)- Diversity - A Privacy Preserving Model to Resist Semantic Similarity Attack, Journal of Computers, 2014; 59-64

[41] Chen B. C, LeFevre K, Ramakrishnan R, Privacy skyline: Privacy with multidimensional adversarial knowledge. University of Wisconsin-Madison Department of Computer Sciences. 2007

[42] Kifer D, Attacks on privacy and deFinetti's theorem", In: Proceedings of the ACM SIGMOD International Conference on Management of data, Rhode Island, ABD, 2009; 127-138, 2009

[43] Wong R. C. W, Fu A. W. C, Wang K, Pei J, Minimality attack in privacy preserving data publishing. In: Proceedings of the 33rd international conference on Very large data bases 2007; (pp. 543-554).

[44] Sanjita B. R, Nipunika A, Desai R, Privacy Preserving In Data Mining, Journal of Emerging Technologies and Innovative Research 2019; vol6 Issue 5