Journal of Artificial Intelligence Practice (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/jaip.2023.060405
ISSN 2371-8412 Vol. 6 Num. 4

Privacy Enhancement with Perturbation Method for Multidimensional Grid

İlker İlter[1,a,*], Safiye Turgay[1,b]

[1]*Department of Industrial Engineering, Sakarya University, Sakarya, Turkey*
[a]*ilkerilter54@hotmail.com*, [b]*safiyeturgay2000@yahoo.com*
[*]*Corresponding author*

Keywords: Privacy, Data mining protecting privacy, Big data, Multi-dimensional grid, Privacy enhancement, Perturbation approach

Abstract: With the development of technology, the use of big data is spreading at an increasing rate. The issues of storing, analysing and securing data have brought along the methods that need to be developed. Ensuring data privacy and data security is the case of partial separation and processing of data with the perturbation method of data with the block chain approach. Within the scope of this study, data analysis performed using normalization, geometric rotation, linear regression and scalar data multiplication and comparative classification in precision data mining.

1. Introduction

Considering the data security structure with the system model proposed in this study, it allowed the analysis of big data in a reliable environment and the perturbation approach [1,2] to ensure the security status of the data before the analysis and to obtain meaningful results from the data as a result of the analysis. The effectiveness of the proposed method used in the Titanic dataset pulled from kaggle.com. The proposed perturbation method, which provides data confidentiality, applied on this data set. In the process of proving the accuracy of the proposed method, decision tree, logistic regression, Naive Bayes, K Neighbours' Classifier, Random Forest, Neural Network, Support Vector Machine and XGboost methods applied. It demonstrated the superiority of the proposed random perturbation method in terms of both classification accuracy and data privacy, by protecting data privacy, predicting data states, I/O states, and independent component analysis. The structure of predicting attack types and taking countermeasures carried out by independent component analysis. Thanks to data mining, the relationships between attributes in multivariate data structures examined and meaningful results drawn. The effect on data concentration analysed with the proposed method.

The remainder of this article organized in the following format. It shows the literature review in the second section, the proposed methodology in the 3rd section and the analysis results in the 4th section. The accuracy of the classification, attack situations and flexible structures, time dimension and scaling situations of the data discussed during the analysis process. In the 5th section, the obtained analysis gives information about the evaluated results and the work that can be done in the future.

2. Literature Survey

The widespread use of big data together with the Internet of Things (IoT) brings the concept of data privacy to the fore in all sectors from manufacturing to health and justice to banking. Data mining is the process of discovering interesting and previously unknown information from big data. Therefore, the necessity of ensuring data confidentiality in the process of data processing also arises. In order to monitor the correlation and autocorrelation structure of multivariate flows efficiently and effectively, and to protect data security, measures taken against attacks against noise addition and fundamental components.

The protection of data privacy and the realization of data mining analysis will provide data analysis in a healthier and more effective environment. The perturbation method was preferred in ensuring the data confidentiality process, and the pre-analysis process applied, taking into account the multidimensional structure of the data before the analysis. For the data privacy process, a precautionary pre-process is required [3,5]. The issue of data privacy has become an important issue in recent days, with the increase in its size and usage area. Traditional techniques such as perturbation, generalization and sampling used to ensure the confidentiality of data [4-7]. It is aimed to clean the database, to protect the information in the data group handled in dynamic data analysis [8], and not to experience any deterioration in data quality during the confidentiality process [9]. Association rules are derived by considering the factors of loss of confidentiality, loss of information [10], cloaking error and database diversity [11,12]. Finally, the key value updated in the proposed study and the suitability, success and failure scenarios taken into account [13]. With the use of software-cantered and virtualized approaches in 5G networks, it also brings security and privacy problems. The development of data flow mining by designing efficient security protocols and researchers have targeted writing algorithms in recent years. In order to protect the data in the perturbation analysis process, it ensured that data corruption prevented by using the correlation coefficient together with the differential privacy issue [14-16]. At the same time, data analysis performed by evaluating the neighbourhood matrix [17]. Association rules are the process of obtaining the rules obtained through correlation in the data [18]. Time series used in data flow analysis and the correlation coefficient should be taken into account in the data analysis process. Here, the short-term collection of event types and temporal correlation without loss of information in event streams discussed [19-23]. In the data analysis process, the pre-processing phase, the analysis phase and the noise-mining phase are performed [24-29]. In this method, random noise added to the privacy sensitive data using a known distribution before the data analysis process. Then, an approximation to the original data distribution created from the distorted data. It uses the reconstructed distribution for post-analysis. Due to the addition of noise, loss of information and protection of confidentiality is always a trade-off in perturbation-based approaches [30].

3. Perturbation Analysis for Multi-Dimensional Grid

In the data perturbation process, input and output perturbation, noise addition and rule hiding obtained by adding or multiplying the noise to the data. Multidimensional perturbation features diversified into condensation, random rotation, geometric perturbation, hybrid perturbation, and multidimensional perturbation. One of the most basic purposes in the data privacy model is the protection of private information. In this process, the system may be vulnerable to attacks against minimalist, composition and foreground information situations. In this regard, it tried to prevent data leakage by applying the local differential privacy approach with the differential privacy method and the perturbation method in the local differential approach.

If the definition of data mining is re-examined, it is the process of extracting competent information from previously unidentifiable data. At the same time, among the data mining tasks, the

availability of the dataset should be close to the original data, while protecting the privacy of the data of individuals and institutions arises (in Fig.1).

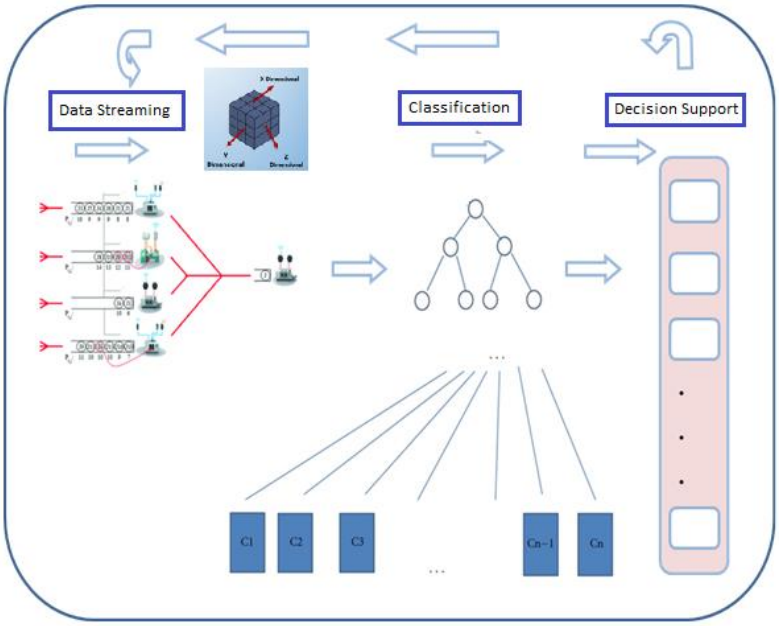

Figure 1: Decision support model considering data privacy analysis

During the analysis of dynamic data in the protection of data confidentiality, a covariance matrix created for each bundle (group) of the data bundles. Geometric rotation applied to these matrices, and the rotated bundles then combined. The bundles then randomly shuffled and released. By applying the perturbation process, the data made ready for analysis, after the process is completed, normalization performed again, and the rules obtained as result of the analysis related to the system, thus contributing to the decision support system. The grouping process especially used to increase the effect of perturbation. The rotation operation applied with a rotation operation equal to the unit matrix. The product of a vector and the identity matrix results in the same vector producing zero perturbations on the initial vector.

Data corruption is the technique of protecting the confidentiality of records, keeping the data values in the database without destroying the meaningfulness and the relationship between the variables. There are various variants of the applied perturbation method, these can counted as: total perturbation, random rotation, geometric perturbation, micro-aggregation and data condensation.

By applying the perturbation method before the data mining analysis process, it aimed to move the data to a more reliable environment. This change made in the original data will result in a more reliable environment and a result value closer to the original data after the analysis. Confidentiality of data provides in a much better way with batch processes.

$$P(G_i)\times (G_i)^T=P(G_i)^T\times P(G_i)=1 \quad (1)$$

$$C(G_i)= P(G_i) \times \Delta(G_i) \times P(G_i)^T \quad (2)$$

In the simplest sense, it expressed as adding a value to the data at a certain coefficient as a perturbation operation.

$$x_i' = x_i + r_i \quad (3)$$

Here, the rendered x_i' refers to the original, i.e. the noisy data added to the original value. The

multidimensional perturbation method applied because the relationship structures of the data and features with each other taken into account. In order to prevent loss of information in the study and to perform the analysis of the data more accurately, the analysis process carried out in large structures and together with groups. This process also expressed as masking the dataset. Here is the process of normalizing the data, rotating the data, with a certain slope and at the same time taking the projection of the data. Data security provided at a higher level by the random rotation of data. The mathematical representation of the rotation process is;

$$g(X) \quad = RX \tag{4}$$

In the structure expressed with the rotation matrix, the features also expressed with g(X). Ensuring data security is possible with some transformations that performed in the data set. In this context, the security of the data also ensured with the transformations performed on the original data set. The proposed method can categorized by applying it to the original centralized and distributed scenarios. Along with the method discussed, it is also important to transform the data back to the original after the process. Perturbation method can considered in two categories as one-dimensional and multi-dimensional. Since the perturbation method, which considered as one-dimensional, only deals with one attribute, the result of the process may be the case of data drift or deterioration of the data properties. However, the random perturbation method includes a four-step process.

3.1. Evaluation Criteria

In this study, the evaluation criteria of the data used in data mining can used after the analysis in the last step. In the final assessment process, precision was the F1 score; the area under the ROC curve used as evaluation criteria. It focuses on the degree of confidentiality of data and reducing data usage loss. The different measurement criteria used here are the confidentiality of the data; value difference (DF); perturbation of the data matrix; Even after perturbation, some of the data items may not change their importance relative to other data (RP), in this case it is expressed as the percentage of the data items' status (RK). In short, it tests whether the importance of the element is lost in the perturbation process. We can express with $Rn_j^{i,}$ whether the element preserves its degree in the perturbation process. Testing whether the mean value of each feature changes after perturbation expressed as the change in the mean value of the features (CP). The RAV function represents the rank of the mean value. CK, on the other hand, refers to the features that can maintain the corresponding order of the mean value after perturbation. Cn_j^i represents changes in the average value of attributes.

3.1.1 Security

It is the comparison of the degree of closeness of the feature to the original data in the degree of confidentiality measurement.

$$Data\ Security = \frac{Variance(x_i - x_i')}{Variance(x_i)} \tag{5}$$

3.1.2 Value Difference (DF)

Represents changes to the data.

$$DF = \frac{\|A - A'\|_F}{\|A\|_F} \tag{6}$$

3.1.3 Perturbation (PA)

It is the process of changing the order of the data in perturbation analysis. The variation of the mean order of the features expressed depending on the degree-based values. R_j^i denotes the state of the data matrix at rest in $R_j^{i'}$, taking into account the order of nm in the original data matrix.

If we consider the data set as D_{mn};

$$A = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left|R_j^i - R_j^{i'}\right|}{nm} \tag{7}$$

3.1.4 Perturbation Percentage Change (PS)

It refers to the process of ordering features after perturbation. It also shows whether an element maintains its degree during the perturbation process. After perturbation, each of the features can maintain their respective order. If the percentage display format of this data is:

$$PS = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} Rn_j^i}{nm} \tag{8}$$

3.1.5 Ranking Change in Mean Value After Perturbation (CP)

Shows the rank change in the mean value of the features after perturbation. . R_{AVi} refers to the rank of the mean value. On the other hand, CK indicates features that can preserve their respective order after perturbation.

$$CP = \frac{\sum_{i=1}^{n}\left|R_{AVi} - R'_{AVi}\right|}{n} \tag{9}$$

$$CK = \frac{\sum_{i=1}^{n} Cn^i}{n} \tag{10}$$

4. Case Study

The Titanic dataset chosen for analysis tasks is multivariate and has different dimensions. The dataset contains only the numeric attributes instead of the class attribute. Table 1 shows a detailed overview of the dataset used. The Titanic dataset is a dataset containing some of the characteristics of the passengers who survived and died during the voyage of the Titanic ship. Titanic dataset data. The dataset contains 1310 records and the total number of records without missing values is 1043. Records with missing values deleted before the data set used in this study. The number of features in the data set is 14, 7 numerical and 7 categorical. Descriptive attributes in this data set are: name, pclass, sex, age, ticket, sibsp, fare, parch, cabin, embarked, boat, body, home dest, survived. A detailed presentation of the Titanic dataset shown in Table X. Name, ticket, cabin, home.dest have been removed from the data set because they are attributes with descriptive properties (in Table 1). The boat and body variables, which are also included in the data set and contain too much missing data, were also removed from the data set. Thus, the number of attributes in the data set to be protected and classified has been reduced to 8. These variables sex, age, pclass, sibsp, fare, parch, embarked, survived were used in the study.

Table 1: Titanic data set

Variables	Data Field Descriptions	Data Types	Data Ranges
sex	Sex	categorical	male.female
age	Age in years	numerical	[0.1667-80]
pclass	Ticket class	numerical	[1-3]
sibsp	Of siblings / spouses aboard the Titanic	numerical	[0-8]
fare	Passenger fare	numerical	[0-512329]
parch	Of parents / children aboard the Titanic	numerical	[0-9]
embarked	Port of Embarkation	categorical	S,Q,C
survived	Survival	numerical	0-1
name	Passenger name	categorical	AllenMiss.Elisabeth Watson
ticket	Ticket number	categorical	113781-24160
cabin	Cabin number	categorical	B5,C22,C23,C24 vs.
boat	Lifeboat(if survived)	categorical	[1-1170]
body	Body number(if did not survive and body was recovered)	numerical	[1-328]
home dest	Home and destination of the passenger	categorical	Montreal, PQ / Chesterville, ON

In the application, open source Anaconda, Jupyter Notebook, Python and Python Numpy, Pandas, Seaborn, Matplotlib, Scikit-learn libraries used. Anaconda Navigator is an open source data science and machine-learning platform. Anaconda parts on the Python programming language and includes many popular libraries to support Python-based data analytics and machine learning projects. Other advantages of Anaconda are that it offers useful tools in data science and machine learning projects such as creating virtual environments, package management, environment management, reporting and sharing. Jupyter Notebook development tool in Anaconda Navigator used in the study. Jupyter Notebook is an open source project and a popular data science, machine learning and scientific computing environment. Jupyter Notebook contains two different types of cells called code cells and text cells. You can write code in Python code or other programming languages in code cells and run it directly. . Jupyter Notebook interface version 4.9.2 used in the study. Scikit-learn also supports key machine learning processes such as model selection, model evaluation, and model optimization. It offers a user-friendly interface in easily perform operations such as cross-validation, hyper-parameter optimization, model performance evaluation, model selection and model tuning. Python library with version 1.0.2 Scikit-learn used in the study (in Figure 2).

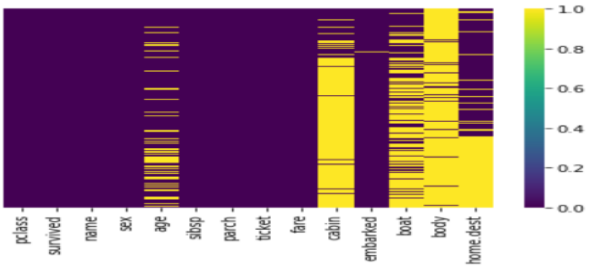

Figure 2: Variables and data density

Age and pclass variables were normalized in the Titanic data set. The results were compared in Figure 3 X1 by applying the cross validation approach to the data that was normalized and the data that was not normalized.

	Algorithm	Accuracy Mean	Accuracy Std
0	Decision Tree	0.693031	0.055407
1	Logistic Regression	0.756456	0.074582
2	Naive Bayes	0.655989	0.167572
3	K Neighbors Classifer	0.709496	0.084796
4	Random Forest	0.733425	0.092685
5	Neural Network	0.751667	0.103950
6	Support Vector Machine	0.757262	0.086035
7	XGboost	0.744844	0.077289

Figure 3: Machine learning accuracy results on Titanic data set

Changes in data between before and after perturbation were tested by applying various machine learning algorithms. The algorithms with the highest validity average obtained after this test were the support vector machine with a value of 0.757262, followed by the logistic regression method with a value of 0.756456, the neural network with a value of 0.751667, and then the XGboost method and the 0.744844 method. The method with the lowest value is Naive Bayes and then the Decision tree method with a value of 0.693031. When we look at the validity standard deviation values, the decision tree, then the logistic regression and then the XGboost method with the value of 0.077289 come. Next comes the K Neighbours classifier method and then the support vector machine method (in Fig.3). As a result of the analysis performed with the confusion matrix, the negative-negative value of the decision tree data is 116, the positive-positive value is 69, the negative-negative value of the logistic regression is 131, the positive-positive value is 77, the negative-negative value is 124 for Naive Bayes, and the number of positive-positive is 78. , 133 positives positive for KNeighboursClass, 64 negative negatives, 135 positives positive for the random forest method, 68 number of negative negatives, 140 positives positive for neural network, 70 negatives-negatives, 132 positives negatives for support vector machine, 73 as obtained. In this context, the number of correct values is highest with 210 in neural network and support vector machine methods, then logistic regression with 208, random forest with 203 and naive Bayes with 202. We see that it comes with 197 (in Figure 4).

In Figure 5, Decision tree, logistic regression, Naive Bayes, K Neighbours, random forest and support vector machine methods applied to Classifier Validity analysis result and standard deviation raw data of cross validation for 100 registered data, data rotated 45 degrees, processing results applied to 45 degrees noisy data are included. At the same time as the classification results, the confidentiality degree was given to the data with 45 degrees noisy and noiseless rotation, and also the attack resistance of the data was determined by testing the pre- and post-process similarity coefficients of the data.

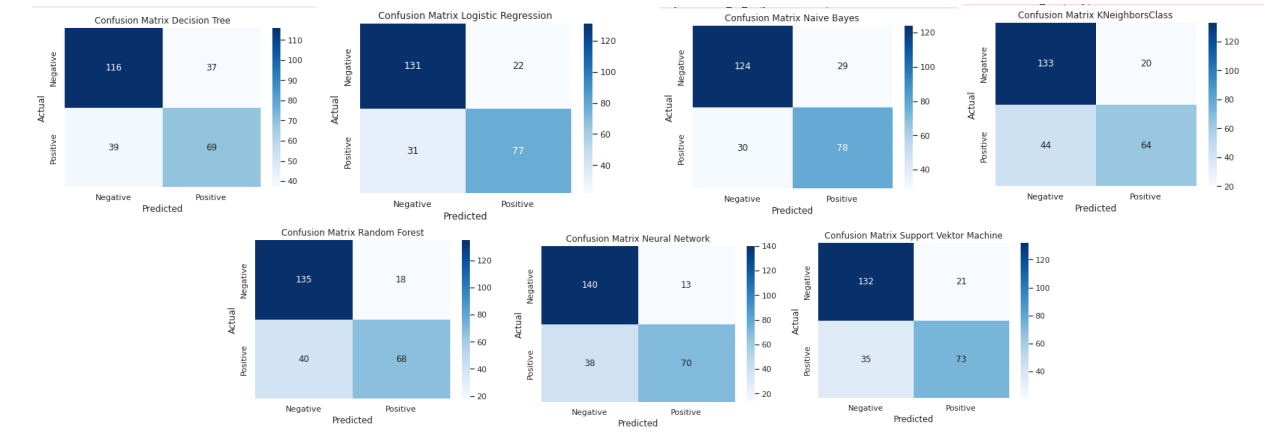

Figure 4: Confusion matrix results on Titanic data set

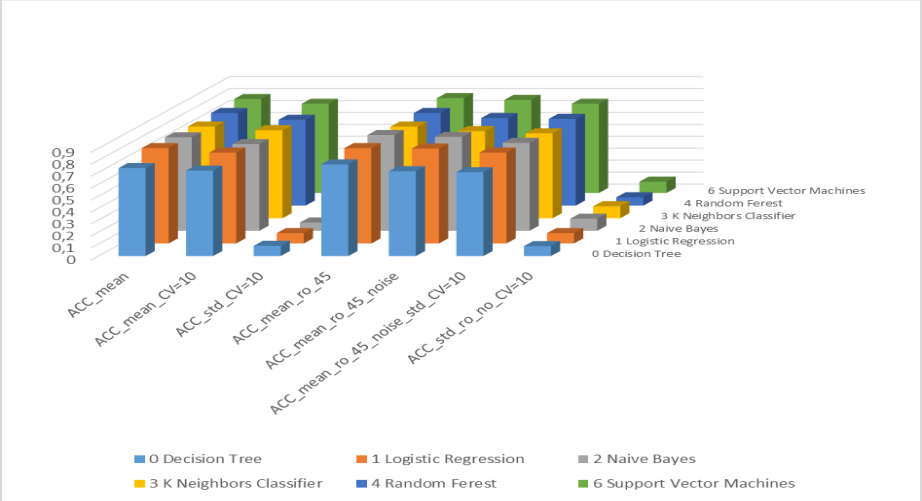

Figure 5: Analysis Results

5. Conclusion

There are singular value decomposition, polar decomposition, etc., as long as the decomposition that create a submatrix with the properties with rotation matrix. Other decomposition methods used to construct the rotation matrix of a particular group. Privasing data security and examining test parameters were discussed in this study. Data corruption is the technique of protecting the confidentiality of records, keeping the data values in the database without destroying the meaningfulness and the relationship between the variables. There are various variants of the applied perturbation method, these can be counted as: total perturbation, random rotation, geometric perturbation, micro-aggregation and data condensation.

References

[1] Chamikara M.A.P., Bertok P., Liu D., Camtepe S., Khalil I., Efficient data perturbation for privacy preserving and accurate data stream mining, Pervasive and Mobile Computing, 48 (2018) 1-19.
[2] Suma B., Shobha G., Fractional salp swarm algorithm: An association rule based privacy-preserving strategy for data sanitization, Journal of Information Security and Applications, 68 (2022) 103224.
[3] Afaq A., Haider N., Baig M.Z., Khan K.S., Imran M.,Razzak I., Machine learning for 5G secuiry: Architecture, recent advances, and challenges, Ad Hoc Networks 123 (2021) 102667
[4] Huang H., Yan Z., Tang X., XiaoF., Li Q., Differential privacy protection scheme based on community density aggregation and matrix perturbation, Information Sciences 615(2022) 167-190.
[5] Kaosar M.G., Paulet R.P., Yi X., Fully homomorphic encrtption based two-party association rule mining, Data & Knowledge Engineering, 76-78 (2012), 1-15
[6] Kulkarni Y., Jagdale B., Sugave S.R., Optimized key generation-based privacy preserving data mining model for secure data publishing, Advances in Engineering Software, 175 (2023) 103332.
[7] Wu T., Wang X., Qiao S., Xian X., Liu Y., Zhang L., Small perturbations are enough: Adversarial attacks on time series prediction, Information Sciences 587 (2022) 794-812.
[8] Qin J., Wang J., Li Q., Fang S., Li X., Lei L., Differentially private frequent episode mining over event streams, Engineering Applications of Artificial Intelligence, 110 (2023) 104681
[9] Wu X., Qi L., Gao J., Ji G., Xu X., An ensemble of random decision trees with local differential privacy in edge computing, Neurocomputing, 485 (2022) 181-195
[10] Wang J., Liu C., Fu X., Luo X., Li X., A three-phase approach to differentially private crucial patterns mining over data streams, Computers & Security, 82 (2019) 30-48
[11] Chamikara M.A.P., Bertok P., Liu D., Camtepe S., Khalil I., Efficient privacy preservation of big data for accurate data mining, Information Sciences 527 (2020) 420-443
[12] Liu L., Kantarcioglu M., Thuraisingham B., The applicability of the perturbation based privacy preserving data mining for real-world data, Data & Knowledge Engineering, 65 (2008) 5-21

[13] Oladeji I., Maolo P., Zamora R., Lie T.T., Density-based clustering and probabilistic classification for integrated transmission-distribution network security state prediction, Electric Power Systems Research, 211 (2022) 106164
[14] Paul M.K., Islam M.R. Sattar A.H.M.S., An efficient perturbation approach for multivariate data in sensitive and reliable data mining, Journal of Information Security and Applications, 62 (2021) 102954
[15] Jangra S., Toshniwal D., Efficient algorithms for victim item selection in privacy-preserving utility mining, Future Generation Computer Systems, 128 (2022) 219-234.
[16] Yang S., Yin D., Song X., Dong X., Manogaran G., Mastorakis G., Mavromoustakis C.X., Batalla J.M.,Security situation assessment for massive MIMO systems for 5G communications, Future Generation Computer Systems,98 (2019) 25-34.
[17] Ahmed U., Srivastava G., Chun-Wei Lin J., A Machine Learning Model for Data Sanitization, Computer Networks, 189 (2021) 107914
[18] Ahmad D., Hameed S.A., Akhtar M., A multi-objective privacy preservation model for cloud security using hybrid Jaya-based shark smell optimization, Journal of King Saud University- Computer and Information Sciences, 34 (2022) 2343-2358
[19] Silva J.C., Giannella C., Bhargava R., Kargupta H., Klusch M., Distributed data mining and agents, Engineering Applciations of Artificial Intelligence 18 (2005) 791-807.
[20] Secretan J., Georgiopoulos M., Koufakou A., Cardona K., APHID: An architecture for private, high-performance integrated data mining, Future Generation Computer Systems (2010) 891-904.
[21] Fahad A., Tari Z., Almalawi A.,Goscinski A., Khalil I., Mahmood A., PPFSCADA: Privacy preserving framework for SCADA data publishing, Future Generation Computer Systems, 37 (2014) 496-511.
[22] Wang J., Fang S., Liu C., Qin J., Li X., Shi Z., Top-k closed co-occurrence patterns mining with differential privacy over multiple streams, Future Generation Computer Systems, 111 (2020) 339-351
[23] Turgay S., Erdoğan S., Security Impact of Federated and Transfer Learning on Network Management Systems with fuzzy DEMATEL Approach, Journal of Artificial Intelligence Practice, Paper ID: 101414
[24] Fazzinga B., Folino F., Furfaro F., Pontieri L., An ensemble-based approach to the security-oriented classification of low-level log traces, Expert Systems with Applications 153(2020). 113386.
[25] Han K., Xia B., Li Y., (AD) 2: Adversarial domain adaptation to defense with adversarial perturbation removal, Pattern Recognition, 122 (2022) 108303
[26] Chamikara M.A.P., Bertok P., Liu D., Camtepe S., Khalil I., An efficient and scalable privacy preserving algorithm for big data and data streams, Computers& Security, 87 (2019) 101570
[27] Ni C., Cang L.S., Gope P., Min G., Data anonymization evaluation for big data and IoT environment, Information Sciences 605 (2022) 381-392
[28] Li S., Mu N., Le J., Liao X., A novel algorithm for privacy preserving utility mining based on integer linear programming, Engineering Applications of Artificial Intelligence 81 (2019) 300-312
[29] Morhukuri V., Parizi R.M., Pouriyeh S., Huang Y., Dehghantanha A., Srivastava G., A survey on security and privacy of federated learning, Future Generations Computer Systems 115 (2021) 619-640
[30] Javid T., Gupta M.K., Gupta A., A hybrid-security model for privacy-enhanced distributed data mining, Journal of King Saud University- Computer and Information Sciences, 34 (2022) 3602-3614