Journal of Soft Computing Paradigm (JSCP) (2021)
Vol.03/ No.01
Pages: 19-28
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2021.1.003

Big Data Analysis and Perturbation using Data Mining Algorithm

Dr. Wang Haoxiang,
Director and lead executive faculty member,
GoPerception Laboratory,
NY, USA.
Email id: hw496@goperception.com

Dr. S. Smys,
Professor,
Department of CSE,
RVS Technical Campus,
Coimbatore, India.
Email id: smys375@gmail.com

Abstract: The advancement and introduction of computing technologies has proven to be highly effective and has resulted in the production of large amount of data that is to be analyzed. However, there is much concern on the privacy protection of the gathered data which suffers from the possibility of being exploited or exposed to the public. Hence, there are many methods of preserving this information they are not completely scalable or efficient and also have issues with privacy or data utility. Hence this proposed work provides a solution for such issues with an effective perturbation algorithm that uses big data by means of optimal geometric transformation. The proposed work has been examined and tested for accuracy, attack resistance, scalability and efficiency with the help of 5 classification algorithms and 9 datasets. Experimental analysis indicates that the proposed work is more successful in terms of attack resistance, scalability, execution speed and accuracy when compared with other algorithms that are used for privacy preservation.

Keywords: Big data, data perturbation, big data privacy, privacy-preservation data mining, information privacy

ISSN: 2582-2640 (online)
Submitted:15.02.2021
Revised: 09.03.2021
Accepted: 02.04.2021
Published: 19.04.2021

Journal of Soft Computing Paradigm (JSCP) (2021)
Vol.03/ No.01
Pages: 19-28
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2021.1.003

1. Introduction

As the advancement of computer technologies and IoT expands, there is a lot of data collected from the surrounding that require proper maintenance and protection. However, this data that is gathered will be useful only if it properly used in decision making. The connection and impact of different data is established with the help of data mining, paving way to proper establishment of relationship between the data to give information to the data users [1]. Moreover, this aspect of obtaining data will also be shared with other parties for proper analysis. During this process, there is lot of possibility for the information to be hacked and breached. The prospect of sharing information while simultaneous ensuring that the personal identifiable information is not disclosed is a mandatory part of information privacy. This also leads to more significant social, ethical, legal and technical challenges. Several commercial and governmental organizations gather large amount of user data apart from personal preferences, financial status, health and individual credit information. Healthcare systems, banking, social networking etc are some of the examples that use privacy information and often overlook the importance of privacy when used indirectly. Similarly, there are many information systems that make use of large quantities of private information for the purpose of analyzing and predicting phenomena based on human behavior such as social physics, epidemics and crimes. Thus, privacy is a crucial part of information gathering and serves to be a complex issue if it doesn't have a steady and reliable solution at hand [2].

Data mining is an optimal method of using the data without the need for disclosing information that is deemed to be private. Privacy-preserving data mining (PPDM) is a solution that is commonly used to address such problems with the help of encryption and data modification (perturbation). To secure data cryptographic methods are commonly used. A number of literature reviews have been studied with examples of cryptographic methods used by PPDM. A good example is its implementation in homomorphic encryption in areas such as sensor networks, cloud computing and e-health. Similarly set intersection, scalar product, secure set union and secure sum are some of the operations that are a part of distributed data mining. But this mechanism involved higher calculation complexity they cannot be used in PPDM [3]. In general data perturbation is set to operate when the computational complexity is low to preserve privacy. Record confidentiality is maintained by data perturbation with the help of systematic modification in the original database elements. The accuracy of PPDM is such that it is almost impossible to distinguish between the original data set and perturbed

ISSN: 2582-2640 (online)
Submitted:15.02.2021
Revised: 09.03.2021
Accepted: 02.04.2021
Published: 19.04.2021

Journal of Soft Computing Paradigm (JSCP) (2021)
Vol.03/ No.01
Pages: 19-28
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2021.1.003

data set so that it cannot be differentiated by third party users. However LDP and GDP on small data sets tend to fail because of their limited tuple count.

There are very few solutions that address the issue of partial/full data usage with LDP and will further involve a lot of issues such as low data utility and addition of noise. Thus both privacy and utility become contradictory factors that usually require compromise of either to perform effectively. The major contribution of this proposed work is the use of a big-data oriented privacy preservation algorithm that uses optimal geometric transformation [4]. This methodology incorporates a new privacy model that has an irreversible perturbation input, enabling release of full data. This methodology ensures that a sturdy system that works against attack during data reconstruction is possible, guaranteeing privacy. When compared with other methodologies, the proposed work is said to be quicker and can be used on multi-dimensional rotation of subplane, noise translation and axis reflection along with random tuple shuffling and randomized expansion. A method to increase the negativeness or positiveness of a specific data is randomized expansion. A memory overhead serves to be the closest solution as it gives excellent efficiency, classification accuracy and good attack resistance with respect to big data. We have used 5 classification algorithms and 9 datasets to test the proposed system and further compared it with geometric perturbation and random rotation perturbation.

2. Related Works

As internet-enabled consumer technologies become more common, it becomes more crucial to protect the privacy of this data. There are number of solutions researched towards finding the optimal solution to address this issue. Some approaches [5] are focused on using different methodologies with individual privacy while some others concentrated on increasing awareness. The biggest issue of them all is the issue of data privacy [6]. Though privacy and security concerns are a common glitch that is faced by all networks, depending on the environments and the dynamics of the devices used will play a big role in holding their impact. Thus such progress will lead to increased complexity that requires more complex privacy preservation and security efforts. In this paper, we have approached three technological approaches namely privacy-enhancing technologies [7], privacy-preserving data mining [8] and disclosure control. In a dynamic environment, location-based and temporal access control, controlling access via authentication and attribute-based encryption

ISSN: 2582-2640 (online)
Submitted:15.02.2021
Revised: 09.03.2021
Accepted: 02.04.2021
Published: 19.04.2021

Journal of Soft Computing Paradigm (JSCP) (2021)
Vol.03/ No.01
Pages: 19-28
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2021.1.003

are the commonly used mechanisms to increase the privacy of the system. Because of the efficiency and simplicity in operation, data perturbation [9] is a preferred form of privacy-preserving data mining technique. Here the input perturbation due to multiplication or addition of noise and output perturbation due to rule hiding and addition of noise are used. Input perturbation can be categorized into multidimensional perturbation and unidimensional perturbation. Some examples of the latter are micro-aggression [10], swapping, randomized response [11] and additive perturbation. Similarly, some examples of multidimensional perturbation include hybrid perturbation [12], random projection, geometric perturbation, random rotation and condensation [13]. The addition of random noise to the original data such that the attributes' statistical properties are kept intact is known as additive perturbation. However, the issue involved is the low utility of the resulting data.

3. Proposed Work

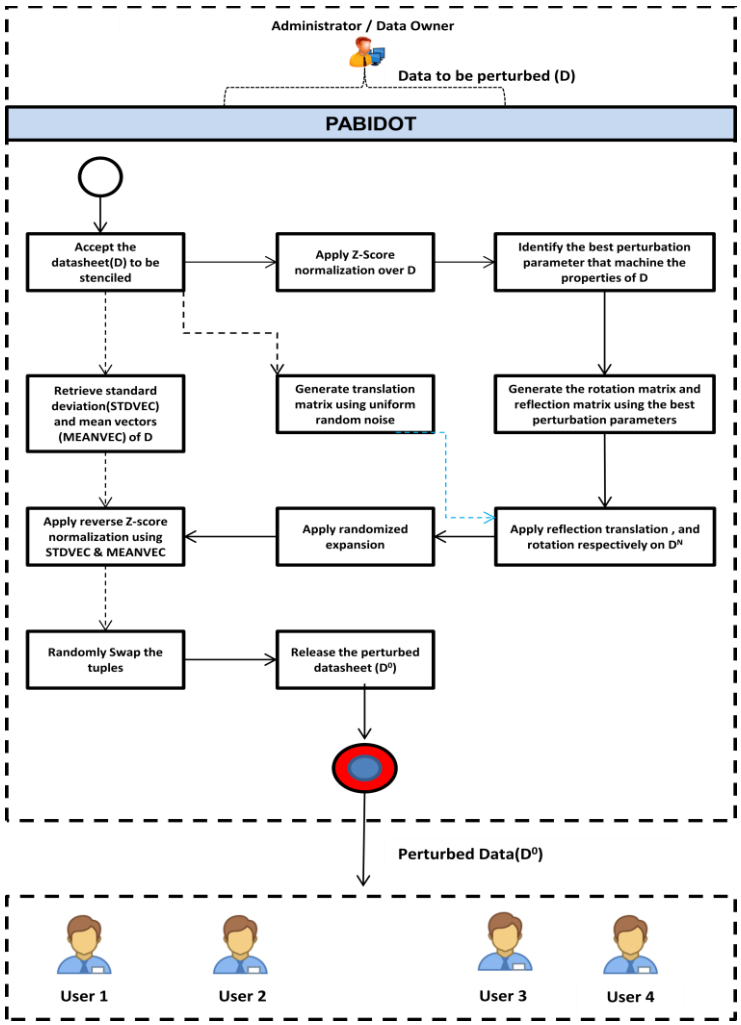

Fig.1 Architecture of the Proposed Methodology

ISSN: 2582-2640 (online)
Submitted:15.02.2021
Revised: 09.03.2021
Accepted: 02.04.2021
Published: 19.04.2021

Journal of Soft Computing Paradigm (JSCP) (2021)
Vol.03/ No.01
Pages: 19-28
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2021.1.003

In this paper, the proposed data mining techniques uses multidimensional rotation, translation, reflection and transformation along with random tuple shuffling and randomized expansion. The Fig.1 represents a simple flow diagram and architecture of the proposed work and accordingly, it enables enhanced privacy protection to the data against attacks. We have used the separation privacy model that executes the task at hand by choosing the apt perturbation parameters depending on the input dataset and their characteristics. The figure also shows the position of the proposed perturbation in big data release environment. We have assumed that only the owner of the dataset will be able to access the original data. Under no circumstance will this data be leaked to third party users. The proposed methodology uses geometric transformation taking into consideration the perturbation parameters and will also use randomized expansion to further elevate randomness. The last step involved is that of random tuple shuffle. Using the privacy model, we ensure that there is a positive difference between the obtained information and the original dataset that is being used. This will determine the apt perturbation possible and will also decrease the search space. On the other hand, reliability and efficiency of the big data perturbation system will continue to increase while simultaneously providing better reconstruction resistance. The standard deviation along with original dataset is the only inputs provided in the proposed work and the output obtained is the perturbed dataset.

3.1 Matrix Data D

A Matrix (D) is used to represent the perturbed data with a size of m x n such that ‘m’ denotes the records and ‘n’ denotes the attributes as columns. This matrix is said to hold only numerical data values such that:

$$D = \begin{bmatrix} a_{11} & a_{12} & a_{1n} \\ a_{21} & a_{2k} & a_{2n} \\ \vdots & & \\ a_{m1} & a_{2k} & a_{2n} \end{bmatrix}$$

These values can represent various values such as gender, height, weight, age etc. This matrix is exposed to multidimensional geometric composite transformation during perturbation process. Every record is taken to be a single Cartesian coordinate point. Similarly, in n-dimensional space, the geometric reflection, rotation and translation is said to be isometric. In such a case, the distance is preserved such that

$$|T(U) - T(V)| = |U - V|, \qquad such\ that\ A, B \in R^n$$

ISSN: 2582-2640 (online)
Submitted:15.02.2021
Revised: 09.03.2021
Accepted: 02.04.2021
Published: 19.04.2021

Journal of Soft Computing Paradigm (JSCP) (2021)
Vol.03/ No.01
Pages: 19-28
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2021.1.003

When many transformation matrices are used, a composite operation is performed such that the matrices $M_1, M_2, M_3, \ldots$ are implemented on the homogeneous matrix as shown below:

$$X' = \left(M_3\left(M_2(M_1X)\right)\right) = .. M_1x\ M_2x\ M_3x\ X$$

On adding new columns, the input data is converted such that it indicates homogenous ccoordinates. This can be represented by:

$$D = \begin{bmatrix} a_{11} & a_{12} & a_{1n} & 1 \\ a_{21} & a_{2k} & a_{2n} & 1 \\ \vdots & & & \\ a_{m1} & a_{2k} & a_{2n} & 1 \end{bmatrix}$$

Similarly, the n-dimensional homogenous translation matrix can be expressed using their Cartesian points such that:

$$T_{ND} = \begin{bmatrix} 1 & 0 \ldots . & 0 & t1\ (n+1) \\ 0 & 1 \ldots . & 0 & t2\ (n+1) \\ \vdots & & & \\ a_{m1} & 0 & 0 & 1) \end{bmatrix}$$

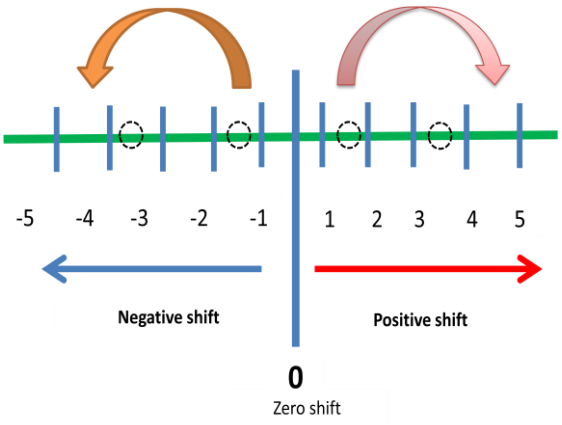

Fig.2. Expansion of Randomized

4. Results and Discussion

For a new value of ϕ, there is a drastic reduction in the time consumed which is evident in Fig.3 and Fig.4 showing the optimized version of perturbation and simple perturbation with data mining. It is evident that the proposed methodology PABIDOT with data mining produces a considerably efficient algorithm compared with the basic version.

ISSN: 2582-2640 (online)
Submitted:15.02.2021
Revised: 09.03.2021
Accepted: 02.04.2021
Published: 19.04.2021

Journal of Soft Computing Paradigm (JSCP) (2021)
Vol.03/ No.01
Pages: 19-28
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2021.1.003

But, time consumption varies and will thus have an improvement of efficiency in the proposed algorithm.

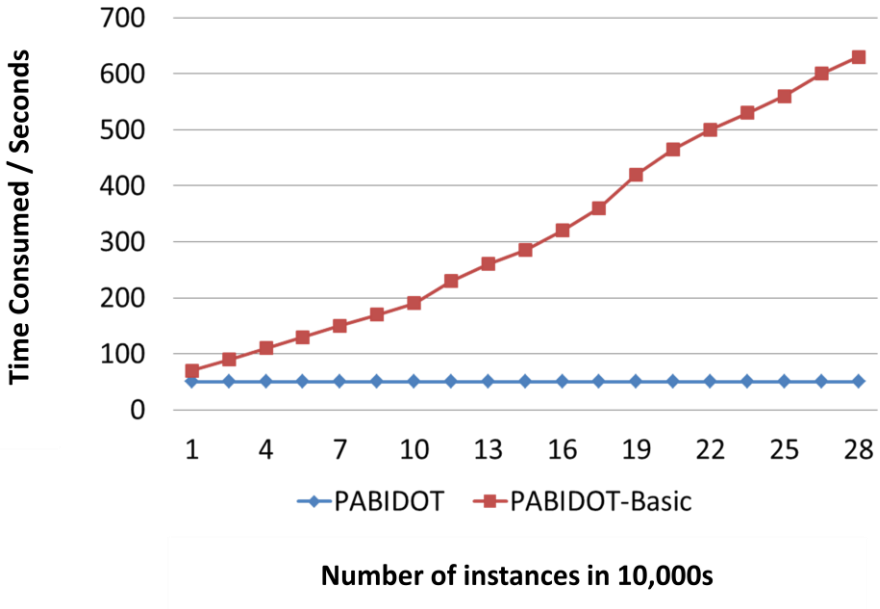

Fig. 3 Time Consumption for 10,000 with respect to the tuples

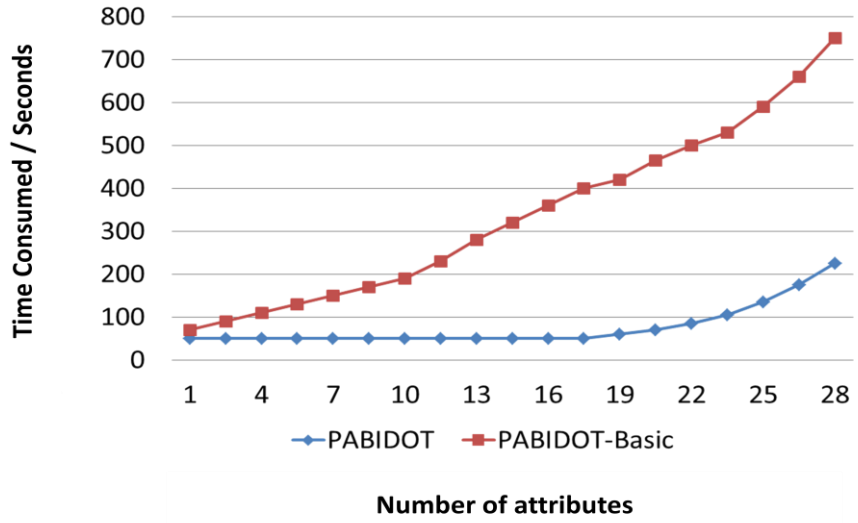

Fig.4 Time Consumption for 10,000 with respect to the number of attributes

Fig.5 and Fig.6 show that the time consumed for GP and RP are is higher when compared with the proposed methodology PABIDOT. Fig.6 indicates a perturbed dataset representation for model validation where testing as well as model training is performed with the same dataset.

ISSN: 2582-2640 (online)
Submitted:15.02.2021
Revised: 09.03.2021
Accepted: 02.04.2021
Published: 19.04.2021

Journal of Soft Computing Paradigm (JSCP) (2021)
Vol.03/ No.01
Pages: 19-28
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2021.1.003

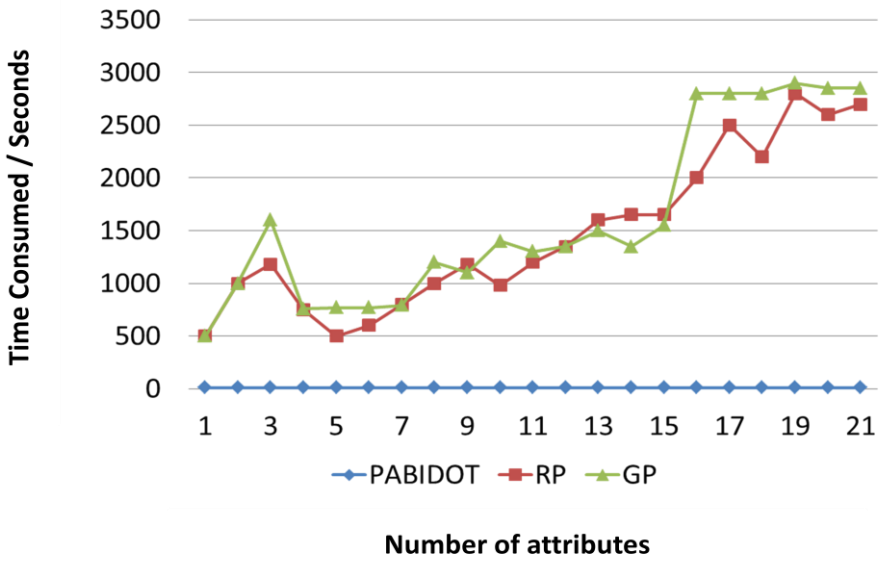

Fig. 5. Time Consumption with respect to the tuples for three methodologies

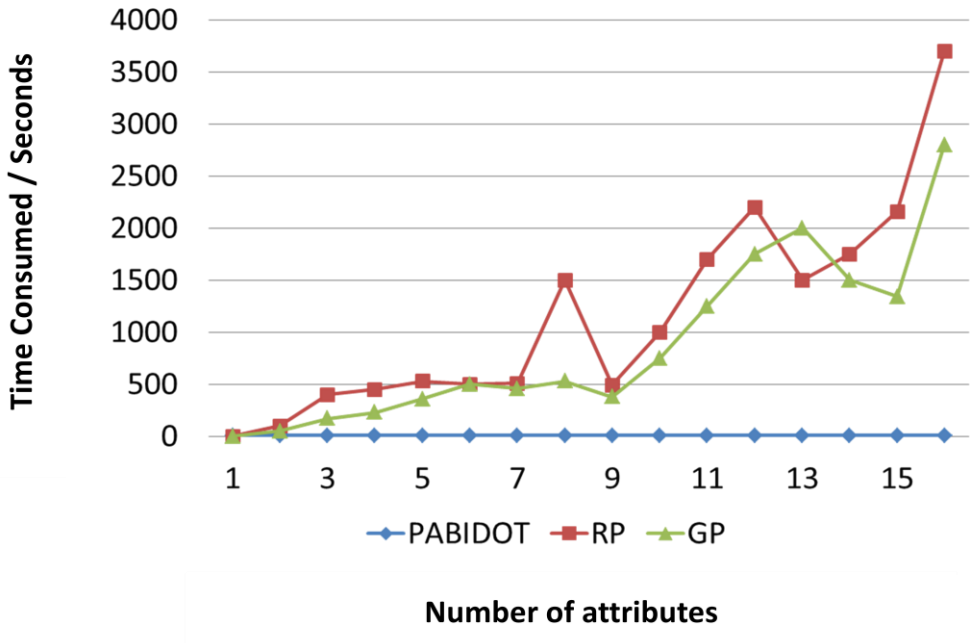

Fig.6 Time Consumption with respect to the number of attributes for three methodologies

5. Conclusion

In this proposed work, perturbation with data mining is proposed to address utility issues, usability, privacy, scalability and efficiency. We have incorperated an empirical separation that is capable of selecting the apt perturbation parameters to identify data. It was also found that the proposed methodology is linear in property with O(m) such that m should

ISSN: 2582-2640 (online)
Submitted:15.02.2021
Revised: 09.03.2021
Accepted: 02.04.2021
Published: 19.04.2021

Journal of Soft Computing Paradigm (JSCP) (2021)
Vol.03/ No.01
Pages: 19-28
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2021.1.003

be particularly higher than that of the attributes, representing the number of instances. The accuracy of this work was determined to be very accurate with respect to the original data and the empirical results indicate high resistance with respect to privacy attacks.

References

[1] Shynu, P. G., Shayan, H. M., & Chowdhary, C. L. (2020, February). A fuzzy based data perturbation technique for privacy preserved data mining. In *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)* (pp. 1-4). IEEE.

[2] Kargupta, H., Datta, S., Wang, Q., & Sivakumar, K. (2005). Random-data perturbation techniques and privacy-preserving data mining. *Knowledge and Information Systems*, *7*(4), 387-414.

[3] Shirley, D. R. A., Ranjani, K., Arunachalam, G., & Janeera, D. A. (2021). Automatic Distributed Gardening System Using Object Recognition and Visual Servoing. In *Inventive Communication and Computational Technologies* (pp. 359-369). Springer, Singapore.

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

[5] Anand, J. V. (2020). A Methodology of Atmospheric Deterioration Forecasting and Evaluation through Data Mining and Business Intelligence. *Journal of Ubiquitous Computing and Communication Technologies (UCCT)*, *2*(02), 79-87.

[6] Chen, K., & Liu, L. (2011). Geometric data perturbation for privacy preserving outsourced data mining. *Knowledge and information systems*, *29*(3), 657-695.

[7] Li, J. Y., Zhan, Z. H., Wang, H., & Zhang, J. (2020). Data-driven evolutionary algorithm with perturbation-based ensemble surrogates. *IEEE Transactions on Cybernetics*.

[8] Kanth, P. C., & Anbarasi, M. S. (2020). A generic framework for data analysis in privacy-preserving data mining. In *Computational Intelligence in Data Mining* (pp. 653-661). Springer, Singapore.

[9] Kataka, E., Zaucha, J., Frishman, G., Ruepp, A., & Frishman, D. (2020). Edgetic perturbation signatures represent known and novel cancer biomarkers. *Scientific reports*, *10*(1), 1-16.

ISSN: 2582-2640 (online)
Submitted:15.02.2021
Revised: 09.03.2021
Accepted: 02.04.2021
Published: 19.04.2021

Journal of Soft Computing Paradigm (JSCP) (2021)
Vol.03/ No.01
Pages: 19-28
http://irojournals.com/jscp/
DOI: https://doi.org/10.36548/jscp.2021.1.003

[10] García, J., Lalla-Ruiz, E., Voss, S., & Droguett, E. L. (2020). Enhancing a machine learning binarization framework by perturbation operators: analysis on the multidimensional knapsack problem. *International Journal of Machine Learning and Cybernetics*, 1-20.

[11] Feyisetan, O., Balle, B., Drake, T., & Diethe, T. (2020, January). Privacy-and utility-preserving textual analysis via calibrated multivariate perturbations. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (pp. 178-186).

[12] Suma, V., & Hills, S. M. (2020). Data Mining based Prediction of Demand in Indian Market for Refurbished Electronics. *Journal of Soft Computing Paradigm (JSCP)*, *2*(03), 153-159.

[13] Shakya, S. (2020). Process mining error detection for securing the IoT system. *Journal of ISMAC*, *2*(03), 147-153.

ISSN: 2582-2640 (online)
Submitted:15.02.2021
Revised: 09.03.2021
Accepted: 02.04.2021
Published: 19.04.2021