



Comparative Analysis on New Variant DE and CSA: Privacy Preservation in Data Mining

¹G.K.Shailaja and ²Dr.C.V.Guru Rao

1Associate Professor

Dept. of IT, Kakatiya Institute of Technology and Science-Warangal,
Telangana, India-506015

2Professor and Director of Evaluation,

Dept. of CSE, S.R.Engineering College, Warangal, Telangana, India-506371

Abstract: In this digital era, the world cannot be imagined without communication. For several purposes, the human beings are required to exchange information. Here, both security and privacy protection was considered as a public policy concern for decades. However, speedy scientific variations are made in case of the fast development of the internet and electronic exchange, which may lead to the leakage of personal information. Hence, this paper intends to formulate an innovative Privacy-Preserving Data Mining (PPDM) technique by following two major stages viz. data sanitization and data restoration. In both the sanitization as well as restoration process, key extraction process plays an important role, where the selection of optimal key must be important. This paper aims to introduce a new Opposition Intensity-based Cuckoo Search Algorithm (OI-CSA) for optimal key selection. Further, with the proposed model, four sanitization research challenges such as hiding failure, information loss, and false rule generation and modification degree are minimized. Finally, the performance of proposed work is analyzed by varying the step size of proposed OI-CSA.

Keywords: PPDM; Sanitization; Restoration; Key Extraction; Cuckoo Search; Step size.

I. Introduction

In the business world, data preservation [1] [2] has turn out to be an increasingly significant task. Certain steps have to also be taken by individuals so as to protect the data from leakage. Data mining is generally described as the practice of discovering patterns from huge quantities of data. The description is considered to be unclear since it has to include the enormous collection of approaches, procedures, and algorithms from several domains such as records, machine learning, and information. Data mining [3] [4] [5] is frequently measured as a significant process in case of knowledge discovery process. In this digital age, it is difficult to imagine the world without communication.

For several purposes, the human beings are required to exchange information's [6] [7]. Here, both security and privacy protection was considered a public policy concern for decades. However, speedy scientific variations are made in case of the fast development of the internet and electronic exchange, and the expansion of most complicated approaches for gathering, investigating, and usage of personal information [8] [9] [10]. This leads to the government concerns which made privacy as a major issue. The data mining [11] [12] field is attaining significant identification with respect to the assessment of huge volumes of information which is gathered and then stored through a computer system. In recent times, the outsized volume of information which is collected from several channels is comprised of much personal information. During the investigation of both personal and sensitive information, there arises a doubt in case of analyzing the confidentiality of individuals [13] [14] [15]. Thus, data mining software is regarded to be as one of the major tools for investigating the privacy of the data [16] [17].

For this motive, several research works are employed on privacy-preserving data mining [18] [19], in which it is considered as an enhanced approach for extracting the knowledge with respect to the protection of privacy of users. Some of these approaches normally aim at the privacy of the individuals while at the same time; several others aim at the privacy of the corporation. Securing the communication is regarded to be as an immense confronts depending on the increasing threats and attacks against network security [20] [21]. Thus, the securing the network is considered to be as the most important challenge in this period depending on the various threats and attacks.

The major contribution of this paper is depicted below.

1. This paper intends to present a novel data privacy preservation model that includes two major phases' viz. data sanitization and data restoration.
2. In both the sanitization and restoration process, key extraction plays an important role, for which OI-CSA model is exploited for optimal key selection.
3. By carrying out optimal key selection, four sanitization research challenges could be reduced.
4. Finally, algorithmic analysis is carried out by varying the step size of adopted OI-CSA and the results are determined.

The arrangement of the paper is as follows. Section II analyzes the literature work. Section III describes the proposed data sanitization and restoration model and section IV portrays the objective function and optimal key generation using improved optimization model. In addition, section V illustrates the results and section VI concludes the paper.

II. Literature review

A. Related works

In 2018, Afzali and Mohammadi [1] have established a novel data anonymization model, which was integrated in the adopted scheme for big data mining. In addition, certain characteristics of big data namely, velocity formulates it crucial to consider all the rules as a confidential one. Further, parallelization schemes included in the developed scheme aids to raise the speed of Association Rule Mining (ARM). Moreover, the adopted system was examined, and noteworthy conclusions were substantiated that shows the enhanced performance of the presented technique over the conventional techniques.

In 2016, Li *et al.* [2] have established two schemes that were depend on distributed ensemble scheme. The chief idea of established model was to learn the distribution of data precisely and moreover to convey the healthcare data devoid of sharing the confidential details of user, thereby protecting the privacy of patients. Here, the presented scheme could efficiently form robust and accurate prediction model. In addition, by means of the developed model, the biomarkers were discovered and also the preferred biomarkers were authorized by biomedical researchers.

In 2016, Tripathi *et al.* [3] have established a significant model for discovering the sensitive hidden contents in huge archives of data mining. There were two kinds of discrimination, initial one was "direct discrimination" and subsequent one was "indirect discrimination". The initial one remains when choices were considered depending on the sensitive features. The later one remains when choices were considered depending on non-sensitive features, which were correlated robustly with the sensitive attributes.

In 2017, Li *et al.* [4] have proposed PMDP model for securing the privacy of the users' data in cloud computing. The introduced model was a completely trusted authority as it had the capacity of decrypting any cipher text that was created by the user. PMDP was constructed depending on the multi key FHE. Further, PMDP framework had split the contestants into 2 classes, honest and semi-honest. The performance of the adopted scheme had exhibited enhanced security and it had also improved the security of the multiparty data aggregation.

In 2019, Abdel *et al.* [5] have formulated Restricted Sensitive Attributes-based Sequential Anonymization (RSA-SA) for preserving the privacy of the data stream. Here, two more privacy constraints such as, Semantic-diversity and Sensitivity-diversity were regarded for conserving the sensitive values in the data streams. The proposed model had generated a better protection against attacks such as similarity attacks, and sensitivity attacks. Further, the adopted scheme was found to be better in offering minimum delay time and information loss in sensitive attributes. The outcomes of the presented model were found to be better in terms of accurate mining outcomes over the existing schemes.

III. Proposed data sanitization and restoration model

Fig. 1 shows the framework of the adopted OI-CSA model. Initially, the data preservation process includes two major processes, namely data sanitization and data restoration. The data sanitization procedure includes a key generation process for preserving the confidential data in a secure manner. The key has to be created in such a way that it must hide the confidential data efficiently. Thus, for optimal key generation, OI-CSA is exploited in this context. The authorized person at the receiver side will then restore the sanitized data by using the same key.

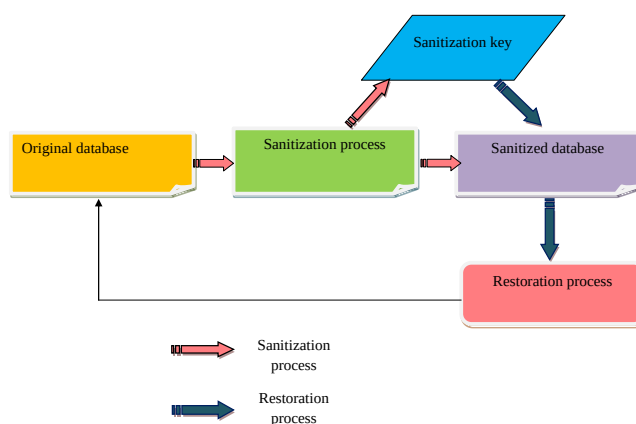


Fig. 1. Structure of the adopted model

A. Sanitization Process

Here the original database is taken as O and the sanitized database is denoted by O' . During sanitization, binarization of O and pruned key matrix, A_2 are done. The resultant key matrix in binarized form is subjected to the rule hiding process repeatedly, where the XOR function is carried out with binarized form of O with the same matrix dimensions and added up with one that generates the O' , as revealed by Eq. (1). In addition, O' attained from sanitization process attains SRs and association rules following the sanitization of B' . Likewise, O extracts the relative association rules before sanitization B thereby attains the aforesaid objectives. The modelling of the proposed sanitization scheme is given by Fig. 2.

$$O' = (A_2 \oplus O) + 1 \quad (1)$$

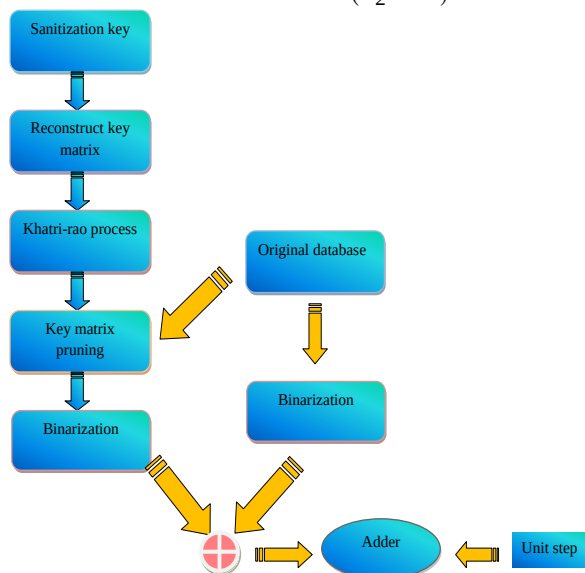


Fig. 2. Sanitization process of the presented model

B. Key Generation

The key generation comprises of solution transformation process, in which key representation, denoted by A is converted using khatri-rao product. At first, A is restructured into A_1 with matrix dimensions $\left[\sqrt{M_O''} \times O_{\max} \right]$ in which $O_{\max}$ point out the maximum transaction length and M_O indicates the transaction counts, the adjacent peak perfect square of M_O is given by M_O'' . For example, the restructured process of $A = \{1,2,1\}$ performs duplication in row-wise and it forms the reconstructed key matrix, A_1 with dimension $\left[\sqrt{M_O''} \times O_{\max} \right]$ as shown in Eq. (2).

$$A_1 = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \\ 1 & 1 & 1 \end{bmatrix}_{\left[\sqrt{M_O} \times O_{\max} \right]} \quad (2)$$

Therefore, the key matrix A_2 with dimensions $\left[\sqrt{M_O} \times O_{\max} \right]$ is obtained by Khatri-rao product of two identically restructured A_1 matrixes (i.e. $A_1 \otimes A_1$) in which kronecker product is denoted by $\otimes$. Depending on Khatri-rao product, the key generation procedure is performed and it produces matrix with dimensions similar to O , which generate A_2 $\left[\sqrt{M_O} \times O_{\max} \right]$. In the end, the process of rule hiding is performed for obtaining O' by hiding the sensitive rules. Here, the optimal key generation is made using OI-CSA scheme.

C. Restoration process

Throughout the restoration process, the O' obtained from sanitization and A_2 from key generation approaches are binarized. The binarized S_d obtained from binarization block is then minimized from unit step unit. In the mean time, the database and key matrix that are binarized performs the XOR function following the subtraction, and consequently the extraction of restored database takes place. The sanitizing key, A_2 is reconstructed by deploying Eq. (2), Eq. (1), Eq. (4) and adopted OI-CSA update. It is employed to generate O' by which the lossless restoring is done by Eq. (3), in which $\hat{O}$ denotes the restored data. The modelling of restoration process is specified by Fig. 3.

$$\hat{O} = (O' - 1) \oplus A_2 \quad (3)$$

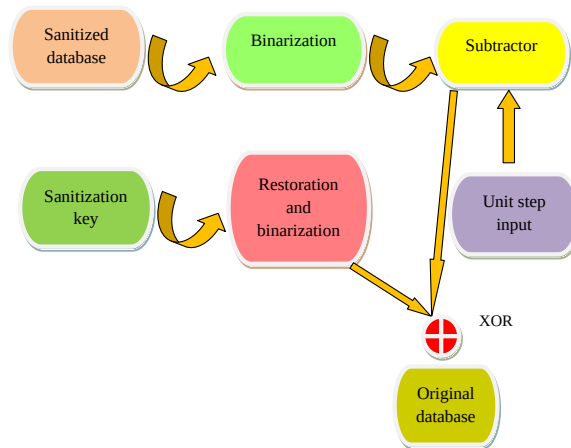


Fig. 3. Restoration process of the of the presented model

IV. Objective function and Optimal key generation using improved optimization model

A. Objective Function

The introduced OI-CSA approach aims to accomplish the objective function for preserving the data as specified by Eq. (1), in which F_1, F_2, F_3 and F_4 denotes the objectives. These objectives portrays the significance of the relevant function as defined by Eq. (5)-Eq. (8). In Eq. (5), F_1 specifies the normalized rate of HF, f_1 indicates the HF rate, in which, $\max(f_1)$ is regarded as the worst f_1 of the entire iterations. In Eq. (6), F_2 signifies the normalized MD rate, f_2 specifies the MD. In Eq. (7), F_3 indicates the normalized IP rate f_3 points out the IP rate and in Eq. (8), F_4 symbolizes the normalized FR rate, f_4 shows the FR.

$$\min F = \max(F_1, F_2, F_3, F_4) \quad (4)$$

$$F_1 = \frac{f_1}{\max(f_1) \forall \text{iterations}} \quad (5)$$

$$F_2 = \frac{f_2}{\max(f_2) \forall \text{iterations}} \quad (6)$$

$$F_3 = \frac{f_3}{\max(f_3) \forall \text{iterations}} \quad (7)$$

$$F_4 = \frac{f_4}{\max(f_4) \forall \text{iterations}} \quad (8)$$

HF rate indicated by f_1 is portrayed as the fraction of sensitive rules which is depicted in O' as specified by Eq. (9). Accordingly, the count of sensitive rules present in O' is portrayed as $f_1 = \frac{|B' \cap SRs|}{|SRs|}$. In Eq. (9), B indicates the association rule produced prior to sanitization, B' signifies the association rules attained from O' and SRs indicates the sensitive rules.

$$f_1 = \frac{|B' \cap SRs|}{|SRs|} \quad (9)$$

IP rate indicated by f_2 is explained as “the rate of non-sensitive rules which are concealed in O' ”. It is given in Eq. (10).

$$f_2 = 1 - \frac{|B - B'|}{|B|} \quad (10)$$

FR denoted by f_3 is portrayed as “the rate of artificial rules produced in O' ” as shown by Eq. (11).

$$f_3 = \frac{|B - B'|}{|B'|} \quad (11)$$

DM specified by f_4 is described as the count of modifications done in O' from O as given by Eq. (12), in which $dist$ signifies the euclidean distance between O and O' .

$$f_4 = dist(O, O') \quad (12)$$

B. Key Encoding

The keys (chromosome) A deployed for sanitization process are provided to OI-CSA model for encoding purpose. The count of keys from key A^1 to key A^M is optimized by means of OI-CSA model, and the optimal key is discovered. The solution encoding process is given by Fig. 4. Here, the length of key is specified by $\sqrt{M_O}$.



Fig. 4. Keys for encoding

C. Cuckoo Search Algorithm

CS approach [22] is dependent on the below three conditions:

1. Every cuckoo chooses a nest in arbitrary style and lays an egg in it.
2. The nests with high quality eggs goes for further generations.
3. For a certain amount of nests, a host cuckoo can realize another egg with a probability $P \in [0,1]$.

The balance among global and local random walks are demonstrated by Eq. (13) and Eq. (14). Here, in Eq. (13), X_i^t and X_k^t specifies the current positions preferred by random permutation, β symbolizes the “positive step size scaling factor”, X_i^{t+1} specifies the net position, s indicates the step size, $\otimes$ symbolizes the element-wise product, F specifies the “heavy side function”, P is a term that switches between random and global walks, ε denotes the uniform distribution. Consequently, from Eq. (14), $N(s, \tau)$ specifies levy distribution.

$$X_i^{t+1} = X_i^t + \beta s \otimes F(P - \varepsilon) \otimes (X_i^t - X_k^t) \quad (13)$$

$$X_i^{t+1} = X_i^t + \beta N(s, \tau) \quad (14)$$

D. OI-CS Algorithm

The traditional CS model is a simple and proficient global optimization model; however it could not be directly exploited to resolve multimodal optimization issues. Therefore, the conventional CS approach is enhanced by modifying the opposition intensity indicated by γ as specified in Eq. (15). In Eq. (14), $X_i^{(w)}$ symbolizes the worst solution, X_i^t specifies the current solution and γ differs from 0 to 1.

$$X_i^{t+1} = X_i^t + \beta N(s, \tau) - \gamma [X_i^{(w)} - X_i^t] \quad (15)$$

V. Results and discussion

A. Simulation procedure

The implemented OI-CSA technique for preserving the sensitive data was simulated in JAVA, and the outcomes were attained. The experimentation was performed by means of four datasets namely, T10, Chess,

Retail, and T40. Furthermore, algorithmic analysis was carried out by varying the step size value (s value) of the proposed algorithm that influences HF, IP, and FR and DM as well for the four adopted datasets.

B. Effect on varying s

The effect of varying s using four datasets namely, chess, retail, T40, and T10 are illustrated by Table I- Table IV correspondingly. The values of s was varied from 0.2, 0.4, 0.6, 0.8 and 1. Accordingly, from Table I, the normalized HF rate, F_1 has attained the values of 0.4025, 0.3705, 0.354, 0.3505 and 0.343 for $s=0.2$, $s=0.4$, $s=0.6$, $s=0.8$ and $s=1$ correspondingly using chess dataset. From Table II, the normalized MD rate, F_2 has attained the values of 0.499756, 0.487805, 0.487758, 0.48771 and 0.48771 for $s=0.2$, $s=0.4$, $s=0.6$, $s=0.8$ and $s=1$ correspondingly using retail dataset. In addition, from Table III, the normalized IP rate, F_3 has attained the values of 0.192686, 0.136847, 0.125905, 0.122117 and 0.112924 for $s=0.2$, $s=0.4$, $s=0.6$, $s=0.8$ and $s=1$ correspondingly using T40 dataset. Also, from Table IV, using T10 dataset, the normalized FR rate, F_4 has acquired the values of 0.231504, 0.207106, 0.156523, 0.13255 and 0.129938 for $s=0.2$, $s=0.4$, $s=0.6$, $s=0.8$ and $s=1$ correspondingly. Therefore, the superiority of the presented OI-CSA approach has been validated effectively.

TABLE I. Effect of Varying s Using Chess Dataset

Measures	$s=0.2$	$s=0.4$	$s=0.6$	$s=0.8$	$s=1$
F_1	0.4025	0.3705	0.354	0.3505	0.343
F_2	0.482697	0.482394	0.481891	0.481072	0.479594
F_3	0.18372	0.146247	0.132245	0.120867	0.079577
F_4	0.175877	0.169263	0.155692	0.122531	0.097969
F	0.482697	0.482394	0.481891	0.481072	0.479594

TABLE II. Effect of Varying s Using Retail Dataset

Measures	$s=0.2$	$s=0.4$	$s=0.6$	$s=0.8$	$s=1$
F_1	0.255	0.25225	0.25225	0.25125	0.25025
F_2	0.499756	0.487805	0.487758	0.48771	0.48771
F_3	0.251286	0.152929	0.145618	0.100095	0.095046
F_4	0.248804	0.159269	0.149178	0.102858	0.096025
F	0.499756	0.487805	0.487758	0.48771	0.48771

TABLE III. Effect of Varying s Using T40 Dataset

Measures	$s=0.2$	$s=0.4$	$s=0.6$	$s=0.8$	$s=1$
F_1	0.253	0.2505	0.25025	0.24925	0.031206
F_2	0.488203	0.488131	0.48813	0.488012	0.487441
F_3	0.192686	0.136847	0.125905	0.122117	0.112924
F_4	0.21206	0.149956	0.14556	0.134422	0.116622
F	0.488203	0.488131	0.48813	0.488012	0.487441

TABLE IV. Effect of Varying s Using T10 Dataset

Measures	$s=0.2$	$s=0.4$	$s=0.6$	$s=0.8$	$s=1$
F_1	0.252	0.25025	0.25025	0.24975	0.24925
F_2	0.477552	0.477508	0.477465	0.477465	0.477289
F_3	0.217936	0.19655	0.151734	0.122342	0.117278
F_4	0.231504	0.207106	0.156523	0.13255	0.129938
F	0.477552	0.477508	0.477465	0.477465	0.477289

VI. Conclusion

This paper has presented a novel PPDM technique using two phases namely, data sanitization and data restoration that commences after the generation of association rules. In both the sanitization and restoration process, key extraction plays a major role that was optimally chosen using OI-CSA model. Accordingly, four objectives such as, HF rate, IP, and FR and DM were minimized by means of the adopted sanitization and restoration processes. In addition, algorithmic analysis was carried out by altering the step size value (s) of the proposed algorithm for the four adopted datasets namely, T10, Chess, Retail, and T40. From the analysis, the

normalized HF rate, F_1 has attained the values of 0.4025, 0.3705, 0.354, 0.3505 and 0.343 for $s=0.2$, $s=0.4$, $s=0.6$, $s=0.8$ and $s=1$ respectively using chess dataset. Thus, the superiority of the presented OI-CSA approach has been validated in an effective manner.

References

- [1] G. A. Afzali and S. Mohammadi, "Privacy preserving big data mining: association rule hiding using fuzzy logic approach," *IET Information Security*, vol. 12, no. 1, pp. 15-24, 1 2018.
- [2] Y. Li, C. Bai and C. K. Reddy, "A distributed ensemble approach for mining healthcare data under privacy constraints", *Information Sciences*, vol. 330, pp. 245–259, 2016.
- [3] K. K. Tripathi, "Discrimination Prevention with Classification and Privacy Preservation in Data mining", *Computer Science*, vol.79, pp. 244-253, 2016.
- [4] J. Li, J. Wei, W. Liu, and X. Hu, "PMDP: A Framework for Preserving Multiparty Data Privacy in Cloud Computing", *Security and Communication Networks*, 2017.
- [5] S. A. Abdelhameed, S. M. Moussa, M. E. Khalifa, "Restricted Sensitive Attributes-based Sequential Anonymization (RSA-SA) approach for privacy-preserving data stream publishing Knowledge-Based Systems, vol.164, pp. 1-20, January 2019.
- [6] S. Upadhyay, C. Sharma, P. Sharma, P. Bharadwaj, K.R. Seeja, "Privacy preserving data mining with 3-D rotation transformation", *Journal of King Saud University - Computer and Information Sciences*, 28 November 2016.
- [7] M. Prakash, G. Singaravel, "An approach for prevention of privacy breach and information leakage in sensitive data mining", *Computers & Electrical Engineering*, vol. 45, pp. 134-140, July 2015.
- [8] T. P. Bhat, C. Karthik, K. Chandrasekaran, "A Privacy Preserved Data Mining Approach Based on k-Partite Graph Theory", *Procedia Computer Science*, vol. 54, pp. 422-430, 2015.
- [9] M. A. Ferrag, L. A. Maglaras, H. Janicke, J. Jiang, L. Shu, "A systematic review of data protection and privacy preservation schemes for smart grid communications", *Sustainable Cities and Society*, vol. 38, pp. 806-835, April 2018.
- [10] A. Romanou, "The necessity of the implementation of Privacy by Design in sectors where data protection concerns arise", *Computer Law & Security Review*, vol. 34, no. 1, pp. 99-110, February 2018.
- [11] D. Sánchez, M. Batet, "Privacy-preserving data outsourcing in the cloud via semantic data splitting", *Computer Communications*, vol. 110, pp. 187-201, 15 September 2017.
- [12] L. Li, R. Lu, K. K. R. Choo, A. Datta and J. Shao, "Privacy-Preserving-Outsourced Association Rule Mining on Vertically Partitioned Databases," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 8, pp. 1847-1861, Aug. 2016.
- [13] J. C. Lin, T. Wu, P. Fournier-Viger, G. Lin, M. Voznak, "Fast algorithms for hiding sensitive high-utility itemsets in privacy-preserving utility mining", *Engineering Applications of Artificial Intelligence*, vol. 55, pp. 269-284, October 2016.
- [14] M. A. P. Chamikara, P. Bertok, D. Liu, S. Camtepe, I. Khalil, "Efficient data perturbation for privacy preserving and accurate data stream mining", *Pervasive and Mobile Computing*, 12 May 2018.
- [15] S. Vennila, J. Priyadarshini, "Scalable Privacy Preservation in Big Data a Survey", *Procedia Computer Science*, vol. 50, pp. 369-373, 2015.
- [16] H. K. Bhuyan, N. K. Kamila, "Privacy preserving sub-feature selection in distributed data mining", *Applied Soft Computing*, vol. 36, pp. 552-569, November 2015.
- [17] L. Liu, M. Kantarcioglu, B. Thuraisingham, "The applicability of the perturbation based privacy preserving data mining for real-world data", *Data & Knowledge Engineering*, vol. 65, no. 1, pp. 5-21, April 2008.
- [18] T. P. Bhat, C. Karthik, K. Chandrasekaran, "A Privacy Preserved Data Mining Approach Based on k-Partite Graph Theory", *Procedia Computer Science*, vol. 54, pp. 422-430, 2015.
- [19] F. González-Serrano, Á. Navia-Vázquez, A. Amor-Martín, "Training Support Vector Machines with privacy-protected data", *Pattern Recognition*, vol. 72, pp. 93-107, December 2017.
- [20] Y. Kokkinos, K. G. Margaritis, "Confidence ratio affinity propagation in ensemble selection of neural network classifiers for distributed privacy-preserving data mining", *Neurocomputing*, vol. 150, Part B, pp. 513-528, 20 February 2015.
- [21] M. Mareli, B. Twala, "An adaptive Cuckoo search algorithm for optimisation", *Applied Computing and Informatics*, 6 September 2017.