IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-6, Issue-5, May 2018 E-ISSN: 2347-2693

Data Transformation Technique for Preserving Privacy in Data

Uma Shankar Rao Erothi[1*], Sireesha Rodda[2]

[1]Department of CSE, RAGHU Institute of Technology, Visakhapatnam, India
[2]Department of CSE, GITAM Institute of Technology, GITAM Deemed to be University, Visakhapatnam, India

[*]*Corresponding Author: umashankar.erothi3@gmail.com, Tel.: +91-8885411100*

Available online at: www.ijcseonline.org

18/May/2018, Published: 31/May/2018

***Abstract*—** The increase of digitization has led to growing concerns over preserving privacy of sensitive data. The ubiquity of sensitive information in data sources such as financial transactions, commercial transactions, medical records, network communication etc., steered towards development of different privacy preserving techniques. In this paper, a novel data transformation technique has been proposed for providing efficient privacy preservation in the data. Inorder to provide privacy to data, the numeric attributes are transformed to the range [-1,1] while the characters or strings are transformed to binary strings. Data analysis over the transformed dataset provides the same result as that of the original dataset. The performance of the data transformation technique is evaluated on the datasets before and after transformation. Experiments on five standard datasets indicate high data utility of the proposed technique. The proposed technique is also evaluated on the standard network intrusion dataset NSL-KDD dataset to study the effectiveness of the proposed technique in intrusion detection domain and the results are analyzed. Privacy measures are evaluated to ascertain the degree of privacy offered by the proposed technique.

Keywords— *Privacy Preservation, PPDM, Data Transformation, Network Intrusion Detection, Data Mining*.

I. Introduction

The advent of digital age in to our lives has brought forward many changes and made itself indispensable. Many personal and sensitive data now relies continually on computers and internet, including applications related to national security communication, finance, transportation, medicine and even education. With the amount of confidential and sensitive information stored under computer networks, it must be ensured that the data resides (or transmitted across networks) securely. Cyber security involves protecting computers, networks and data from unintended, unauthorized access, change or destruction. However, increasing volume and sophistication of cyber security threads necessitates the need for developing intelligent techniques which can not only anticipate attacks but also identify new attack types.

PPDM (Privacy preserving data mining) methods have been used for extracting relevant and useful patterns from huge amount of data while protecting sensitive or private information. Several methods including Anonymization-based (K-anonymity [2,3], L-Diversity [1,2] etc), Perturbation-based (Adding-noise [4,5,6], Randomization [4,5] etc) Cryptography-based (Pseudonymization[7,8], Secure Multi-Party Computation [7,8]), Normalization [9,10] have been studied to upheld privacy in the data. The existing approaches are often very complex and time consuming to execute and suffer from problems such as excessive generalization and suppression.

In this paper, a novel data transformation technique has been proposed in this paper for converting the dataset in to a new form while hiding sensitive information and without any data loss.

The remainder of the paper is structured as follows: Section 2 discuss related work in the area. Section 3 includes methodology of the proposed data transformation. Section 4 includes overview of datasets used for experimentation. Section 5 analyzes the experimental results obtained from five popular classifiers. Section 6 summarizes the features of the proposed technique and suggests for the extension.

II. Related Work

Data mining techniques have been successfully used to tackle the privacy preservation. The majority of methods show different ways to transform numerical data to preserve user privacy. These methods include anonymization, data shuffling, perturbation, cryptography, data swapping and normalization.

Jain and Bhandare [11] proposed a data perturbation method based on min-max normalization. This method normalizes an attribute value in to a small specified range. The performance of the method is evaluated on four benchmark datasets. Weka tool [12] was used to test the accuracy before and after distortion and privacy parameters are measured using java code.

Vatsalan et al [13] presented a novel encoding technique for sensitive/private information using Counting Bloom Filters and scalable protocol for privacy preserving record linkage. The experimental results conducted on standard datasets are analyzed in terms of disclosure risk, privacy protection, scalability and linkage quality.

Kargupta and Datta [14] proposed a new Spectral Filtering method using covariance matrix. An analytical framework for the development of new PPDM techniques and steps to reconstruct original data was presented. The performance of proposed technique is analyzed with other popular filter techniques.

Muralidhar and Sarathy [15] compared the working of various data perturbation and data shuffling techniques, the underlying theoretical basis in terms of disclosure risk and data utility. The evaluation of perturbation method for numerical data is discussed in two sections linear and non linear models. Finally, Data shuffling method was proposed to protect numerical data by combining swapping or perturbation approaches.

Samarati and Sweeney [16] used k-anonymity approach to avoid disclosure of entity specific information so that released data cannot be linked to any other external data. The concept of minimal generalization and procedure to transform data are presented to achieve k-anonymity. The anonymized data was a result of generalization and suppression on the original information. The effectiveness of proposed technique is analyzed on medical database.

Chen and Liu [17] proposed a rotation based perturbation method. Method is evaluated on different classifiers to maintain similar accuracy without any data loss. The results from the proposed technique improved the accuracy and privacy quality.

Huang et al [18] presented methods for data reconstruction based on Bayes Estimate and Principal Component Analysis. The proposed techniques are evaluated experimentally and theoretically to analyze the amount of private data disclosed and correlation between them. Experimental results showed that disclosing sensitive information is very high when the correlation between the attribute is high. It was shown empirically that BE-based techniques outperform PCA-based techniques.

Agrawal and Srikanth [19] modified the original values by integrating value distortion and value class-membership methods. A novel method had been proposed to reconstruct original data from perturbed distribution. The perturbation method converts original distribution by using swapping method or simply by adding noise to it. The proposed technique is evaluated using quantitative measures to measure the amount of privacy achieved. The perturbed data with randomization and unperturbed data without randomization were compared in terms of classification accuracy.

Liu et al [20] presented a novel representation for numerical data using two phase data perturbation technique. This technique is applied on various data reconstruction methods to the transformed datasets. Naïve Bayes and decision tree classifiers were applied on reconstructed data using WEKA tool. The proposed two phase perturbation method performance was compared with decision tree and Naive Bayes.

Mohana et al [21] proposed attribute suppression and generalization technique for anonymization. Particle swarm optimization (PSO) algorithm is used to extract optimal attribute set, on which k-anonymity (KA) method is used for classification. Experiments were conducted on UCI dataset to compare the results obtained from KA and PSO technique. The results obtained from NaiveBayes classifier has shown that, the KA outrages PSO in terms of accuracy, precision and recall.

Nguyen and Choi [22] evaluated performance of different classifier algorithms on KDD_CUP'99 dataset [24]. Best algorithms for Probe, Dos, U2R and R2L attack category are identified based on empirical results. Two models for classifier selection were proposed. The experiment results from two models showed minor improvement in True Positive Rate (TPR) for DoS and Probe attacks and low False Positive Rate (FPR) for all categories. Major improvement in R2L and U2R categories were observed.

Sangkatsanee et al [23] proposed an Intrusion Detection System which incorporated a Decision tree approach to classify normal and attack types for online network data. The simulated results showed that proposed technique achieved higher detection rate, low CPU consumption and less time to detect attacks.

III. METHODOLOGY

The original dataset is first tested for the type of attributes it contains as shown in algorithm. The proposed data transformation technique transforms numeric and non-numeric attribute values appropriately as shown in function1 and function2. The distorted dataset thus obtained, is evaluated by popular classifiers. The overall process of Network Intrusion Detection System (NIDS) architecture is depicted in Figure 1.

The proposed data transformation (shown in algorithm) method may be applied for numeric as well as non-numeric attributes to preserve the user privacy of sensitive information as shown in Figure 4. Initially, the attribute values are tested to check whether they are of the type numeric (Integer/Real) or non-numeric (Nominal/Strings). If the attribute is of numeric type, the data is transformed within the range [-1, 1] as shown in Figure 3 and non-numeric data is transformed to binary strings of only 0's and 1's as shown in Figure 2.

If the attribute is of numeric type, the data is transformed within the range [-1, 1] by using formula provided in Eq. (1) where *TV* denotes transformed value, *Max* denotes the maximum value in the training data.

$$\text{TV} = \left(\frac{\text{Value}}{\text{Max+1}}\right) * c_1 - c_2 \qquad (1)$$

Choosing the highest value in the training data helps to transform numeric data in a uniform manner and also to obtain original value from the transformed value. For the transformed value (TV) to lie in the range [-1, 1], the parameters c_1 and c_2 are empirically set 2 and 1 respectively.

Numeric Type Conversion: Let us take the numeric value under consideration as 0.32. Assume *Max* value in dataset is 97, the values of parameters c_1 and c_2 are 2 and 1 respectively. Substituting the values in Eq. (1) gives transformed value of -0.9934021.

Non-Numeric Conversion: For nominal value, the ASCII values of the individual digits are added. The individual digits of the sum thus obtained are again added and the result is subsequently represented in equivalent binary form. For example, let the nominal value be *TCP*. Adding the corresponding ASCII values of individual characters, we get, Sum= T+C+P = 84 + 67 + 80 = 231; representing 231 in binary format takes more number of bits, so to convert in to reduced form, a technique called "Reduced ASCII SUM" method is used. In this process, 231 converted as 2+3+1 = 5. Then, finally this reduced ASCII SUM i.e., 5 is represented as binary string i.e., 101.

The transformed dataset is then provided as input to the classification technique. The classification model thus obtained will be used to classify incoming traffic as normal or attack type. The performance of the NIDS is evaluated using various performance measures viz. accuracy, detection rate (DR), precision, false alarm rate (FAR) and F-Score.

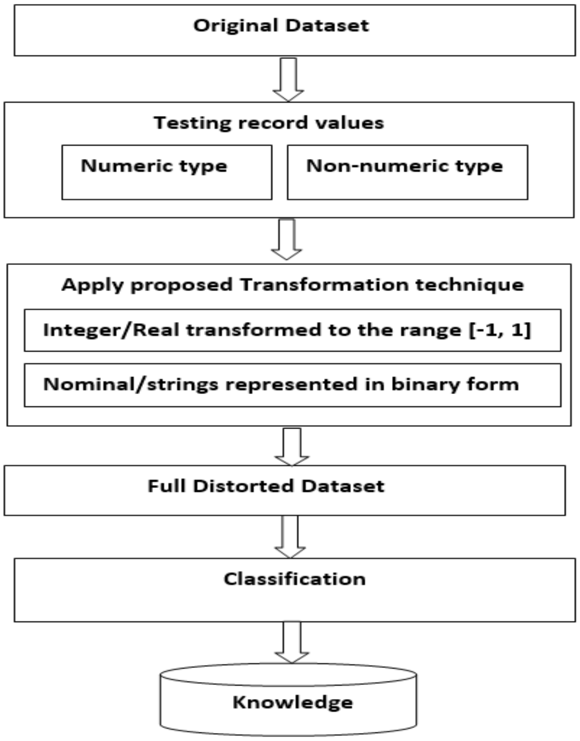
Original Dataset
Testing record values
Numeric type
Non-numeric type
Apply proposed Transformation technique
Integer/Real transformed to the range [-1, 1]
Nominal/strings represented in binary form
Full Distorted Dataset
Classification
Knowledge

Fig. 1 Architecture of NIDS for Sensitive Data

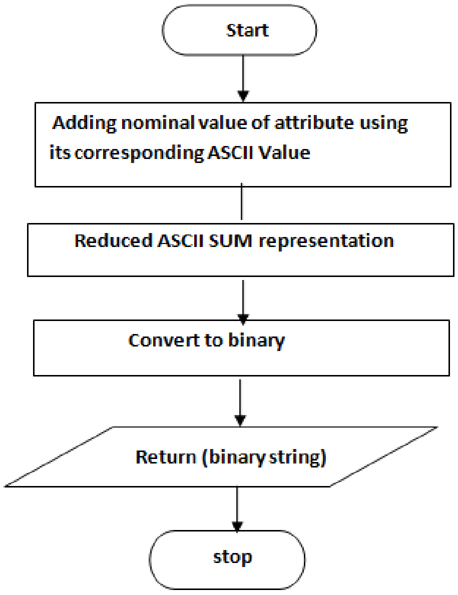
Start
Adding nominal value of attribute using its corresponding ASCII Value
Reduced ASCII SUM representation
Convert to binary
Return (binary string)
stop

Fig. 2 Non-Numeric attribute transformation

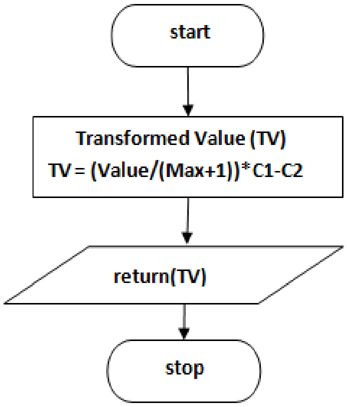

Fig. 3 Numeric attribute transformation

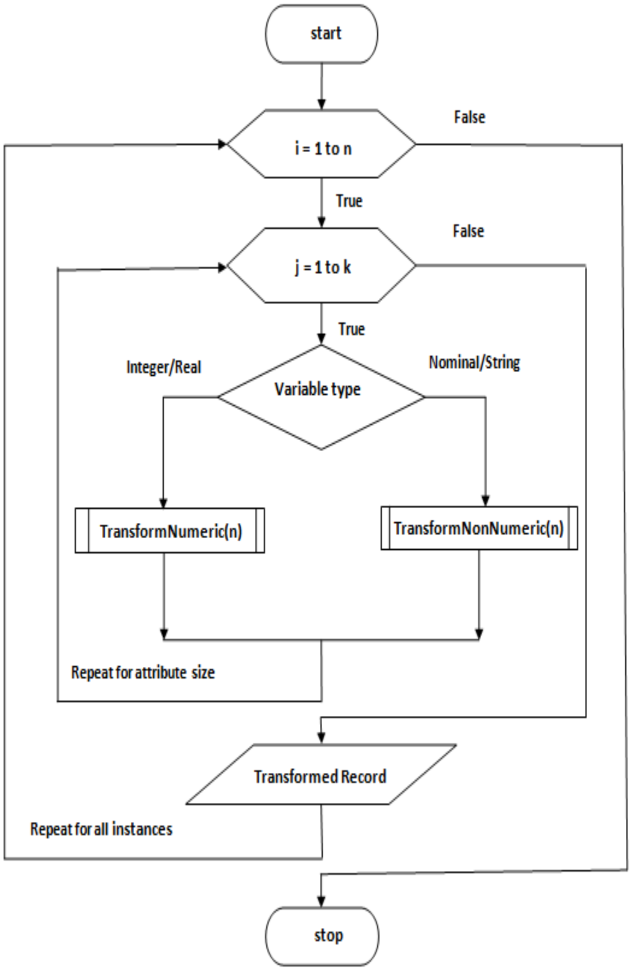

Fig. 4 Data Transformation Process

To assess how the proposed data transformation technique affects the performance of the classifier involved. Data Utility for each dataset is also computed. The evaluation measures such as classification accuracy must remain unchanged with or without applying data transformation. Then the data transformation technique is said to possess good data utility.

Whenever data is transformed, privacy measures [26, 27, 28] must be evaluated to appraise how the data transformation affects behaviour of the original data. Let the original matrix be *A* and its transformed matrix be *TA* both of which are *n*m* matrices. Five privacy measures are used to identify the change in behaviour of the original matrix with respect to the transformed matrix.

Value Difference (VD)

After applying data transformation technique, the values of dataset change. The VD is the difference in relative value of the frobenius norm i.e., the difference of transformed matrix TA from original matrix A as provided in Eq.(2).

$$VD = \frac{||A-TA||_F}{||A||_F} \qquad (2)$$

The frobenius norm of an *n*m* matrix A, with j^{th} row and k^{th} column data denoted by a_{jk}, is calculated as given in Eq.(3).

$$||A||_F = \sqrt{\sum_{j=1}^{n}\sum_{k=1}^{m} a_{jk}^2} \qquad (3)$$

Rank Position (RP): RP indicates the change in rank positions for each feature after applying data transformation technique. The rank positions of each entry in original and transformed dataset may change after the elements are sorted in ascending order. Assume that original dataset *A* has *n* data samples and *m* features.

$Rank_k^j$ denotes the rank position of the k^{th} entry in feature *j*, and $TRank_k^j$ denotes rank position of transformed data A_{kj} .Then RP is calculated as shown in Eq.(4).

$$RP = \frac{\sum_{j=1}^{m}\sum_{k=1}^{n}|Rank_j^i - TRank_k^i|}{m*n} \qquad (4)$$

Rank Maintenance (RK) is defined as the proportion of elements maintaining their column rank positions after transformation. RK is calculated as shown in Eq.(5).

$$RK = \frac{\sum_{j=1}^{m}\sum_{k=1}^{n} Rk_k^j}{m*n} \qquad (5)$$

$$where\ Rk_k^j = \begin{cases} 1 & Rank_k^j = TRank_k^j \\ 0 & otherwise \end{cases}$$

Change of Rank of Features (CP) is used to measure the change of each attribute in its average value and their rank position before and after data transformation as shown in Eq.(6).

$$CP = \frac{\sum_{j=1}^{m} |RankAV_j - TRankAV_j|}{m} \qquad (6)$$

RankAV_j and *TRankAV_j* denotes the rank of attribute *j* before and after transformation.

Maintenance of Rank of Features (CK)

This metric measures the proportion of the features that maintain their ranks of average value after the transformation as shown in Eq.(7).

$$CK = \frac{\sum_{j=1}^{m} Ck_j}{m} \qquad (7)$$

$$where\ Ck_j = \begin{cases} 1 & RankAV_j = TRankAV_j \\ 0 & otherwise \end{cases}$$

The training dataset is initially provided to the Data Transformation algorithm which transforms numeric as well as non-numeric attributes (string).

Algorithm: Data Transformation

```
for each instance $I$ in dataset $D$
for each Attribute $A_K$in instance $I$
where K=1,2,3…no. of attributes - 1
        testValueType($A_K$)
if $A_K$ is Numeric
        transformNumeric($A_K$)
else
        transformNonNumeric($A_K$)
end if
end for
 Write($I$) //store transformed data
end for
```

Function1: transform Numeric (A_K)

```
for each instance $I$ in dataset $D$
for each Attribute $A_k$in instance
where   k=1,2,3…no. of attributes -1
if $A_K$ is Numeric
        $NV = \left(\frac{A_K}{(MAX+1)}\right) * C_1 - C_2$
end if
end for
return ($NV$) //transformed numeric value
end for
```

Function2 : transform Non-Numeric(A_K)

```
for each instance $I$ in dataset $D$
for each Attribute $A_K$in instance I
where K=1,2,3…no. of attributes -1
if $A_K$ is Non-Numeric
for each character $C_j$ in string
   $N = N + C_j$  //ASCII sum
end for
while $N > 0$
$R = N\%10$ ;
   $BS[i++] = R$ ;
   $N = N/2$;
end while
end for
end if
return($BS$) //transformed binary string
end for
```

IV. Dataset description

The proposed technique is analyzed on five benchmark datasets from University of California Irvine, Machine Learning Repository(UCI-ML)[25] viz., Iris, Glass identification, Ionosphere, mush room and zoo dataset. The description of the UCI-ML datasets used in the experiments is provided in Table 1.The number of training/test samples in NSL-KDD dataset for different types of attacks is provided in Table 2. The detailed summary of the attacks in the NSL-KDD datasets is shown in Table 3.

Table 1 : Summary of UCI-Datasets

S.No	Dataset	#Instances	#Attribute	#class
1	Iris	150	4	3
2	Glass Identification	214	10	7
3	Ionosphere	351	34	2
4	Mush room	8124	22	2
5	Zoo	101	17	7

Table 2: Attack-wise Summary of the NSL_KDD Dataset

Attacks	Training samples	Testing Samples
Normal	6503	2569
DoS	4487	1885
Probe	1072	509
U2R	7	1
R2L	93	37
Total samples	12162	5001

Table 3: NSL_KDD Class-wise distribution

Attack	SNO	Class	Training	Testing
Dos	1	land	0	1
	2	neptune	4041	1716
	3	teardrop	87	32
	4	pod	18	9
	5	smurf	258	94
	6	back	83	33
	Total		4487	1885
Probe	7	ipsweep	321	151
	8	portsweep	273	136
	9	nmap	145	56
	10	satan	333	166
	Total		1072	509
U2R	11	buffer_overflow	4	0
	12	rootkit	3	1
	Total		7	1
R2L	13	warezclient	81	33
	14	guess_passwd	3	4
	15	multihop	0	0
	16	ftp_write	0	0
	17	imap	2	0
	18	phf	2	0
	19	warezmaster	5	0
	Total		93	37
Normal	20		6503	2569

V. EXPERIMENTAL RESULTS AND DISCUSSION

All the experiments were conducted on Intel Core i5-2400 CPU 3.10 GHz PC with 4GB of RAM running on 32-bit windows operating system. The effectiveness of the proposed method is evaluated on five benchmark UCI-ML datasets as well as standard NSL_KDD Network Intrusion Detection Dataset.10-fold cross validation is applied on the five UCI-ML datasets whereas the performance of the classifiers on NSL_KDD dataset is validated on the NSL_KDD test set.

Table 4 presents the accuracy obtained from five UCI-ML datasets. Naïve Bayes, IBk, C 4.5 and Random Forest classifiers are evaluated before and after applying the proposed data transformation technique. It can be observed that the accuracy measures for Iris and Ionosphere datasets remain unchanged because both the datasets contain numerical attributes alone. There is slight variation in the accuracy measures of Mushroom and Zoo datasets which can be attributed to the presence of string-valued attributes along with numerical attributes.

Table 4: Accuracy before and after data transformation

Measures	Classifiers	Accuracy Before	Accuracy After
Iris	NaiveBayes	96	96
	IBk	100	100
	C 4.5	98	98
	Random Forest	100	100
Glass Identification	NaiveBayes	85.98	91.12
	IBk	100	100
	C 4.5	100	99.72
	Random Forest	100	99.53
Ionosphere	NaiveBayes	85.98	85.98
	IBk	100	100
	C 4.5	100	100
	Random Forest	99.53	99.53
MushRoom	NaiveBayes	95.88	94.69
	IBk	100	100
	C 4.5	100	100
	Random Forest	100	100
Zoo	NaiveBayes	100	99.01
	IBk	100	100
	C 4.5	99	98.02
	Random Forest	100	100

Table 5: Privacy Measures on the transformed datasets

Datasets	VD	RP	RK	CP	CK
Iris	0.97	0	1	0	1
Glass Identification	1	0	1	0	1
Zoo	0.39	0.79	0.95	0.79	0.94
Inosphere	1.59	0	1	0	1
Mush room	0.90	0	1	0	1
NSL_KDD	1	0.14	0.98	0.001	0.66

Table 5 shows the privacy measures obtained on the datasets. The value difference for the datasets is generally high indicating no correlation between the original value and the transformed value. This offers more privacy to the data. The Rank Position values are also minimum indicating that position of ranks of the attributes has not changed much due to data transformation. The rank maintenance values are close to 1 indicating that most of the elements in all the columns maintained their ranks. The CP values are also close to 0 which show that the rank of the average value of attributes has not altered much during data transformation. High values of CK indicate that most of the attributes retain the ranks of the average value of attributes. The exceptions occurring in the values of VD, RP and CP in Zoo dataset may be attributed to the presence of more number of nominal attributes. In toto, the values of privacy measures indicate that the proposed data transformed technique provides minimal loss of information.

Figure 5 presents the variation in performance measures on the standard NSL_KDD dataset. The variation of accuracy before and after data transformation is applied on the

NSL_KDD dataset. Except for the Naïve Bayes classifier, there exists slight decrease in the accuracy for the other classifiers. Figures 5 also depict similar behaviour among other evaluation measures. It can be observed that the evaluation measures provide almost similar results before and after the proposed transformation technique are employed.

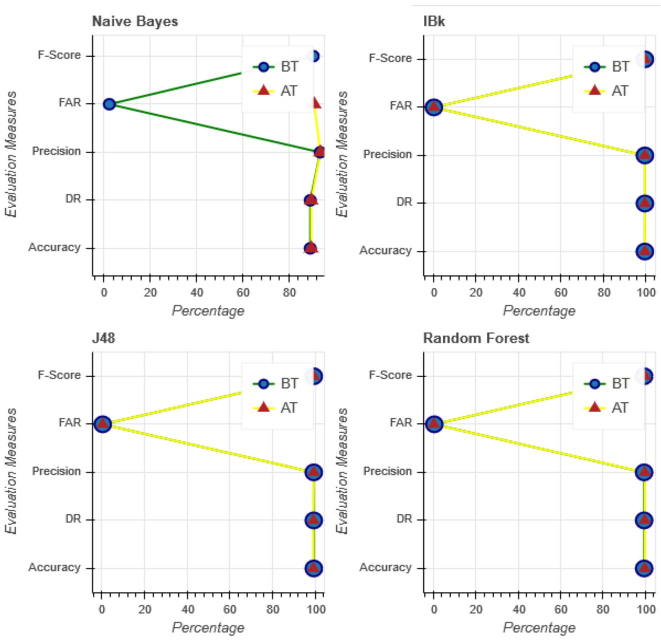

Fig. 5 Performance measures of NSL_KDD Dataset

Figure 6 and Figure 7 shows the time taken in milliseconds for building the classifiers with as well as without using data transformation technique. The minor increase in time taken indicates that the proposed data transformation technique does not put additional overhead on the classifiers.

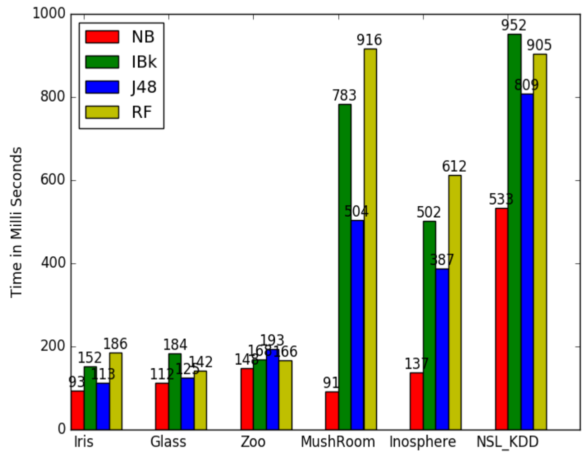

Fig. 6 Classifier building time before Data Transformation

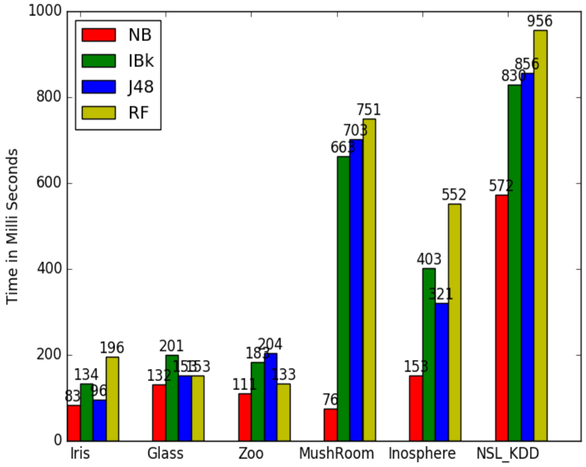

Fig. 7 Classifier building time after Data Transformation

Figure 8 provides the comparison of memory requirements with or without using the data transformation technique. It may be observed that the memory requirements did not alter much, or even decreased in some cases. The presence of real values and binary strings reduce the memory required for building the classifiers. Low overhead on time and memory requirements is especially beneficial for network intrusion detection datasets which are generally high-dimensional in nature.

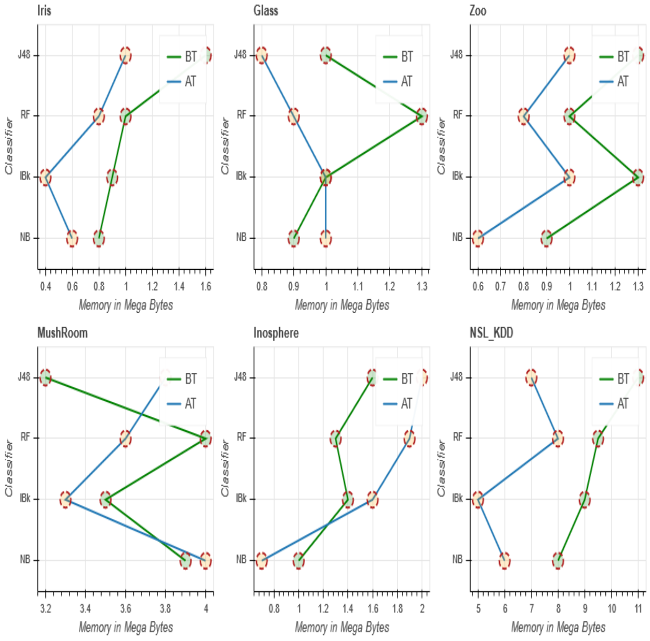

Fig. 8 Memory required for classifiers before (BT) and after Data Transformation (AT).

VI. CONCLUSION AND FUTURE WORK

In this paper, a novel data transformation technique is proposed inorder to preserve sensitive information in the datasets. The proposed technique achieved high degree of data distortion and maintained similar accuracy before and after transformation for all the considered data mining algorithms. The performance of the proposed technique is evaluated in terms of data utility, privacy measures, and time and memory requirements using six benchmark datasets. Results indicate good performance when the method is evaluated using popular classifiers. In future, improved methods for transforming nominal attribute values may be explored.

ACKNOWLEDGEMENT

The authors would like to express their gratitude to the Science and Engineering Research Board (SERB), Ministry of Science & Technology, Govt. of India under grant No. SB/FTP/ETA -0180/2014 for providing support.

REFERENCES

[1] Ashwin Machanavajjhala, Johannes Gehrke, Daniel Kifer,Muthuramakrishnan Venkitasubramaniam,

"*l-diversity: Privacy beyond k-anonymity*", ACM Transactions on Knowledge Discovery from Data (TKDD), Vol.1,No.1,pp.1-12,2007.

[2] Ninghui Li, Tiancheng Li and Suresh Venkatasubramanian,"*t-closeness: Privacy Beyond k-anonymity and l-diversity*", IEEE 23rd International Conference on Data Engineering,IEEE, pp.1-10, 2007.

[3] A. Hussien, N. Hamza and H. Hefny, "*Attacks on anonymization-based privacy-preserving: a survey for data mining and data publishing*",Journal of Information Security, Vol. 4, No. 2, pp. 101-110, 2013.

[4] Yu Zhu and Lei Liu, "*Optimal randomization for privacy preserving data mining*", Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining,ACM, pp.761-766, 2004.

[5] Swapnil Kadam and Navnath Pokale, "*Preserving Data Mining through Data Perturbation*", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), Vol. 4, No. 11, pp. 4128-4131,2015.

[6] Ashish. E. Mane and Sushma Gunjal, "*Privacy preserving using additive perturbation based on multilevel trust in relational streaming data*", Multidisciplinary Journal of Research in Engineering and Technology(MJRET), Vol. 2, No. 2, pp. 392-397,2015.

[7] Wenliang Du and Mikhail J.Atallah, "*Secure multy-party computation problems and their applications: a review and open problems*", Proceedings of the 2001 workshop on new security paradigms, ACM, pp. 13-22, 2001.

[8] Benny Pinkas,"*Cryptographic techniques for privacy-preserving data mining*", ACM Sigkdd Explorations Newsletter,Vol. 4,No. 2, pp. 12-19, 2002.

[9] Syed Md. Tarique Ahmad, Shameemul Haque and Prince Shoeb Khan, "*Privacy Preserving in Data Mining by Normalization*", International Journal of Computer Applications, Vol. 96, No. 4, pp. 14-18, 2014.

[10] C.Saranya and G.Manikandan. "*A Study on normalization techniques for privacy preserving data mining*", International Journal of Engineering and Technology (IJET), Vol. 5, No.3, pp. 2701-2704, 2013.

[11] Yogendra Kumarjain and Santoshkumar Bhandare," *Min max normalization based data perturbation method for privacy protection*", International Journal of Computer & Communication Technology (IJCCT), Vol. 2, No. 8, pp. 45-50, 2011.

[12] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten ,"*The WEKA Data Mining Software: An Update",* SIGKDD Explorations, Vol. 11, No. 1, 2009.

[13] Vatsalan, Dinusha, Peter Christen and Erhard Rahm, "*Scalable Multi-Database Privacy-Preserving Record Linkage using Counting Bloom Filters*", arXiv preprint arXiv:1701.01232, 2017.

[14] Hillol Kargupta,Souptik Datta,Qi Wang and KrishnaMoorthy, "*Random-data perturbation technique and privacy-preserving data mining*", IEEE International Conference on Data Mining,IEEE, pp. 1-19, 2003.

[15] K.Muralidhar and R.Sarathy, "*Perturbation methods for protecting numerical data: Evolution and evaluation*", Proceedings of the 5th Security Conference, 2006.

[16] Pirangela Samarati and Latanya Sweeney, "*Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression*", Technical report, SRI International, pp. 1-19, 1998.

[17] Keke Chen and Ling Liu, "*Privacy preserving data classification with rotation perturbation*", In Proceedings of the 5th IEEE International Conference on Data Mining (ICDM'05), IEEE, pp. 589–592, 2005.

[18] Zhengli Huang, Wenliang Du and Biao Chen." Deriving private information from randomized data", In Proc. of ACM SIGMOD'05, pp. 37-48, 2005.

[19] Rakesh Agrawal and RamaKrishnan Srikant, "*privacy preserving data mining*", Proceedings of the 2000 ACM SIGMOD international conference on Management of data, Vol. 29, No. 2, pp. 439-450, 2000.

[20] Li Liu, Murat Kantarcioglu and Bhavani Thuraisingham, "*The applicability of the perturbation model-based privacy preserving data mining for real-world data*", Sixth IEEE International Conference on Data Mining - Workshops (ICDMW'06), pp. 6-21, 2006.

[21] Mohana, S., S. A. Sahaaya and Arul Mary, "*A COMPARITIVE FRAMEWORK FOR FEATURE SELCTION IN PRIVACY PRESERVING DATA MINING TECHNIQUES USING PSO AND K-ANONUMIZATION*", Emerging Technologies in Networking and Security (ETNS), 2016.

[22] Huy Anh Nguyen and Deokjai Choi, "*Application of data mining to network intrusion detection: classifier selection model*", Asia-Pacific Network Operations and Management SymposiumSpringer Berlin Heidelberg, pp. 399-408, 2008.

[23] Phurivit Sangkatsanee, Naruemon Wattanapongsakorn and Chalermpol Charnsripinyo, "*Real-time Intrusion Detection and Classification*", IEEE network, 2009.

[24] KDDcup99, "*Knowledge discovery in databases DARPAarchive*", http://www.kdd.ics.uci.edu/databases/kddcup99/ task.html, 1999.

[25] Blake, Catherine, and Christopher J. Merz, "*{UCI} Repository of machine learning databases*", 1998.

[26] Shuting Xu,Jun Zhang,Dianwei Han and Jie Wang, "*Data distortion for privacy protection in a terrorist analysis system*", International Conference on Intelligence and Security Informatics, Springer Berlin Heidelberg, pp.459-464, 2005.

[27] Wang, Jie, Weijun Zhong, and Jun Zhang, "*NNMF-based factorization techniques for high-accuracy privacy protection on non-negative-valued datasets*", Sixth IEEE International Conference on Data Mining-Workshops (ICDMW'06), IEEE, 2006.

[28] Jie Wang, Weijun Zhong,Shuting Xu and Jun Zhang, "*Selective Data Distortion via Structural Partition and SSVD for Privacy Preservation*", IKE, pp.1-7, 2006.

Authors Profile

Dr. Sireesha Rodda is a Professor in Department of Computer Science and Engineering at GITAM Deemed to be University, Visakhapatnam, India. She has published more than 20 papers in refereed National and International Journals. Her research interests include machine learning and big data analytics.

Mr.E.Uma Shankar Rao presently working as Assistant Professor in RAGHU Institute of Technology, Visakhapatnam, Andhra Pradesh, India. He received his M.Tech (CSE) degree in the year 2013.He received B.Tech (CSE) degree in the year 2007.His research interest Data mining, Image Processing, Data Structures.