Journal of Web Engineering, Vol. 16, No.3 & 4 (2017) 277-292

A COMPLETE PRIVACY PRESERVATION SYSTEM FOR DATA MINING USING FUNCTION APPROXIMATION

V. RAJALAKSHMI

Assistant Professor, School of Computing,Sathyabama University,
Chennai, India – 600119.
rajalakshmi.it@sathyabamauniversity.ac.in

M. LAKSHMI

Professor, School of Computing, Sathyabama University,
Chennai, India – 600119.
laks@icadsindia.com

V. MARIA ANU

Assistant Professor, School of Computing,Sathyabama University,
Chennai, India – 600119.
Mariaanu18@gmail.com

January 12, 2015
December 17, 2016

Data privacy has become the primary concern in the current scenario as there are many pioneering methods for efficient mining of data. There are many algorithms to preserve privacy and handle the trade-off between privacy and utility. The ultimate goal of these algorithms is to anonymize the data without reducing the utility of them. A Privacy preserving procedure should have a minimum execution time, which is the overhead of clustering algorithms implemented using classical methods. There is also no single procedure that completely handles the trade-off and also updates itself automatically. In this work, the anonymization is implemented using Radial Basis Function [RBF] network, which provides both maximum privacy and utility with a proper tuning parameter specified between privacy and utility. The network also updates itself when the trend of data changes by controlling the maximum amount of error with a threshold value.

Keywords: Privacy preservation, Radial basis function, Function approximation, Data anonymization
Communicated by: B. White & M. Gaedke

1 Introduction

The usage of internet and social networking web sites has increased the amount of globally available data. Because of this, we are drowning in data but starving for knowledge and privacy. These data are provided for mining to retrieve non-trivial knowledge for future decision making. As techniques for revealing non-trivial patterns using various data mining algorithms are explored, the threat towards the data is also increased [35]. When such data are provided for mining in their original form, it forms a

threat for the privacy of an individual. Typical example includes disease of a patient, credit card balance of a customer, purchase details from a departmental store, government weapon details in military, etc., Anonymization issues also occur in surveying, statistical databases, cryptographic computing, access control, social networking and so on. Hence data need to be modified before they areprovided for mining or to any third party for processing. The main part in this process is that the modification done on the data should not affect the mining result and other statistical parameters about the data.

There is an inverse relationship between privacy and utility of the data as shown in Figure 1.For mining, as the exact data is not required, a perfect approximation is sufficient, the modification is accepted. The research which alters the data without modifying the mining results is termed as PPDM [Privacy Preservation in Data Mining].

Figure 1: Trade-off between privacy and utility

Attributes in a database are divided into three types – unique identifying attributes, sensitive attributes, quasi identifying attributes. When data are given for mining unique identifying attributes like patient ID, credit card number, Employee ID, etc., are removed completely from the database. Sensitive attributes like disease, credit card balance, salary, etc., are the primary concerns for mining and they are not altered. Quasi identifying attributes like age, zipcode, height, married, gender, etc., are also available in a public database like voter's list. These are the values which are altered so that the exact individual of the record is not identified.

The information disclosure is categorized into two types [36], Identity disclosure that specifies which record is associated with which individual in a released table and in Attribute disclosure, new information about some individuals is revealed by the released table. In this work, Identity disclosure is handled i.e., the association between an individual and a particular record is tried to be hidden.

PPDM techniques are implemented by Randomization or Perturbation [37]. In randomization, random numbers are generated with less variance and zero mean. These random numbers are then added with the data in additive perturbation and multiplied in multiplicative perturbation. Cryptographic methods are used for multi party handling of data. Perturbation techniques includes anonymization, permutation, swapping, slicing, etc., K-Anonymity is one of the widely implemented technique using generalization and suppression. Generalization refers to altering a value with a less specific but semantically acceptable value, while suppression refers to not releasing a value by hiding partially or completely.

2 Organization of the Paper

Section 3 specifies the various literatures and their related work. Methods which have led to the development of this paper are discussed in this section. Section 4 defines the problem and the objectives of this work. Section 5 describes the usage of RBF network for PPDM, a variation of the network with its coding and corresponding output. Section 6 states an experimental setup, which gives the database used, the software used and the initial cluster reference. Section 7 explains in detail about the various performance measuring parameters related to PPDM techniques and the comparison of methods based on each of the parameters. Section 8 provides the conclusion and future work.

3 Related Work

PPDM was done initially using different anonymization methods like randomization [9]. These methods increase the error as it completely relies on randomly generated values. This can be done as simple as adding random noise values as in [16]. [5] Performs anonymization using a geometric perturbation method and [14] performs using a generic method which do not consider the relationship among the attributes.[22] provides a multi level solution in which one of them can be done using cryptography. Cryptography is directly used in [25] which use a key value to perform anonymization. Data anonymization is done using various methods like generalization [31][33], suppression [17]. K-anonymity is an important step towards PPDM techniques. This has been implemented using decision trees [32], classification [10] or clustering [4][6][13].

There are some variations of K-anonymity in [12] [20]. Since there were drawbacks in K-anonymity, l-diversity was introduced which was improvised as t-closeness in [19] and [29]. But t-closeness is inefficient in terms of time complexity. There came a wise methodology to first cluster the data and then apply anonymization according to the association of each data to a cluster [13][26]. In [18], fuzzy technique is used to implement clustering. Clustering is implemented using genetic algorithms in [24]. Every cluster is rotated using isometric rotation in [7][15][27], which is efficient but is susceptible to similarity attacks and the reconstruction of original data is also straight forward. [16] Performs a noise addition after clustering. In [21], slicing is used to group similar data and then perform anonymization.

Neural networks are efficient systems which can identify similar items naturally and then they can be anonymized easily. In [34], neural network is used to implement clustering. In [30] , back propagation network is used to implement classification of data and in [23] radial basis function[RBF] network is used for privacy preservation.

4 Problem Definition

In order to reduce the execution time of anonymization, to enhance the efficiency in terms of both privacy and utility and also to set a tunable parameter, RBF network is used as a function approximation network which can group similar data along with approximating them. The main objectives of this work are,

- To efficiently anonymize the data

- To provide proper tuning parameter between privacy and utility.
- To provide more privacy with less error and train the system itself when error increases.

5 Implementation

A neural network is a feed forward network and has a single hidden layer of sigmoid function which is capable of approximating uniformly any continuous multivariate function, to any desired degree of accuracy. Clustering was implemented using neural networks, which are efficient systems to identify similar items naturally and their output can undergo anonymization. When Neural networks are used to implement clustering, it reduces the execution time of the procedure. When any usual clustering algorithm is used every data is calculated for its presence in every cluster and after its membership to a cluster is found, the centroid of that cluster changes. This does not happen in a neural network and hence reduces the time of execution.

5.1 Radial Basis Function Network

Radial basis function networks (RBFNs), are special type of neural networks which are being applied for problems such as function approximation, pattern recognition and time series prediction, etc., RBF networks when used directly as an universal approximators of desired accuracy, provide a solution for PPDM but with a less efficiency. The standard RBF network consists of three layers, i.e., the input, hidden and output layers. The hidden layer of an RBF network can be viewed as a function that maps the input patterns from a nonlinear separable space to a linear separable space. In the new space, the responses of the hidden-layer neurons then form a new feature vectors for pattern representation. Each output vector can be assumed as a representation of a group of input patterns. The architecture of a radial basis network and its function approximation for anonymization is specified in figure 2.

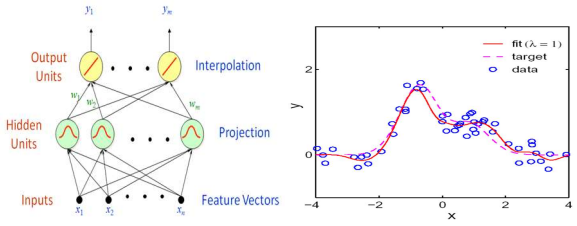


Figure 2: RBF network and its output as function approximators

Given a data set X with size N x m and the output vector Y of same size is shown below:

$$X = \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,m} \\ x_{2,1} & x_{2,2} & \dots & x_{2,m} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ x_{N,1} & x_{N,2} & \dots & x_{N,m} \end{pmatrix} \quad (1)$$

$$Y = \begin{pmatrix} y_{1,1} & y_{1,2} & \dots & y_{1,m} \\ y_{2,1} & y_{2,2} & \dots & y_{2,m} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ y_{N,1} & y_{N,2} & \dots & y_{N,m} \end{pmatrix} \quad (2)$$

Now each row of $X_i = [x_{i,1}\ x_{i,2} \quad \dots \quad x_{i,m}]$ targets a row of $Y_j = [y_{j,1}\ y_{j,2} \ \dots\ y_{j,m}]$. We want to find a target function f (x_i) , that produces the lowest error when predicting the unknown related values Yj. This is equivalent to determining the weight vector W for finding Y with minimum error.

$W = [w_1, w_2, \dots w_p]$. (3)

$Y = f(x)$ (4)

Using Radial Basis Network the function is chosen as a radial basis function as follows.

$$f(x) = \sum_{k=1}^{p} w_k \varphi_k(x) \quad (5)$$

The radial function can be specified as

$$\varphi_k(x) = \varphi(\|x - x_k\| \quad (6)$$

where x_k is the center of the activated neuron.

The three main parameters of a radial basis function are

- Centre X_k
- Distance Measure $r = \| x - x_k \|$
- Shape of the radial basis function

Figure 2 specifies the how an RBF network is given input for anonymization and gives a sample output of function approximators. It shows that the output of the network follows the input and they are not similar values as input.

For training the network, a training set of data is chosen and the network is built.

$$T = \{(x^k, y^k)\}_{k-1}^{p} \quad (7)$$

The goal is set as

$$Y^{(k)} \approx f(x^{(k)}) \quad (8)$$

5.2 A Direct Function approximation (FUNAP)

The function approximation in a radial basis network produces an output similar or close to that of the input. This quality is used directly for data anonymization. Every data undergoes a modification by the radial function based on its distance from the activated neuron center. By this, data are clustered and they remain close to their center. As we do not want the exact values of the input, the goal is set to a nominal value which will produce an approximate value to that of the input. The parameters of the network are varied for different levels of performance of the network. The spread and the number of neurons are used for variation. Spread controls the amount of error produced by approximation and hence the amount of privacy preserved for the data. Number of neurons controls the inputs responding to a closer centroid neuron. This is related to the amount of error and hence the mis-classification error. This explains the quality of anonymized data.

A directfunctionapproximationisverymuchsensitiveonlytonetworkparameters. If the parameters are identified, the anonymization becomes reversible and the data gives same output for every similar group of inputs. FUNAP is also susceptible to homogeneity problem. To overcome these issues, a variation in FUNAP is done using isometric rotation. After choosing the neuron which responds for the input the data is rotated with respect to the neuron centre using randomly generated angles. Isometric rotation ensures that the proximity of data with respect to the centre remains the same. The method also responds differently for the same inputs and it is also not reversible since we use random values for angles.

5.3 Function Approximation towards Center (FUNAPTOC)

Using random angles the mobility of data also becomes random, which increases error in anonymized data. Network parameters also have limited range for the best performance. Hence an optimal method includes both displacement and rotation of data. After rotation using the Euclidean distance between the data and centroid, every data is moved towards the centre which reduces the error. The method provides best results in terms of privacy and utility. The procedure is also not reversible as it is a two-fold process.

Figure 3explains the architecture of FUNAPTOC system, which trains a network, rotates with respect to the activated center and moves them towards its center. This way a space for data alteration is

increased and reduces the error. The displacement has both positive and negative values depending on the original value of the attribute.

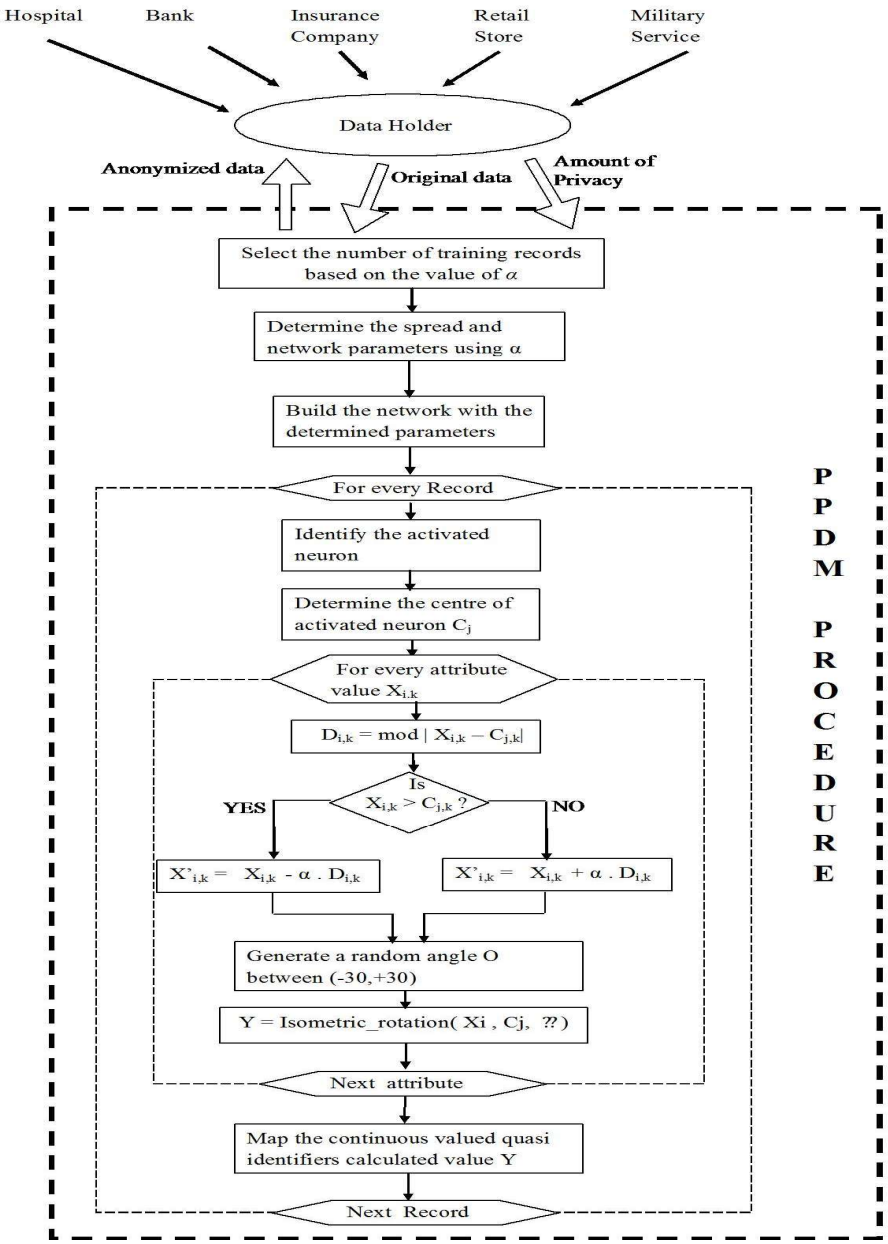


Figure 3: Flow diagram of FUNAPTOC system

Figure 4 gives the code of FUNAPTOC, which moves the data towards center, which is based on the distance between the data and the center.

```
randum1 =rand(1000,1);
for(i=1:n)
     y3=y1(:,i);
     y3=y3';
     out=distfcm(y3,w1);
     ans=min(out);
     j=find(out==ans);
     d1=(w1(j,1)-y3(1,1))/2;
     d2=(w1(j,2)-y3(1,2))/2;
     Y(i,1)=w1(j,1)+d1 * randum1(i);
     Y(i,2)=w1(j,2)+d2 * randum1(i);
end
```

Figure 4: Code for FUNAPTOC

If FUNAPTOC is only used without Isometric rotation, then data are bound to symmetric attacks, where there is a possibility for all the data to be similar to that of the centre value. The anonymization does not perform any change to the data. Hence data are first rotated and then moved towards their centers.

Though an RBF network can be designed to handle any number of inputs with similar number of outputs, only two attributes are shown for anonymization and clustering.

6 Experimental Setup

The Adult data set similar to the one from UCI repository is used for implementing the method. The data set contains 30,162 records after preprocessing. Attributes which are selected for the processing and their properties are shown in Table 1. The attribute severity of disease is chosen to be the sensitive attribute. Matlab is used to implement radial basis network and verify the results. The design of the network is done for different spread, goal and number of neurons and the optimal output is selected based on the performance metrics. Matlab is chosen because it provides flexibility in altering the RBF network parameters for tuning and also provides graphical representation of outputs.

The graphs show the performance of three methods for various sizes of training vectors and compared for various performance measuring parameters. The formulae used for calculating the parameter values are specified in section 6. The results are compared with the greedy based sequencing and rotation based on clustering the data [28].

Table 1: Description of adult data set

.

S.No	Attribute name	Attribute type	No. of distinct values
1	Age	Continuous	74
2	Work-class	Categorical	8
3	Education	Categorical	16
4	Country	Categorical	41
5	Marital-status	Categorical	7
6	Race	Categorical	5
7	Gender	Categorical	2
8	Hours per week	Continuous	58
9	Severity of disease	Categorical	3

Data are initially clustered into three groups and are termed according to their severity of disease as Low, Moderate and High.

7 Performance Measurement

The performance of the procedures is measured using the following parameters:

1. Bias in centroid values
2. Sum of Squared error
3. Rate of classification error
4. Amount of privacy preservation
5. Amount of Information distortion

For all the comparisons the base-line algorithm is chosen to be rotation on greedy based sequencing of data.

7.1 Bias in centroid values

Input data are first clustered and then undergoes anonymization. The utility of the data is maintained if data are centered to a similar place. Hence the centroid values are compared before and after anonymization. If the centers are closer the bias will be less exhibiting that the data has similar utility with that of the original data. As the bias increases the utility of the data reduces. The following table compares all the methods with respect to centroid values.

Table 2: Centroid values based on two attributes – Age and Hours Per Week (HPW).

Cluster	**Original Centroid**		**After Greedy**		**After FUNAP**		**After FUNAPTOC**	
Attribute	**Age**	**HPW**	**Age**	**HPW**	**Age**	**HPW**	**Age**	**HPW**
Low	26.76	33.83	25.62	35.63	25.27	32.70	24.01	37.58
Moderate	39.20	56.59	39.31	47.71	35.95	47.24	39.02	53.62
High	53.30	39.99	56.51	39.22	54.78	42.66	52.44	42.30

7.2 Sum of Squared error

The global error function is the residual sum of squared error is given by

$$SSE = \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2 \qquad (9)$$

Table 3: Comparison of Error Values

Method	**SSE value**
Greedy	1168
FUNAP	1333
FUNAPTOC	653

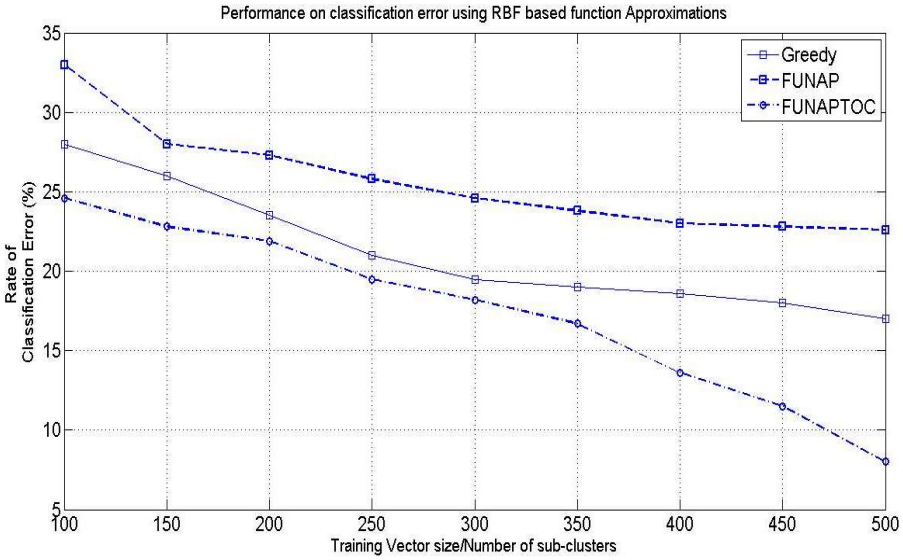
Table 3 specifies the amount of variation of SSE for the proposed methods. The variation should be controlled and minimum amount of variation is considered to be the best. This value is directly proportional to the bias in centroid values. Using the analysis, it is shown that after adding the error of all the 30K data FUNAPTOC performs better than the other methods.

7.3 Rate of classification error

Original data is clustered into three clusters as low, medium and high as they are named according to the severity of diseases using FCM based clustering algorithm. After anonymization, data are given for same clustering algorithm and the outputs are compared. For comparing the performance, any data mining method can be chosen and in this work clustering is chosen. The results are compared with respect to the shifting of data between the clusters. The data shifting between low and high is considered more serious compared to shifting between low and medium, medium and high. Using the analysis, FUNAPTOC performs best compared to other two methods. Table 4 explains different methods with change of their methods for each of the clusters. Graph in figure 4 specifies the performance of methods for classification error for different tunable parameters.

Table 4: Comparison of classification error percentage between different clusters

Method	Low to Mod.	Low to High	Mod. To Low	Mod. To High	High to Low	High to Mod.	Total % of Error
Greedy	0.08	0.02	0.01	0.01	0.02	0.03	17
FUNAP	0.09	0.005	0.08	0.03	0.001	0.02	22.6
FUNAPTOC	0.02	0	0.02	0.02	0	0.02	8



7.4

Figure 4: Comparison on Classification error

A database is privacy preserved if there is less probability for associating any transaction with its sensitive attribute. This is a very important parameter as it specifies the main requirement of the algorithm. It is measured by the number of data altered from its original value to the total number of data considered. If the records remain unaltered, then they are said to be unprotected. The amount of privacy preservation is calculated using

Hidden Failure, HF = U/n (10)

.

U = Number of unaltered records which are bound to insecurity

n = Total number of records.

The graph in figure 5 shows the performance based on privacy for different methods.

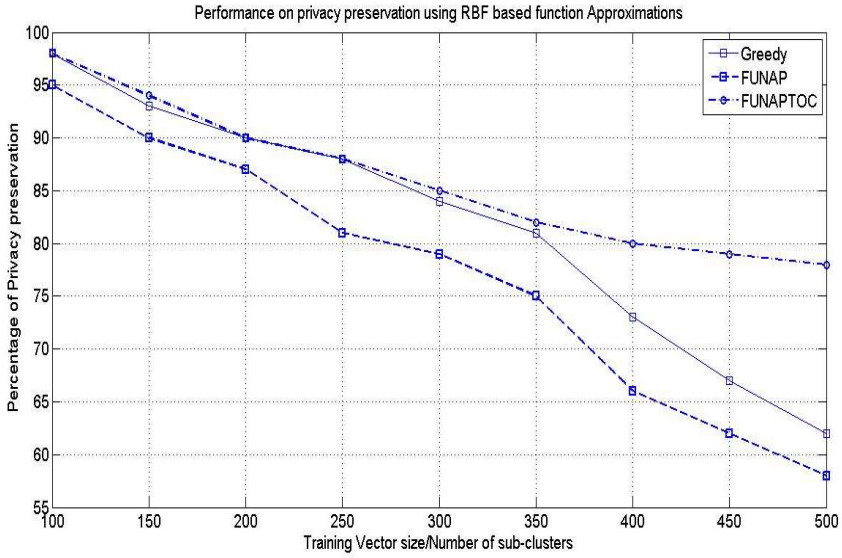


Figure 5 : Comparison on amount of Privacy

The graph shows that more privacy is achieved for FUNAPTOC compared to other methods. It also shows that the performance comparatively increases if the size of the training vector is increased.

7.5 Amount of Information distortion

Information distortion can be calculated from the difference between the original table and the anonymized table. It can also be calculated as the distribution of data with respect to the centroid. The information distortion can be calculated using the following equations. The dissimilarity of record i in j^{th} attribute with respect to centroid $c_k$is given by,

$$\text{diss}\,(r_{i,j},c_{k,j}) = [r_{i,j} - c_{k,j}]^2 \qquad (11)$$

The distortion of all records is given by,

$$D = \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{j=1}^{m} u_{ik} * diss(r_{ij}, c_{kj}) \qquad (12)$$

Where u_{ik} specifies the membership of i^{th} record in k^{th} cluster.

$$\sum_{k=1}^{K} u_{ik} = 1 \qquad (13)$$

$u_{ik} \in \{0,1\}$

The graph in figure 6 shows the performance of methods based on information distortion.

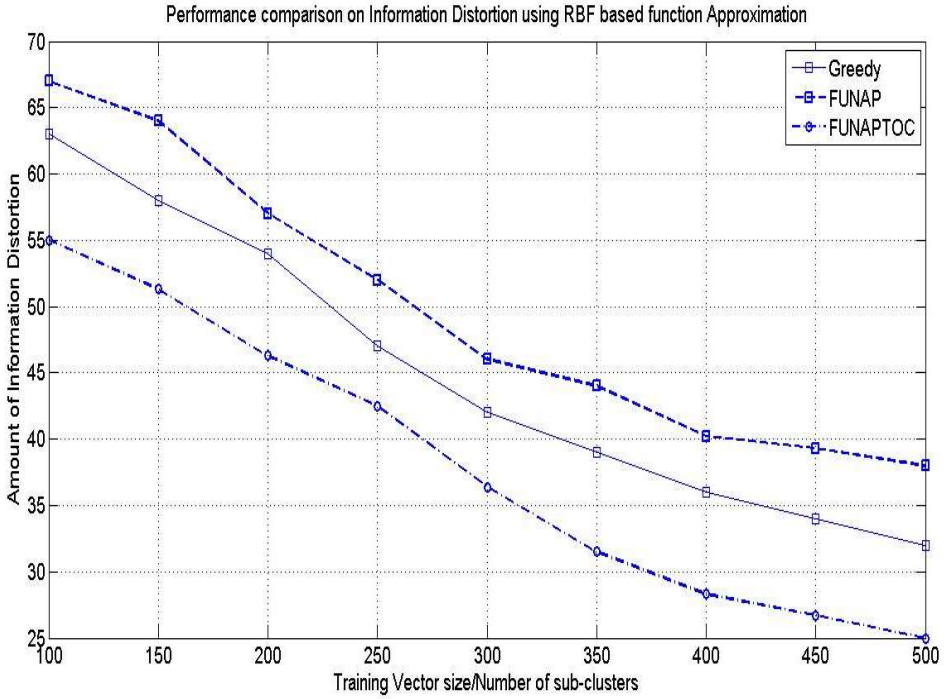


Figure 6: Comparison on amount of Information distortion

8 Conclusion and Future Work

Thus the proposed work explains the use of RBF network as function approximators in PPDM techniques. The variation done on the network performs better than the direct approximation. Implementing and training the RBF network is a onetime process, which can be updated by fixing a threshold error value. FUNAPTOC performs better as it provides a space for altering the network parameters like spread and number of neurons to provide accepted results. With less information distortion and less classification error, FUNAPTOC provides more privacy. The methods are compared with the existing greedy method of clustering and shown that using RBF network gives a better result in terms of information distortion, classification error and amount of privacy. The amount of privacy and utility are controlled with the parameters of the radial basis network – Spread, number of neurons and training vector size. Hence they are used as tuning parameters to choose the required amount of privacy and utility. As a future work, the network can be implemented to handle different types of data since neural network can easily adopt different data types. By monitoring the performance periodically and updating the RBF network improves the efficiency of the system.

REFERENCES

[1] Aggarwal. C.C, Yu.P.S (2004) : A CondensationBasedApproachtoPrivacyPreserving Data Mining.*Proceedings of the EDBT Conference*, pp. 183--199.

[2] Aggarwal, Charu C., and S. Yu Philip (2008): A general survey of privacy-preserving data miningmodels and algorithms. *Springer US*, DOI : 10.1007/978-0-387-70992-5_2

[3] Buratovic, Ines, Mario Milicevic, and KrunoslavZubrinic. (2012):Effects of data anonymizationonthe data miningresults. *MIPRO, Proceedings of the 35th International Convention. IEEE*, pp.1619 – 1623.

[4] Byun, Ji-Won (2007): Efficient k-anonymizationusingclusteringtechniques. *Advances in Databases: Concepts, Systems and Applications*. *SpringerBerlin Heidelberg*, pp. 188-200.

[5] Chen, K., &Liu, L. (2009): Privacy-preservingmultipartycollaborativeminingwithgeometric data perturbation. *Parallel and DistributedSystems, IEEE Transactionson*, vol.*20*(12), pp. 1764-1776.

[6] Chiu, Chuang-Cheng, and Chieh-Yuan Tsai (2007): A k-anonymityclusteringmethodforeffective data privacypreservation. *Advanced Data Mining and Applications*. SpringerBerlin Heidelberg, pp. 89-99.

[7] Dhiraj, S. S., Khan, A., Khan, W., &Challagalla, A. (2009): Privacypreservation in k-meansclusteringbyclusterrotation. *In TENCON 2009-2009 IEEE Region 10 Conference*, pp. 1-7.

[8] Evfimievski. A., Srikant. R., Agrawal. R. and Gehrke. J. (2004), "Privacypreservingmining of association rules of InformationSystems, Vol. 29, No.4, pp. 343-364.

[9] Fong. P. K., and Weber-Jahnke. J. H. (2012), "Privacy preserving decision tree learning using unrealized data sets", IEEE TransactionsonKnowledge and Data Engineering, Vol. 24, No.2, pp.353-364.

[10] Fung, Benjamin CM, Ke Wang, and Philip S. Yu. (2007): Anonymizingclassification data forprivacypreservation. *Knowledge and Data Engineering, IEEE Transactions on*19.5, pp. 711-725, DOI: 10.1109/TKDE.2007.1015.

[11] Ghinita, Gabriel, PanosKalnis, and Yufei Tao (2011): Anonymouspublication of sensitivetransactional data.*Knowledge and Data Engineering, IEEE Transactionson 23.2*,pp. 161-174,DOI: 10.1109/TKDE.2010.101.

[12] Gionis, Aristides, and TamirTassa.(2009): k-Anonymizationwithminimalloss of information. *Knowledge and Data Engineering, IEEE Transactionson* 21.2, DOI: 206-219,10.1109/TKDE.2008.129.

[13] Guo, Kun, and Qishan Zhang (2013) : Fastclustering-basedanonymizationapproacheswith time constraintsfor data streams.*Knowledge-Based Systems*46, pp. 95-108.

[14] Han, S., Ng, W. K., Wan, L., & Lee, V. (2010). Privacy-preservinggradient-descentmethods. *Knowledge and Data Engineering, IEEE Transactions on*,*22*(6),pp. 884-899.

[15] Hong, D., &Mohaisen, A. (2010). AugmentedRotation-BasedTransformationforPrivacy-Preserving Data Clustering. *ETRI Journal*, *32*(3), pp.351-361.

[16] Islam, M. Z., &Brankovic, L. (2011). Privacypreserving data mining: A noiseadditionframeworkusing a novel clusteringtechnique. *Knowledge-BasedSystems*, *24*(8), pp. 1214-1223.

[17] Kisilevich, Slava, et al. "Efficient multidimensional suppressionfor k-anonymity."*Knowledge and Data Engineering, IEEE Transactionson* 22.3 (2010): 334-347,DOI : 10.1109/TKDE.2009.91.

[18] Kumar, Pradeep, KishoreIndukuriVarma, and AshishSureka (2011): Fuzzybasedclusteringalgorithmforprivacypreserving data mining. *International Journal of Business InformationSystems* 7.1,pp. 27-40.

[19] Li, Ninghui, Tiancheng Li, and SureshVenkatasubramanian.(2010): "Closeness: A new privacymeasurefor data publishing." *Knowledge and Data Engineering, IEEE Transactionson* 22.7 ,pp. 943-956, DOI:10.1109/TKDE.2009.139.

[20] Li, T., & Li, N. (2008): Towardsoptimal k-anonymization. *Data &KnowledgeEngineering*, *65*(1), pp.22-39.

[21] Li, T., Li, N., Zhang, J., &Molloy, I. (2012): Slicing: A new approachforprivacypreserving data publishing. *Knowledge and Data Engineering, IEEE Transactionson*, *24*(3), pp.561-574.

[22] Li, Yaping, et al.(2012): "Enablingmultilevel trust in privacypreserving data mining." Knowledge and Data Engineering, IEEE Transactionson 24.9, pp. 1598-1612,DOI:10.1109/TKDE.2011.124

[23] Lin, C. J., Chen, C. H., & Lee, C. Y. (2004): A self-adaptive quantum radial basisfunctionnetworkforclassificationapplications. In *Neural Networks, 2004. Proceedings. 2004 IEEE International JointConferenceon* ,Vol. 4, pp. 3263-3268.

[24] Lin, Jun-Lin, and Meng-ChengWei.(2009): Geneticalgorithm-basedclusteringapproachfor k – anonymization. *ExpertSystemswithApplications* 36.6, pp. 9784-9792.

[25] Liu. Q., Shen. H., and Sang, Y. (2015), "Privacy-preserving data publishingformultiplenumericalsensitiveattributes", TsinghuaScience and Technology, Vol. 20, No.3, pp.246-254.

[26] Ni, Weiwei, and Zhihong Chong.(2012) : Clustering-orientedprivacy-preserving data publishing. *Knowledge-BasedSystems* , Vol.35 , pp. 264-270.

[27] Oliveira, Stanley RM, and O. Zaiane. (2004) : Data perturbationbyrotationforprivacy-preservingclustering.*TechnicalReport TR04*, Vol. 17.

[28] Rajalakshmi.V, &AnandhaMala.G. S.(2014): IsometricRelocation of Data bySequencing of Sub-ClustersforPrivacyPreservation in Data Mining. *International Journal of Engineering&Technology)*, Vol.6, Issue 2.

[29] Rebollo-Monedero, David, Jordi Forne, and Josep Domingo-Ferrer.(2010): From t-closeness-likeprivacyto post randomizationviainformationtheory. *Knowledge and Data Engineering, IEEE Transactionson* 22.11, pp.1623-1636,DOI: 10.1109/TKDE.2009.190.

[30] Samet, S., &Miri, A. (2012). Privacy-preserving back-propagation and extreme learning machine algorithms. *Data &KnowledgeEngineering*, Vol. *79*, pp. 40-61.

[31] Sun. X., Sun. L., and Wang. H. (2011), "Extended k-anonymitymodelsagainstsensitiveattributedisclosure", ComputerCommunications, Vol.34, No.4, pp. 526-535.

[32] Tamersoy, A., Loukides, G., Nergiz, M. E., Saygin, Y., &Malin, B. (2012). Anonymization of longitudinal electronic medical records. *InformationTechnology in Biomedicine, IEEE Transactionson*, *16*(3) , pp. 413-423.

[33] Tao, Yufei, HekangChen, XiaokuiXiao, ShuigengZhou, and Donghui Zhang.(2009): Angel: Enhancingtheutility of generalizationforprivacypreservingpublication.*Knowledge and Data Engineering, IEEE Transactionson* 21, no. 7 ,pp. 1073-1087,DOI: 10.1109/TKDE.2009.65.

[34] Tsiafoulis, S. G., &Zorkadis, V. C. (2010): A Neural Network ClusteringBasedAlgorithmforPrivacyPreserving Data Mining. In*ComputationalIntelligence and Security (CIS), 2010 International Conferenceon,* pp. 401-405.

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

[36] Wahlstrom. K., Roddick. J. F., Sarre. R., Estivill-Castro. V., and de Vries. D. (2009), "Legal and technicalissues of privacypreservation in data mining", Encyclopedia of Data Warehousing and Mining, SecondEdition**,** pp. 1158-1163**.**

[37] Wang, Jian, (2009): "A surveyonprivacypreserving data mining." *DatabaseTechnology and Applications, First International Workshopon. IEEE,* DOI : 10.1109/DBTA.2009.147.