Sensitive Data Anonymization Using Genetic Algorithms for SOM-based Clustering

Fatemeh Amiri[1,2], Gerald Quirchmayr[1,2], Peter Kieseberg [3]
[1]University of Vienna, Vienna, Department of Computer Science, Austria
[2]SBA Research Institute, Vienna, Austria
[3]St. Poelten University of Applied Sciences, St. Poelten, Austria
Email: amirif86@univie.ac.at, gerald.quirchmayr@univie.ac.at, pkieseberg@sba-research.org

Abstract—**Improving privacy protection by using smart methods has become a major focus in current research. However, despite all the technological compensations through analyzing privacy concerns, the literature does not yet provide evidence of frameworks and methods that enable privacy protection from multiple perspectives and take into account the privacy of sensitive data with regard to accuracy and efficiency of the general processes in the system. In our work, we focus on sensitive data protection based on the idea of a Self-Organizing Map (SOM) and try to anonymize sensitive data with Genetic Algorithms (GAs) techniques in order to improve privacy without significantly deteriorating the accuracy and efficiency of the overall process. We organize the dataset in subspaces according to their information theoretical distance to each other in distributed local servers and then generalize attribute values to the minimum extent required so that both the data disclosure probability and the information loss are kept to a negligible minimum. Our analysis shows that our protocol offers clustering without greatly exposing individual privacy and causes only negligible superfluous costs and information loss because of privacy requirements.**

Keywords-Privacy-preserving; Big Data; Clustering; Kohonen's map; SOM; Genetic Algorithms.

I. INTRODUCTION

Owing to communal advantages, privacy-preserving data mining in e-business applications is very attractive [3][26]. The huge volume of data in e-business holds by big data and data mining methodologies. Data mining tasks can lead to the identification of data subjects as well as the disclosure of personal data. To address this problem, at first sight, contradicting requirements, privacy-preserving data mining techniques have been proposed [1][11][21]. Presenting privacy measures within data mining tasks enable them to become more popular and widespread; however, such measures may bring considerable costs and some difficulties concerning the topic of privacy-preserving systems. First, privacy measures require extra computational and storage costs that contribute to the scalability issues. Also, due to the privacy-preserving measures, it becomes an issue to run protected operations with reasonable accuracy [4].

Among the most popular algorithms in the data mining research community address, soft computing methods seem to be more capable to bring optimal solutions [2]. They apply generalization and suppression methods to the original datasets in order to preserve the anonymity of individuals data refer to. Privacy-Preserving on distributed data is important for both online companies and users due to mutual rewards. However, companies do not want to give up competitive knowledge advantages or violate anti-trust law [5].

Among data mining tasks, SOM as an unsupervised competitive learning works well on dividing an input data into closest clusters. SOM cluster approach improves the online computational complexity and expands the scalability of the recommendation process [24]. To implement SOM safely, we design our method based on the Genetic Algorithm. GAs [15] have recently become increasingly important for researchers in solving difficult problems. GAs could provide reasonable solutions in a limited amount of time. They are adaptive heuristic search algorithms derived from the evolutionary ideas of natural selection and genetics [6]. In this study, we propose a method for hiding sensitive data on Horizontally Distributed Data (HDD) among multiple parties without greatly jeopardizing their uniqueness. We assume that n users' preferences for m items are horizontally partitioned among L parties. Users are grouped into various clusters using SOM clustering off-line. After determining n's cluster, those users in that cluster are considered the best similar k users to each other. As off-line costs are not critical to the success of overall actions, our scheme performs GA reducing computations off-line. We analyze the scheme in terms of privacy and performance and perform real-data-based experiments for accurate analysis. Using our method, the local servers can overcome coverage and accuracy problems through partnership. Additionally, as they do not reveal their private data (by running our GA method) to each other, they do not face privacy issues. Let T be the whole data which is partitioned between K companies. Each local unit L holds TL, where TL is a nL×m matrix, k = 1, 2, ..., L; and nL shows the number of users whose data held by the unit L. Thus, each local unit L holds the ratings of nL users for the same m items. Figure 1 shows the first glance of the proposed model in this paper.

The contributions of the paper can be listed, as follows: (i) we propose a novel SOM method utilizing hiding sensitive items of information to alleviate privacy-preserving

 ISBN: 978-1-61208-661-3

problems. (ii) We employ privacy-preserving measures to provide a sufficient level of privacy to individuals. (iii) We show the applicability of soft clustering techniques to the distributed framework to overcome scalability issues. (iv) We also show a comparison among utilized Traditional SOM technique with the proposed method. To the best of our knowledge, our paper presents the first analyses and evaluation on hiding sensitive information in SOM-based clustering on a distributed framework using GAs.

The remainder of this paper is organized as follows: Related work on privacy-preservation in SOM computing is reviewed in Section II. Section III discusses some technical preliminaries employed in the sequence of this paper. The presented protocol to protect transaction data against sensitive item disclosure based on Genetic Algorithms is described in Section IV. In Section V, we evaluate the data utility of the proposed protocol with real datasets. Finally, in Section VI, we summarize the conclusions of our study and outline future research directions.

II. RELATED WORK

To preserve privacy for partitioned data some methods have already proposed. Such studies help data owners cooperate when they own inadequate data and need to combine their fragmented data for improved facilities. A privacy-preserving ID3 algorithm based on cryptographic techniques for horizontally partitioned data is proposed by Lindell and Pikans [22] and followed by Clifton [5]. Vaidya and Clifton [27] presented privacy-preserving association rule mining for vertically partitioned data based on the secure scalar product protocol involving two parties. Privacy-preserving Naïve Bayes classier is also another common method to solve privacy issue in partitioned data [8][19][30].

SOM suffer from its considerable amount of communications in training steps that account for some security and privacy gaps. The number of studies on privacy-preserving in SOM is limited. The first study on solving this issue on SOM has been done by Han [13] that proposed a protocol for two parties each holding a private, vertical data partition to jointly and securely perform SOM. Kaleli and Polat[18] proposed a Homomorphic encryption privacy-preserving scheme to produce SOM clustering-based recommendations on vertically distributed data among multiple parties. They use this encryption, which is employed to privately encrypt and decrypt user vectors to avoid exposing of individual data. Bilge and his partners [4] focus on privacy-preserving schemes applied on clustering-based recommendations to produce referrals without greatly jeopardizing users' privacy. They investigate the accuracy and performance consequences of applying RPTs to some clustering-based CF schemes. Kaleli in [17] proposes offline SOM clustering with least jeopardizing the secrecy. He used the offline local server to run SOM independently in order to decrease the number of communications.

Soft computing methods in recent years brought novel results in privacy-preserving issue in different scenarios. One of the novel soft techniques is GAs. GAs are the search techniques, which are designed and developed to find a set of feasible solutions in a limited amount of time [29]. Fewer studies have adopted GAs to find optimal solutions to hide sensitive information. Han and Ng [12] presented a privacy-preserving genetic algorithm for rule discovery for arbitrarily partitioned data. To achieve data privacy of the participant parties, secure scalar product protocols were applied to securely evaluate the fitness value. Dehkordi [6] introduced a new multi-objective method for hiding sensitive association rules using GAs. The objective of their method is to support the security of database and to keep the effectiveness and certainty of mined rules at the highest level. In the proposed framework, four sanitization strategies were proposed with a different criterion. Lin et al.

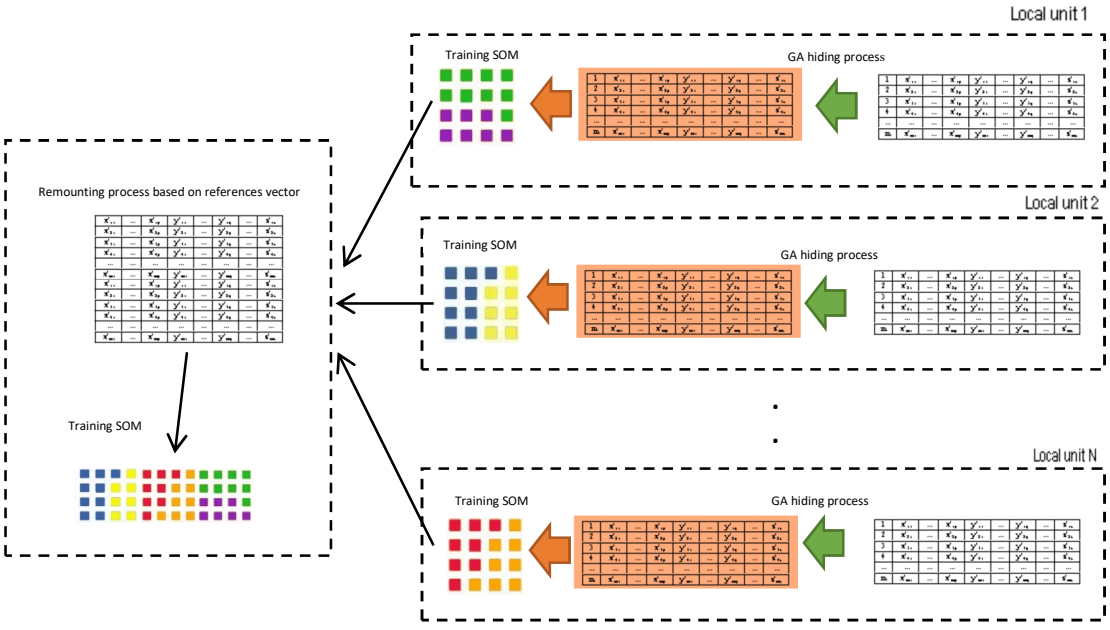

Figure 1. Overall scheme of GASOM protocol.

[20] compact pre-large GA-based algorithm to delete transactions for hiding sensitive items is thus proposed. Their method tries to refine the limitations of the evolutionary process by adopting the compact GA-based mechanism and also the pre-large concept.

To the best of our knowledge, there is no work to date on a privacy-preserving version of SOM using GAs in distributed servers. In this paper, we propose a protocol for multiple parties each holding a private, horizontal data partition to jointly and securely perform SOM. We prove that our protocol is correct and safe in front of some defined privacy attacks.

III. PRELIMINARY CONCEPTS

In this section, we review some technical preliminaries employed in our designed algorithms that are used in the sequence of this paper.

A. Problem definition

One of the most common techniques used to protect personal knowledge from disclosure in data mining is Hiding sensitive data [20]. In this paper, a hiding utility algorithm is proposed to hide sensitive items through optimal transaction deletion. To evaluate whether the transactions are required to be deleted for hiding the sensitive items, the hiding failure parameter is thus concerned. The transactions with any of the sensitive items are first evaluated by the GA algorithm designed to find the minimal hiding failure values among transactions. These transactions will be directly removed from the database. The procedure is thus repeated until all sensitive items are hidden. The reduced dataset is then sent for SOM training by local servers.

Definition 1. (SOM training) The SOM architecture entails of two fully connected layers: an input layer and a Kohonen's layer also called topology-preserving maps [31]. The steps of SOM clustering algorithm and the constants used in the algorithm are described in the following [9].

Based on the constants defined by Haykin [14], to find the Kohonen's layer neuron a random object x is selected from input data X and the winning Kohonen's Neuron (KNi) is determined by the computed minimum Euclidean distance between x and Wj using (1) as follows. Wj represents initial weights chosen randomly among objects in X for j = 1, 2, …, T, where T shows a number of neurons in Kohonen's layer and s show an iteration:

$$KN_i^{(s)} = \min \left\| x^{(s)} - W_j^{(s)} \right\| \quad (1)$$

Update the weight vectors of all neurons by using (2), as follows:

$$w_j^{(s+1)} = w_j^{(s)} + \eta(s) h_{j.i}(s) \left(x - W_j^{(s)} \right) \quad (2)$$

where $\boldsymbol{h_{j.i}(s)}$ is the neighborhood function g(s) and $\boldsymbol{h_{j.i}(s)}$ are computed using (3) and (4), as follows:

$$\eta(S) = \eta_0 \exp(-s / \tau_2)). \quad s = 0.1.2. \ldots \quad (3)$$

$$h_{j.i}(s) = \exp\left(-\frac{d_{i.j}^2}{2\sigma^2(s)} \right) \; and \quad \sigma(s) = \sigma_0 \left(-\frac{s}{\tau_1} \right) \quad (4)$$

Repeat from all these steps until no noticeable change in the future map.

Definition 2. The input and output of the proposed protocol GA-based SOM (GASOM) including two algorithms are defined as:
Let *T* be the original database, a minimum support threshold ratio *MST*, and a set of sensitive items to be hidden $\boldsymbol{SX} = \{\boldsymbol{SX1}.\boldsymbol{SX2}.\cdots \boldsymbol{SXn}\}$. Let all of these parameters be input values, and $\boldsymbol{T*}$ be reduced database with least and hided sensitive information as the output of genetic algorithm and the input dataset for SOM clustering algorithm.

Definition 3. (hiding failure value) To evaluate the hiding failures of each processed transaction in the sanitization process, the α parameter is used to evaluate the hiding failures of each processed transaction in the sanitization process. Figure 2 shows the relation of the main dataset and its intersection with reduced datasets.

When a processed transaction contains a sensitive item, the Sum of the α value for the processed transaction $\boldsymbol{T_j}$ is calculated as:

$$\alpha^j(S_x) = \frac{MAx_{sx} - freq(S_x) \; + \; 1}{MAx_{sx} - \lceil |T| \times MST \rceil + 1} \quad (5)$$

where $\boldsymbol{MST}$ is defined as the percentage of the minimum support threshold, sensitive items $\boldsymbol{S_x}$ is from the set of sensitive items SX, MAX is the maximal count of the sensitive items in the set of sensitive items $\boldsymbol{SX}$, $|\boldsymbol{T}|$ is the number of transactions in the original database, and $\boldsymbol{freq(S_x)}$ is the occurrence frequency of the sensitive items $\boldsymbol{S_x}$. The overall α value for transaction j is calculate as:

$$\alpha^j = \frac{1}{\sum_{i=1}^{n} \alpha^j(s_i) + 1} \quad (6)$$

Definition 4. (fitness function) to find the optimal transactions including sensitive items to be deleted, the genetic algorithm needs a novel fitness function. Base on the[16] the fitness function calculates as:

$$\text{Fitness function} = W1\alpha + W2\beta + W3\gamma \quad (7)$$

where w1, w2, w3 are weighting parameters, defined by users. α value calculate by formula 2. β is another factor as the number of missing items and γ is the number of artificial items. Based on the power of SOM clustering in safely training phase and keeping complexity simple in distributed execution, we define W1=1 and ignore the other factors.

B. Privacy attacks

User's data is considered to be protected effectively when an adversary could not identify a particular user's data through linkages between a record owner to sensitive feature in the published data [25]. Thus, these linkage

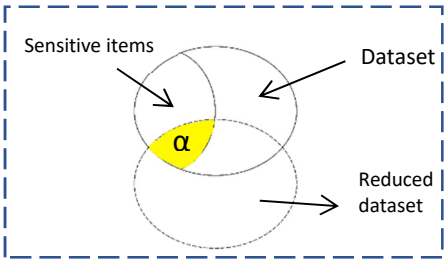

Figure 2. Hiding failure parameter.

attacks can be classified broadly into three types of attack models namely Record Linkage, Attribute Linkage and Table Linkage [28]. The proposed protocol in this paper aims to resist in front of Table Linkage sort of attacks.

Definition 5. (privacy attack) In all types of attacks, it is assumed that opponent knows the QIs (Quasi-identifiers) of the victim. If an opponent is talented to link a record holder to the published data table then such kind of privacy risk is known as Table linkage [28]. In this scenario, the attacker tries to govern the occurrence or nonappearance of the victim's record in the released table. To prevent table linkages privacy models such as δ-presence, ε-Differential privacy, (d, γ)-privacy and distributional privacy can be used. Our strategy in designing the method is trying to finding the optimal subsets to be deleted from the dataset, in order to preserve the data in front of such an attack.

C. Loss metric

Preventing sensitive item revelation may reduce the utility of data, as it involves data transformation [10]. One way is by measuring the difference between original data and transformed data, also called general purpose metrics, such as Generalization Cost, Normalized Average Equivalence Class Size, Normalized Certainty Penalty, and Information Loss Metric. For this paper, general purpose metrics apply to evaluate the information loss in this paper.

Definition 6. The information loss (IL) for a distributed GASOM partitioned and refined protocol is defined as $\sum_i len(T^*) - len(T)$ where T^* is the optimized and reduced dataset of T, and $len(T)$ denotes the number of transactions contained by T.

Definition 7. The information loss for a sensitive item i is defined as $SL(i) = len(i.T^*)$ where $len(i.T^*)$denotes the number of transactions that contain sensitive item i . Accordingly, the information loss for an anonymized integration dataset T is defined as:

$$IL(T) = \sum_{tran \epsilon T} IL(T) + \sum_{sen} IL(i) \quad (8)$$

where sen is a set of items to be hide in the transactions defined by administrator. The IL measures the information loss of item hiding process through the number of sensitive items. The larger the IL is, the less certain the results are relating to the true information of trajectories and transactions.

IV. PROPOSED SOLUTION

In this section, we represent our distributed SOM-based protocol (GASOM) to protect transaction data against sensitive item disclosure based on Genetic Algorithms. It consists of two phases. First, eliminating sensitive items disclosure through our Genetic Algorithm designed for this purpose. Second, securely SOM training datasets by applying a horizontally distributed map in order to reduce the number of communication among local units. We assume that adversaries hold limited knowledge of the victim, such as the work-class that the victim has previously filled in tax forms and also know the corresponding public items that the victim purchased.

A. Eliminate sensitive items disclosure

In this paper, a sensitive data hiding approach GASOM based on the genetic algorithms is thus proposed to find the appropriate transactions to be deleted for hiding sensitive items. The sensitive items to be hidden can be defined as $SX = \{SX1. SX2. \cdots SXn\}$. In the proposed GASOM for hiding the sensitive items through transaction deletion, the support count of a sensitive item must be below the minimum support threshold (MST), in which each transaction to be deleted must contain any of the sensitive items in SX. Base on this concept, we assume each transaction from T as a chromosome. A chromosome with m genes is thus designed that is compatible with the m attribute in the real dataset to be solved. Each gene represents a positive integer of transaction ID (TID) value as a possible transaction to be deleted.

The general steps of this algorithm are as follows:

Algorithm 1. GA dataset reducing

```
INPUT:    T, SX, MST
OUTPUT:  a reduced dataset T*
1.  Define the sensitive items as SX
2. for all the transaction Ti in T
    If Si ∈ Ti
        Project Ti from T to T '
    End if
End for
# initialize probability vector for each transaction Ti in T '
3. for all transaction Ti in T '
     Define p[i]=1
End for
4. Repeat
   # call GA function to compete for two transactions with
default crossover and mutation approach
     from T '
    1. randomly selecting TA and TB from T '
    2. compete for TA and TB by the fitness function
     For all transaction in T '
       Increase p[i] by 1/[ T '] for winner transactions
       decrease p[i] by 1/[ T '] for loser transactions
     End for
Until termination condition is not satisfied
#termination condition is reaching MST threshold
```

In competition process, each time two individuals are used for competition (in step 4). This approach can reduce the population size to speed up the evaluation process. As long as the termination condition is not satisfied, two other chromosomes are then generated again and compete on the probability of selected transactions in the winner chromosome. The final vector P as the output of this algorithm represents the probability of each transaction to delete from the main dataset.

B. Applying SOM clustering on a reduced dataset

The corporations, exclusively malicious ones, participating in distributed services attempt to derive information about each other's data. They can try to obtain useful information from interim results or final predictions. To protect data owners' confidentiality, our proposed scheme has to overcome privacy attacks. We use a two-step approach, where we cluster data off-line using SOM clustering (horizontally distributed) and utilize a genetic algorithm to hide sensitive items. We perform as many works as possible off-line to improve online efficiency. Also, with this technique, we reduce the number of communications in a network that known as on the most challenges in SOM. After determining local units online, clustering is estimated based on the users' data in local clusters.

The basic steps of our proposed protocol are as follows:

Algorithm 2. HDD SOM

```
INPUT:  main dataset T
OUTPUT:  - Index and reference vectors
(up to the request by central unit)
            - Local SOM clusters
1. Each local unit apply Algorithm 1 to get
 reduced safe database T*
2. Local unit i apply SOM algorithm on Ti*on the
 local data to obtain local clusters and also a reference
vector (to send to central unit)
3. In case of a request from the central unit, the local
index i
 send reference vector to the central server that will
represent the original data
4. The central unit remounts the dataset based on the
reference vector sent by local units and applies SOM
algorithm again to obtain a final output.
```

In step 1, algorithm 1 applies on each local dataset to get a reduced dataset T* which hide sensitive data. Applying genetic algorithm locally reduce the execution time which is a crucial factor especially in distributed networks. Then, in step 2, traditional clustering applies in each local dataset. These datasets are horizontally held same attributes. Thus, the algorithm applies to each subset, obtaining a reference vector and also locally trained clusters. This is the first time of applying SOM on the dataset and later in central unit another SOM training run to identifying the existing clusters. In case of a request from the central unit, in step 3, an index vector corresponding to the closest vector will be select and store in reference vector. This vector is very similar to the original object and in this way, data topology which is important will be kept. These vectors will be sent to the central unit and finally, in step 4, central unit combine these partial results and remount the dataset to obtain the main topology which is partially different with the original object but is very similar and more importantly protected. By applying another traditional SOM clustering method, the central unit could reach final output which all the clusters in all the unit exist and also sensitive data are hidden without losing accuracy.

V. EXPERIMENTS

In this part, we evaluated the data utility of the proposed protocol with real datasets. Also, privacy protection and information loss of the algorithm were tested. It should be noted that all the experiments accomplished on a local server and the idea of Algorithm 2 will be test in future works.

A. Experimental data

The test environment used for our initial Experiments was a VM/ Linux Ubuntu platform with 4 vCPU in Intel(R) Xeon (R) E5-2650 v4 processor and 4 GB memory. Two real database Adult [7], and Bank Marketing Dataset [23] is

used to evaluate the performance of the proposed algorithms in terms of the privacy and also the execution time as well as the accuracy of clustering operations. The details of these databases are shown in Table 1.

TABLE 1. EXPERIMENTAL DATASETS

Database	Transactions	Attributes	Area	Missing value
Adult	48842	14	Social	Yes
Bank Marketing	45211	17	Business	N/A

At first level, we weighed the execution times of proposed GA method that is a discussing topic in privacy issues. Genetic Algorithms are time-consuming and this factor significantly influences toward the goodness of the protocol. We tried to apply an optimal fitness function to promote the complexity. The execution times obtained using the proposed genetic algorithm are then compared under different minimum utility thresholds with a fixed rate of sensitive percentage 5% for the database is shown in Figures 3 and 4.

With increasing factor of MST Runtime is reduced, which naturally means reducing the level of data safety. In this experiment, the number of transactions is relatively equal, but the features and conditions used to define sensitive data are more complex in the Bank Marketing dataset. So, results amount to a significant increase in runtime in Figure 3.

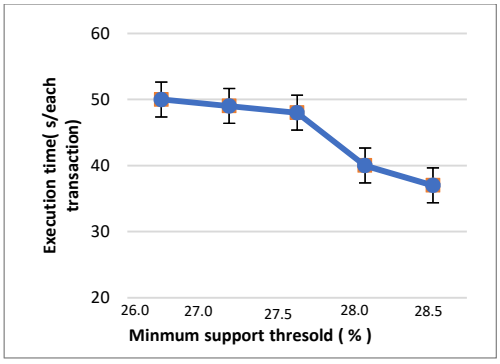

Figure 3. Execution time for adult data set with various minimum support thresholds.

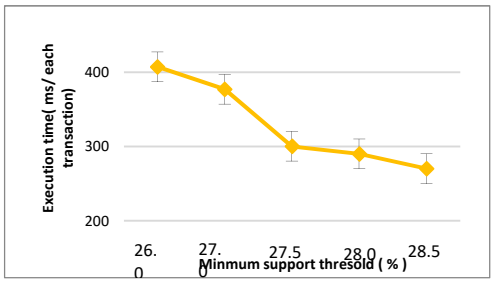

Figure 4. Execution time for adult data set with various minimum support thresholds.

Runtime is affected by the number of sensitive properties and validating conditions, but in general, the complexity of the algorithm is acceptable, especially as it runs locally on the server and does not add a bootloader to the system.

To evaluate the precision of the proposed algorithm, results are compared with those of traditional SOM clustering. Experiments were carried out using MATLAB 8.5 as well as SOM TOOLBOX. we set the radius of lattice to 3/2; and network topology to the hexagonal lattice, which is default topology in the MATLAB toolbox, and the optimum cluster number as three [17].

For the first experiments of our protocol, a test set of adult datasets has been used, which includes 1000 tuples with 14 attributes. Age and work-class are considered as sensitive attributes. A similar implementation of the Bank Marketing database was created with 3000 tuples and 17 properties. Sensitive features in this experiment were defined on three items of gender, age, and occupation, in order to verify the accuracy of the output of the genetic algorithm defined by the complexity of the sensitive items.

The neural network was implemented through MATLAB with the SOM toolbox and the attributes were represented in numeric format. The approach followed by firstly selecting the tuples matched with sensitive criteria, optimally hide those records with the genetic algorithm proposed. On our initial experiments, we cluster the data set only at the begging of the algorithm. In that case, the time needed for the hiding of the sensitive items in the dataset is depicted in Figures 3 and 4.

Afterward, the use of the neural network for training the partitioned dataset has been tested. Figures 5 and 6 show the U-Matrixes of clustering scheme of the two databases before and after data hiding task.

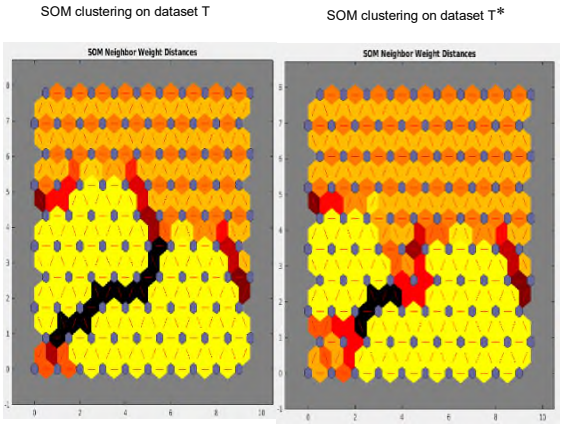

Figure 5. The U-Matrixes of traditional and proposed SOM clustering simulated on the Adult dataset.

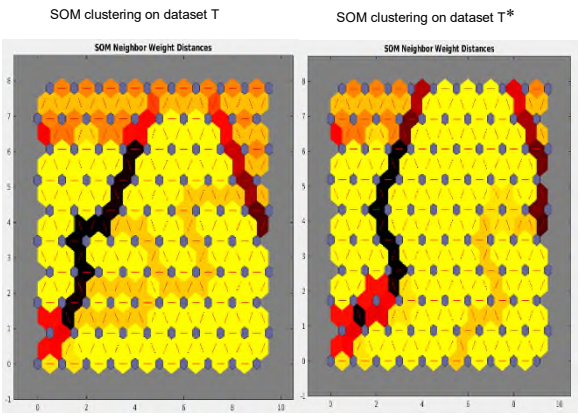

Figure 6. The U-Matrixes of traditional and proposed SOM clustering simulated on Bank Marketing dataset.

It is clearly shown that the difference between weighting distance in found clusters is not too much, however, it is affected by the size of the database. Figures 7 and 8 demonstrate the certainty penalty of the weight positions which is significantly increased by parameter weight. A new well-promising algorithm which takes into account the above assumption with less penalty in similarity of results is being studied and it is expected to be even more efficient.

To validate the proposed algorithms, besides the visual comparison of the trained map and U-Matrix between classic SOM and proposed approach, some other comparative criteria were used including average quantization error between data vectors and BMUs on the map and topographic error counting of errors obtained in the application of the algorithm over the datasets. In the next section, these accuracy measure results are presented.

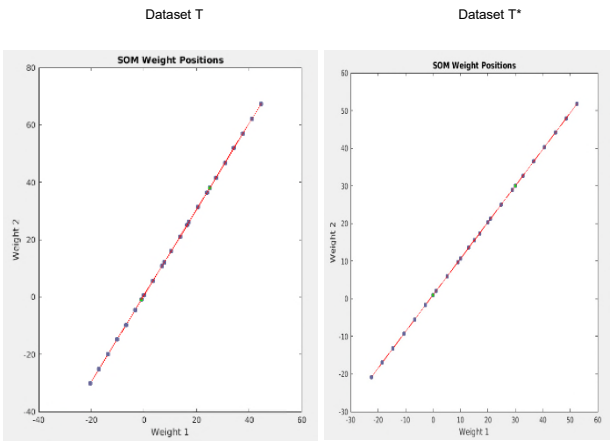

Figure 7. The Weight Position vectors of traditional and proposed SOM clustering simulated on Adult dataset.

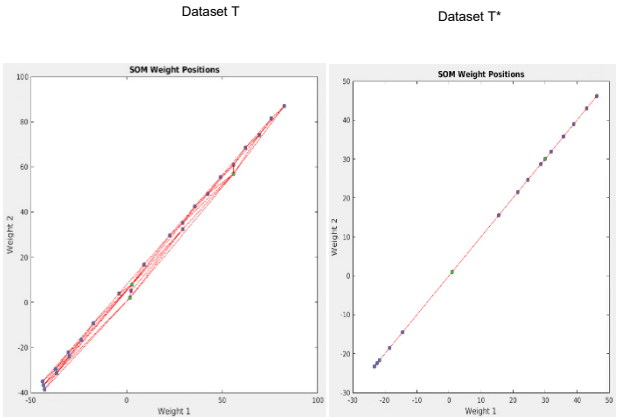

Figure 8. The Weight Position vectors of traditional and proposed SOM clustering simulated on Bank Marketing dataset.

B. *Analysis of information loss and privacy*

The well-known datasets Adult and Bank Marketing were used in a variety of privacy-preserving studies. Adult dataset It presents 48842 instances containing personal data with 14 attributes. We defined the sensitive items in this trial experiment as the age of the people under 30 with work-class 'Private'. Firstly, the dataset was analyzed initially using proposed GA method to extract a probability vector indicating the rate of failing in hiding sensitive items. The user then could decide about the rate of deletion from the dataset, which defined as MST in the proposed algorithm. Dataset was then horizontally partitioned, each containing all the attributes. In this phase, we implement experiment on a local server with 1000 tuples, both plan, and hexagonal lattice. For Bank Marketing dataset with 45211 instances and 17 attributes a similar criterion defined for age, job and marital attribute to test the influence of complexity of sensitive criteria on final results. 3000 tuples used for this experiment with the equal condition on the local server.

Maps size was defined by SOM Toolbox, based on data distribution in input space. In our experiments, maps were randomly initialized and batch SOM was used. We defined the constant fixed for both classic and proposed SOM as sigma initial=2 and sigma final=1 and trainlen defined to 1 epochs. Table 2 summarizes the clustering quality measures.

TABLE 2. COMPARING THE ACCURACY OF CLASSIC SOM AND GASOM

Dataset	Method	QE	TE
Adult	Classic SOM	0.0798	0.2290
	GASOM	0.0943	0.1435
Bank Marketing	Classic SOM	0.193	0.042
	GASOM	0.135	0.088

 ISBN: 978-1-61208-661-3

QE represents the average quantization error and TE represents the topological error. These results prove the usefulness of GASOM in keeping the clustering quality beside the improvement of protection. Also, to prove the protection level of the dataset, the difference between the number of sensitive items before and after hiding task in genetic algorithm calculated as follows:

$$\frac{Senfreq|T^*|}{Senfreq|T|} \qquad (9)$$

where T* is the reduced dataset and T is the main dataset before hiding task. The result of (9) is always near to zero which proves the goodness of the proposed method of hiding the sensitive items. Although the results of experiments prove the usefulness of proposed protocol, try to refine the methods in order to keep the accuracy of clustering and the execution time sounds imperative. Figure 9 represents a comparison between the relation of hiding factor and Minimum Support Threshold (MST), which demonstrates privacy protection decrease with increasing the MST.

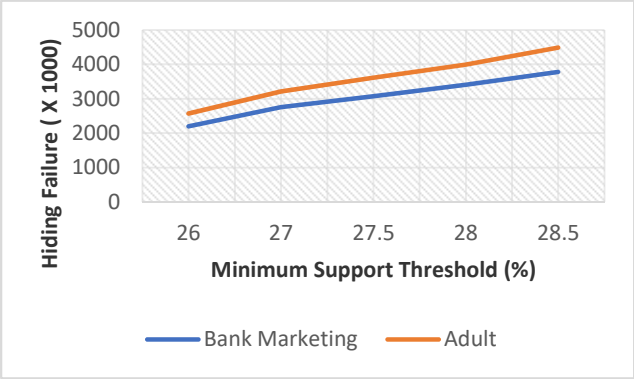

Figure 9. The relation between two factors of MST and Hiding Failure to check the information loss in front of privacy level.

Therefore, the boundary of information loss and privacy level are opposite and should be calculated and selected according to the requirements and conditions of the current database. Linkage attack by applying this protocol is completely deniable. The quasi-identifier for our method is defined as the whole subset of attributes that can uniquely identify a record. So, an attacker cannot find the complete quasi-identifier which we have already change with our GA method. However, this method is just a try to check the goodness of proposed GA methods in finding the sensitive items. Although the accuracy of GASOM is relatively acceptable, some other techniques like fuzzifying the optimal subset found by fitness function seem to be useful to implement to avoid of eliminating those transactions from the database.

VI. CONCLUSION AND FUTURE WORK

The crucial need for smarter approaches to analyze data distributed among several sites is obvious. This issue beside the increasing importance of privacy-preserving becomes much complicated. In this paper, a hiding sensitive technique for SOM clustering approach for partitioned data is thus proposed to hide the sensitive items using Genetic Algorithms. To determine the goodness of a transaction, a flexible fitness function with adjustable weights is also designed to consider the general side effect of hiding failure. Our offerings in this paper can be summarized as follows: First, a sanitization process to find the sensitive items from the main dataset will be done in order to shape a probability vector indicating the chance of each transaction to be deleted from dataset to hide the sensitive data. Second, the reduced dataset will be trained by local SOM to shape the topological map and finally the central unit merge the results of the local unit based on the reference vectors sent by local units to integrate the final clusters. Experiments are conducted to show that the proposed GASOM protocol beats better than classic algorithms considering the criteria of side effects but the execution time.

Final results demonstrate that the proposed protocol obtained similar results to those of classic clustering algorithms. The results of privacy protection prove the power of proposed GA methods. However, it is still necessary to find a more effective solution to keep the privacy with less information loss. Further research will include applying this protocol on distributed units and also trying a different soft based method like swarm intelligence to compare with the results of GA method form the privacy point of view. In this version of protocol, sensitive items defined by users which is context-based. In future works, we want to consider more details about these sensitive items regarding ownership, personal and semi-context sensitive data. Also, it makes a lot of sense to propose some way to change just a small portion of the database instead of deleting those records to reach all the goals defined in this paper at the same time. In this way, we want to integrate our protocol to some other machine learning techniques, such as fuzzy sets to refine the triple goals of privacy, accuracy, and speed. It should be noted that all the experiments accomplished on a local server and the idea of Algorithm 2 will be test in future works.

ACKNOWLEDGMENTS

This work was partially supported by SBA Research Institute, Vienna, Austria.

REFERENCES

[1] R.Agrawal and R.Srikant," Privacy-preserving data mining" ACM Sigmod Record, ACM, pp. 439-450, 2000.

[2] F. Amiri and G. Quirchmayr, "A comparative study on innovative approaches for privacy-preserving in knowledge discovery", ICIME, ACM, 2017.

 ISBN: 978-1-61208-661-3

[3] F. Belanger and R.E. Crossler,"Privacy in the digital age: a review of information privacy research in information systems", MIS quarterly vol. 35, no. 4, pp.1017-1042, 2011.

[4] A. Bilge and H. Polat, "A comparison of clustering-based privacy-preserving collaborative filtering schemes", Applied Soft Computing vol. 13, no. 5, pp. 2478-2489, 2013.

[5] C. Clifton, et al. , "Privacy-preserving data integration and sharing", Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery ACM, pp.19-26, 2004.

[6] M.N. Dehkordi, K. Badie, and A.K. Zadeh, "A novel method for privacy preserving in association rule mining based on genetic algorithms", Journal of software vol. 4, no. 6, pp. 555-562, 2009.

[7] D.K.T. Dheeru, "UCI Machine Learning Repository. University of California", Irvine, School of Information and Computer Sciences, 2017.

[8] W. Du, and M.J. Atallah, "Privacy-preserving cooperative statistical analysis", Computer Security Applications Conference, ACSAC , Proceedings 17th Annual IEEE, pp. 102-110, 2001.

[9] G. Gan , C. Ma, and J. Wu, "Data clustering: theory, algorithms, and applications", Siam , 2007.

[10] A. Gkoulalas-divanis, G. Loukides, and J. SUN, "Publishing data from electronic health records while preserving privacy: A survey of algorithms", Journal of biomedical informatics vol. 50, pp. 4-19, 2014.

[11] A. Gkoulalas-divanis and V.S Verykios, "An overview of privacy preserving data mining", Crossroads vol. 15, no. 4, p. 6, 2009.

[12] S. Han and W.K. Ng, "Privacy-preserving genetic algorithms for rule discovery", International conference on data warehousing and knowledge discovery Springer, pp. 407-417, 2007.

[13] S. Han and W.K. Ng, "Privacy-preserving self-organizing map", In DaWaK Springer, pp. 428-437, 2007.

[14] S. Haykin, "Neural Networks: A Comprehensive Foundation", Vol. 2. Segundo, Prentice Hall. España, 1999 .

[15] J.H. Holland, "Adaptation in natural and artificial systems. An introductory analysis with application to biology, control, and artificial intelligence", Ann Arbor, MI: University of Michigan Press, pp. 439-444, 1975.

[16] T.P. Hong, K.T. Yang, C.W. Lin and S.L. Wang, "Evolutionary privacy-preserving data mining", World Automation Congress (WAC), IEEE, pp. 1-7, 2010.

[17] C. Kaleli and H. Polat, "Privacy-preserving SOM-based recommendations on horizontally distributed data", Knowledge-Based Systems vol. 33, pp. 124-135, 2012.

[18] C. Kaleli and H. Polat, "SOM-based recommendations with privacy on multi-party vertically distributed data", Journal of the Operational Research Society vol. 63, no. 6, pp. 826-838, 2012.

[19] M. Kantarcoglu, J. Vaidya and C. Clifton, "Privacy preserving naive bayes classifier for horizontally partitioned data", IEEE ICDM workshop on privacy preserving data mining, pp. 3-9, 2003.

[20] C.W. Lin, B. Zhang, K.T. Yang, and T.P. Hong, "Efficiently hiding sensitive itemsets with transaction deletion based on genetic algorithms", The Scientific World Journal, 2014.

[21] Y. Lindell and B. Pinkas, "Privacy preserving data mining", Annual International Cryptology Conference Springer, pp. 36-54, 2000.

[22] Y. Lindell and B. Pinkas "Privacy preserving data mining. Journal of cryptology vol. 15, no. 3, 2002.

[23] S. Moro, P. Cortez, and P. Rita, "A Data-Driven Approach to Predict the Success of Bank Telemarketing", Decision Support Systems, Elsevier, vol. 62, pp. 22-31, 2014.

[24] T.H. Roh, K.J. Oh, and I. Han, "The collaborative filtering recommendation based on SOM cluster-indexing CBR", Expert systems with applications vol. 25, no. 3, pp. 413-423, 2003.

[25] P.N. Tan, M. Steinbach, and V. Kumar, "Association analysis: basic concepts and algorithms", Introduction to Data mining, pp. 327-414, 2005.

[26] E.C. Turner, and S. Dasgupta, " Privacy And Security In E-Business", Taylor & Francis, 2003.

[27] J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data", Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining ACM, pp. 639-644, 2002.

[28] K. Wang, R. Chen, B. Fung, and P. Yu, " Privacy-preserving data publishing: A survey on recent developments", ACM Computing Surveys, 2010.

[29] X.-Z. Wang, Q. He, D.-G. Chen, and D. Yeung, "A genetic algorithm for solving the inverse problem of support vector machines", Neurocomputing vol. 8, pp. 225-238, 2005.

[30] R. Wright and Z. Yang, "Privacy-preserving Bayesian network structure computation on distributed heterogeneous data",Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining ACM, pp. 713-718, 2004.

[31] Y. Yang, W. Tan, T. Li, and D. Ruan , "Consensus clustering based on constrained self-organizing map and improved Cop-Kmeans ensemble in intelligent decision support systems". Knowledge-Based Systems vol. 32, pp. 101-115, 2012.

 ISBN: 978-1-61208-661-3