A Machine Learning Approach for Privacy-preservation in E-business Applications

Fatemeh Amiri[1,2], Gerald Quirchmayr[1,2], Peter Kieseberg[3]
[1]*University of Vienna, Vienna, Department of Computer Science, Austria*
[2]*SBA Research Institue, Vienna, Austria*
[3]*St. Poelten University of Applied Sciences, St. Poelten, Austria*

Keywords: Privacy-preserving, E-business, Big Data, Data Mining, Machine Learning.

Abstract: This paper aims at identifying and presenting useful solutions to close the privacy gaps in some definite data mining tasks with three primary goals. The overarching aim is to keep efficiency and accuracy of data mining tasks that handle the operations while trying to improve privacy. Specifically, we demonstrate that a machine learning methodology is an appropriate choice to preserve privacy in big data. As core contribution we propose a model consisting of several representative efficient methods for privacy-preserving computations that can be used to support data mining. The planned outcomes and contributions of this paper will be a set of improved methods for privacy-preserving soft-computing based clustering in distributed environments for e-business applications. The proposed model demonstrates that soft computing methods can lead to novel results not only to promote the privacy protection, but also for retaining performance and accuracy of regular operations, especially in online business applications.

1 INTRODUCTION

Privacy-preservation in e-business that holds big data is an essential topic of discussion and, requires addressing several aspects (Verykios et al., 2004a). Applying a privacy improvement task may bring side effects to other running tasks like data mining applications and administration programs.

Managing big data is itself an extensive field and is made possible by involving a knowledge discovery concept which is the process of extracting useful hidden information from raw data (Fayyad et al., 1996, Narwaria and Arya, 2016). Data mining, as the core part of the knowledge discovery approach, enables the practitioner to extract useful knowledge from the data to better understand and serve their customers to gain competitive advantages (Chen, 2006).

From the privacy-preserving point of view, confidential and sensitive data should be protected at the same time when running the data mining process (Jahan et al., 2014). Therefore, the long-term aims in this paper are specific attainments in privacy-preserving data mining (PPDM) in the e-business environment that can be achieved by a certain number of steps.

In the first studies on PPDM, the primary objective of privacy-preserving algorithms was only protecting sensitive data at the time of publishing (Westin, 1999). Lots of approaches are adopted to solve this issue. Some of the proposed algorithms try to hide the sensitive data and some others are designed to modify and add extra noise in a manner to protect raw data (Samet and Miri, 2012). Also, data mining tasks are time-consuming.

So, if privacy preserving approaches reduce the performance significantly, the final combined algorithms will most likely not be useful. This issue is even more problematic in distributed environments than centralized models.

Hence, a strong privacy-preserving method cannot present real practical achievements without satisfying complexity constraints.

As Figure 1 shows, privacy of data should be protected as vigorously as possible while at the same time assuring an acceptable level of performance and accuracy.

However, the accuracy of data mining methods should be fixed and overheads must be as low as possible. Thus, privacy-preservation is not the only issue anymore. Accuracy and the performance of the data mining tasks have the same necessity. In this

Amiri, F., Quirchmayr, G. and Kieseberg, P.
A Machine Learning Approach for Privacy-preservation in E-business Applications.
DOI: 10.5220/0006826304430452
In *Proceedings of the 15th International Joint Conference on e-Business and Telecommunications (ICETE 2018) - Volume 2: SECRYPT*, pages 443-452
ISBN: 978-989-758-319-3

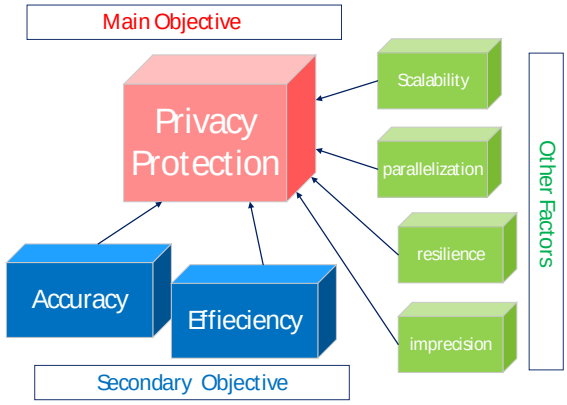

Figure 1: Relationship between PAE keywords (Privacy-Accuary-Efficiency).

paper, equal attention will be paid to these three key parameters - PAE (Privacy, Accuracy, Efficiency). It should be noted that in e-business applications, some other factors like those mentioned in Figure 1 come into action and need to be taken into account by design approaches. Lots of models and algorithms were proposed in order to close privacy gaps. However, privacy is not the only point, and other factors need to be taken into account. To the best of our knowledge, a method with such a description that satisfies all PAE factors has not yet been proposed.

Solving this issue needs a new approach to reach significant improvements. A considerable number of scholars using conventional methods indicate the requirement for a new technique.

Soft computing as a collection of various methodologies could provide flexible knowledge processing facilities to handle ambiguous real-world problems in a synergistically manner (Zadeh, 1994). The idea of using the word *soft* indicates the focus of using methods that are not conventional or *hard*. However, in many publications soft computing is also used as synonym for machine learning too. Soft computing can focus on achieving robustness, tractability, and at the same time provide less expensive solutions regarding time and space complexity (Maimon and Rokach, 2007). As a result, the challenge of developing acceptable solutions for privacy-preserving data mining are well-addressed by soft computing methods (Mitra et al., 2002) (Malik et al., 2012).

Recently, privacy-preserving using soft computing methods was discussed and some novel approaches proposed by the authors of this paper (Amiri and Quirchmayr, 2017). These methods yield better results in all PAE aspects defined in Figure 1.

This paper will build on our previous work by further exploring and describing privacy gaps in PPDM and then propose some methods to close privacy gaps in selected data mining tasks.

Regarding the problem of closing privacy gaps and harmonizing some other factors, it is expected to make a number of improvements that contribute to advancing the state of the art in privacy-preservation in data mining in distributed environments:

- *A privacy-preserving model for soft computing-based clustering.* This model consists of several methods for improving the PAE factors. Each method uses a different soft computing approach to find a reasonable level of performance for each factor.

- *The proposed algorithms bring a better level of privacy.* With implementing soft computing methods, we expect to reach better results than traditional methods of PPDM. The definition of a privacy breach in this paper is: Disclosing not only the sensitive information of individuals without their consent, but also releasing information in such a way that the individuals may be linked with totally unrelated information even with a very small probability.

- *The accuracy of data mining task results*. The proposed algorithms shall not sacrifice accuracy of data mining tasks. These methods protect the privacy of data before the mining task and, keep the utility and do not alter raw data.

- *The efficiency of overall operations*. As soft computing methods are natively fast, we hope that the computation and communication cost would not significantly increase through the addition of privacy-preserving components. Figure 2 shows the overall plan.

As Figure 2 shows, SOM(Self Organizing Map) clustering data mining tasks is used for data mining purposes. This clustring could be horizontally or vertically in distribution. The primary objective is accuracy and efficiency of the final output of data mining tasks with a high privacy level. To achieve this purpose, we use a combination of soft computing methods – depicted in the left box of the picture. As discussed before, none of the unique techniques could satisfy all PAE factors. So a hybrid model will be used in the proposed algorithms. We apply a Neural Network based clustering like SOM and try to promote its PAE factors. According to the current state of research, SOM clustering suffers from lack of privacy, as well as a lack of some communication aspects in the distributed interaction between clusters.

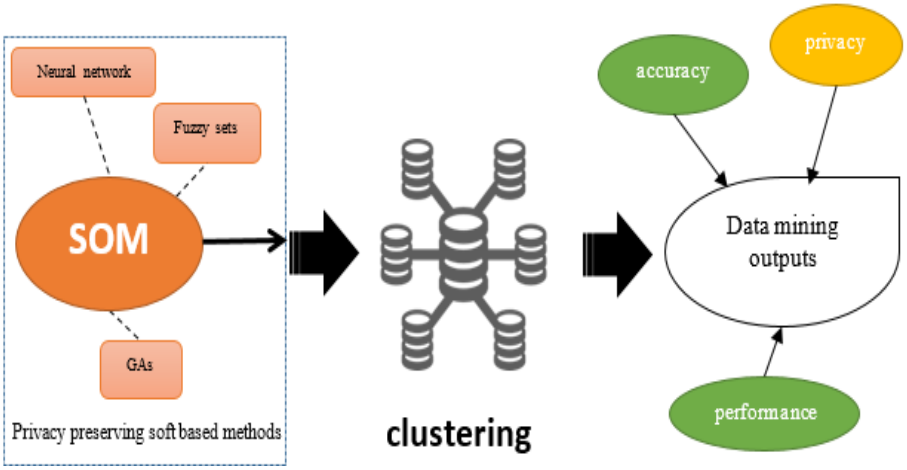

Figure 2: The overall plan of proposed approach in this paper including the soft methods in the left box, selected data mining task and expected results.

We try to close these gaps by implementing soft computing methods like fuzzy sets and Gas(Genetic Algorithms) in our proposed plans.

The core of this paper is organized as follows: Related work on privacy-preservation in data mining and soft computing is reviewed in Section 2. Section 3 discusses the definition of the problem with the parameters involved. The applied methodology to design the model and related methods is also covered in this section. Model schemes for implementing the new privacy-preserving mechanisms are described in Section 4. In Section 5, we discuss the application of our mechanisms to protect privacy when using machine learning and also discuss expected results. Finally, in Section 6, we summarize the major conclusions of our research and outline future research directions.

2 RELATED WORKS

Privacy-preserving data mining was introduced by (Agrawal and Srikant, 2000). Related studies present several useful methods with different views on the PAE factors. Some scholars focus only on one kind of database, like (Agrawal and Srikant, 2000) that concentrate on statistical databases and promotions brought by reconstructing the perturbation before constructing the mining process. However, in e-business applications, data transitions usually happen in distributed environments. (Wu et al., 2007) proposed a simplified taxonomy that divides the data mining process into two centralized and distributed environments. Despite its importance, the methods presented are much less suited for distributed environments than central servers. Another important issue lies in the selection of the method for achieving privacy (e.g., k-anonymity through generalization), as this selection directly influences the PAE factors. The capability of candidate methods to improve privacy should not be the only factor in this selection. Some other factors like accuracy and complexity should be taken into account as well. However, this point is rarely investigated by scholars. (Aggarwal and Philip, 2008) used a general classification to explore current research, including randomization methods, the k-anonymity model, and distributed privacy preservation. A peer contribution of their study (that has not been investigated before) is paying particular attention to the side effects of data mining approaches that may influence the privacy of individuals. They explore this side effect as Downgrading Application Effectiveness. Although they propose an effectiveness factor for the first time, the analysis of the methods is not extensive.

A comprehensive study on popular methods in PPDM accomplished in (Amiri and Quirchmayr, 2017). They defined a broader classification in both categories (soft and traditional ways of privacy-preserving) to reach an innovative idea about research gaps and the state of the art of the topic. K-Anonymity, Perturbation, Cryptographic & Distributed methods, Association rule-based PPDM, Classification based PPDM, Fuzzy Logic, Neural Networks, Genetic Algorithms and Rough sets are the main categories.

Analysis of state of the art shows that, among popular methods used for privacy-preserving, soft computing methods may yield better results. In this paper four best popular soft methods selected for designing the algorithms, including Neural Networks,

Fuzzy sets, Genetic Algorithms and Rough Sets. This selection is based on their unique characteristics and demonstrated the capabilities of these methods.

Soft clustering is one of the most widely used approaches in data mining with real-life applications.

Mingoti and Lima compared SOM and other clustering algorithms like fuzzy c-means, k-means according to their clustering capabilities on different data structures. Their results show that there are situations in which SOM performs well, even though it is not the best (Mingoti and Lima, 2006). (Roh et al., 2003) apply SOM clustering to their e-business application in which empirical results demonstrate that the SOM-based recommender system applied in e-business designs supplies higher-quality referrals than other comparative models. Hence, we also use SOM clustering to enhance the privacy attribute while obtaining the dual secondary goal of better accuracy and performance.

3 PROBLEM DEFINITION

The privacy-preservation problem in e-business is still an open problem and needs more secure methods to improve the protection level of sensitive data. However, improving privacy may account for sacrificing the accuracy of data mining operations and the efficiency of the overall process. So, proposed methods should consider all these factors.

3.1 Problem Statement

The problem of privacy-preserving in data mining has three viewpoints and, requires a solution which keeps a balance in all the aspect defined in Figure 1:

- Improving privacy
- Keeping efficiency
- No information loss

1) Type of Environment

The focus of this paper lies on distributed environments, since not only privacy and optimization issues in this type of environments seem to be more crucial, but also fewer studies have been done on this definition. Moreover, with the explosion of data, and requiring better performance, resorting to distributed computing and finding gaps and challenges of new issues in this environment seems to be critical (Saxena and Pushkar, 2017).

2) Type of Methods for Closing Privacy Gaps

Verykios and his partners demonstrated that privacy-preserving problem might be an NP-hard problem (Verykios et al., 2004b).The particular capability of soft computing methods is their power of deriving knowledge, as well as extracting patterns and trends from complex data, which are otherwise hidden in many applications (Malik et al., 2012).

On the other hand, the comparison between two factors of performance in Figure 3 and privacy level demonstrates that none of the classic methods even soft methods are satisfactory.

So, we will define our solutions with hybrid soft computing methods. We will identify and discuss new secure mining computations that enable preserving privacy.

3) Selected Data Mining Task to Improve Privacy

As different data mining task may have different feature and behaviour, to limit the scope of the problem, soft clustering is selected to implement the proposed privacy-preserving algorithms base on the gaps that are have already found.

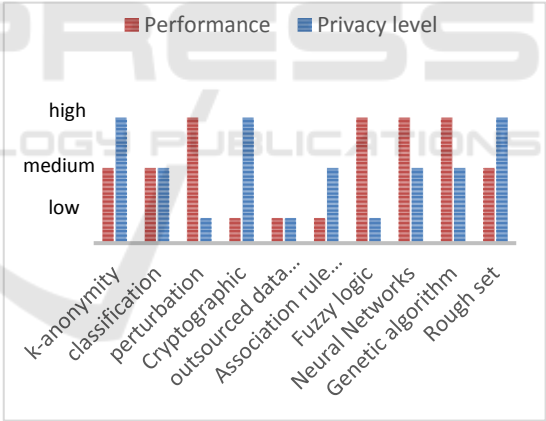

Figure 3: A comparison between two factors of privacy and performance, in all methods, to show this point that none of the soft nor hard methods are completely satisfying the PAE factors.

We will augment a SOM clustering from different aspects according to the PAE factors. The reason of appointing SOM is based on the results of evaluating clustering capability of SOM state where it was shown that the SOM network is superior to hierarchical clustering algorithms in e-business applications.

In this way, different proposed methods improve SOM using a selection of different soft computing methods. Some of our proposed methods are entirely new and make use of procedures that have not been utilized in this context before. Some others build on already existing techniques

and improve them. In section 6, we explain the details of these methods.

3.2 Methodology

The "design science research methodology" is used in this paper which is typically applied to categories of artifacts including algorithms, human/computer interfaces, design methodologies (including process models) and languages.

It is fundamentally a problem-solving paradigm. Design science, as the other side of the IS research cycle, creates and evaluates IT artifacts intended to solve identified organizational problems. A mathematical basis for design allows many types of quantitative evaluations of an IT artifact, including optimization proofs, analytical simulation, and quantitative comparisons with alternative designs(Von Alan et al., 2004).

We follow this methodology with these steps:

1) Identification of the criteria for the solution to meet
2) Model development
3) Implementation
4) Test
5) Evaluation of results
6) Comparison with existing literature definitions

The first two parts discussed in this paper and following parts will be complete in the future.

4 A PRIVACY -PRESERVING SOM BASED MODEL - SPE

In this part, the fundamental concepts are discussed which mainly work based on the Neural Networks as a popular soft computing method and then, the SPE model designed for this problem explained.

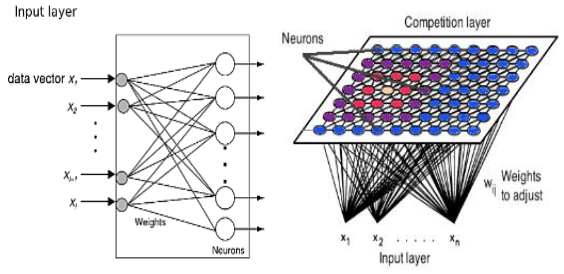

Figure 4: SOM network layers (Kohonen and Maps, 1995).

4.1 Prerequisite Definitions

Neural networks applied to solve the problems where the solutions are either too complicated or impossible to find a solution for a problem like privacy-preserving that is uncertain and imprecise (Zadeh, 1994). Popular Neural Network models for data mining purposes that seem suitable for the solution of the mentioned problem include Hopfield networks, Multilayer feedforward networks, and Kohonen's map.

The soft clustering method selected in this paper to solve the privacy gaps is Kohonen's map known as SOM. We improve the PAE factors using soft computing methods like fuzzy and GAs.

Kohonen's self-organizing maps (SOM) is one of the essential Neural Networks models for data mining tasks, especially for data clustering and dimension reduction (Kohonen, 1982). SOM can learn from complex, multidimensional data and transforms them into a topological map of much fewer dimensions, typically one or two (Zhang, 2009). These characteristics make SOM a suitable scheme to apply for e-business applications, especially in recommender systems.

A typical SOM network has two layers of nodes, an input layer and an output layer called Kohonen's layer. Each node in the input layer fully is connected to nodes in the two-dimensional output layer. Figure 4 shows a simplified SOM network with several input nodes in the input layer and a two-dimensional output layer. SOM executes a series of iterations that each consists of two phases: the competition and cooperation phases (Kohonen and Maps, 1995). At iteration t, during the competition phase, for each input data X(t) = [Xi(t), X2(t), …, Xd(t)] from the data set, the Euclidean distance between X(t) and each neuron's
weight vector Wj(t)= [Wj,1(t),Wj,2(t), … , Wj,d(t)] ($1 \leq j \leq K$, where K is the total number of neurons in the grid) is computed to determine the neuron closest to X(t) as follows:

$$\|X(t) - W_j(t)\| = \sqrt{\sum_{i\equiv 1}^{d} (X_i(t) - W_{j.i}(t))^2} \quad (1)$$

The neuron c with weight vector Wc(t) that has the minimum distance to the input data X(t) is called the winner neuron:

$$c = arg_j \min \|X(t) - W_j(t)\| \quad (2)$$

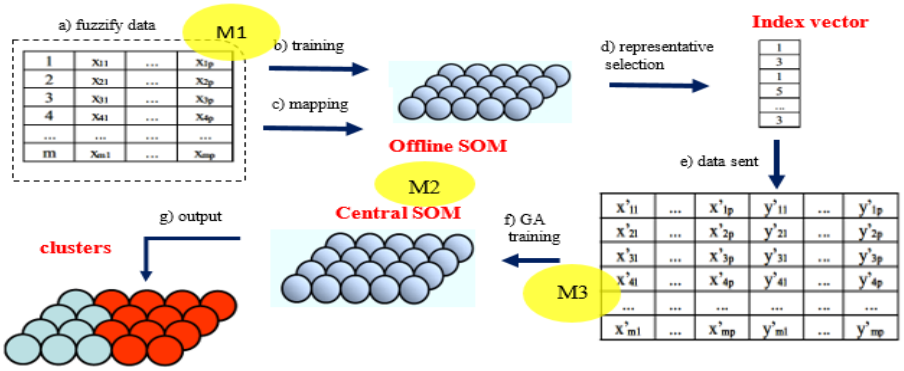

Figure 5: First plot of SPE model as a privacy-preserving SOM clustering.

During the cooperation phase, the weight vectors of the winner neuron and the neurons in the neighborhood G(rc) of the winner neuron in the SOM grid are shared towards the input data, where rj is the physical position in the grid of the neuron j. The magnitude of the change declines with time and is smaller for neurons that are physically far away from the winner neuron. The function for change at iteration t can be defined as Z(rj , rc, t)

The update expression for the winner neuron c and the neurons in neighborhood G(rc) of the winner neuron is shown as follows:

$$\forall j \epsilon\, G(r_c).W_j(t+1) \;=\; W_j(t) + Z(r_j.r_c.t)\big[X(t) - W_j(t)\big] \quad (3)$$

SOM suffers from some privacy gaps in the communication parts:

- securely discover the winner neuron from data privately held by two parties
- safely update weight vectors of neurons
- firmly determine the termination status of SOM (Han and Ng, 2007)

To overcome these challenge, some add-ons to SOM designed to identify and address the privacy gaps without information loss or significant overhead. These hybrid techniques are based on soft methods and will help to fix the breaches. The general model(SPE) is shown in Figure 5. The yellow circles in different parts of the picture indicate the proposed methods which try to solve the issue. These methods declared in following parts which try to improve privacy with close a privacy breach that SOM suffer from.

4.2 SPE Model

Privacy-preservation with attention to accuracy and efficiency is an NP-hard problem (Bonizzoni et al., 2009, Blocki and Williams, 2010). So, finding a unique approach which could satisfy all these factors is impossible. In this way, the SPE model proposed to improve the privacy level of sensitive data regarding other angles too. SPE consists of some independent methods, each one tries to augment PAE factors. These methods proposed to implement in different parts of SOM transactions proposed according to the privacy bugs founds. In this manner, each method could focus to fix a privacy breach without any distortion in common operations. In SPE some independent methods are defined. Each method tries to protect sensitive data without jeopardizing in the system.

Method 1: Protecting sensitive data with fuzzifying before training.

The idea of using fuzzy sets in the privacy-preservation problem is trying to apply fuzzy functions to preserve the private information of individuals while revealing the details of the aggregation in public. Results of analyzing related works proved that fuzzy sets work for privacy preservation and only introduce low additional costs in the different cases that were explored above. Due to its low cost, it may be applied in both types of environments, especially in distributed servers. However, using fuzzy sets as a single method of privacy-preservation is not sufficient to cover new attack(Li et al., 2017). Due to the progression of attacks in prediction of the results of applying fuzzy functions, using Fuzzy transforming function cannot protect the sensitive data strongly. So, with using fuzzy sets alone, there is no guarantee to reach the best case of privacy-preservation. However, as a

complementary approach, fuzzy methods are robust tools.

Method 1. Fuzzify data before training

INPUT:
Dataset S consists of sensitive attribute data in m rows and n columns.
OUTPUT:
Distorted Datasets that each S' consist of m rows and n columns.

```
//Fuzzy membership functions such as Triangular-shaped,
 S-shaped, and Gaussian used for transformation.
Begin
{
      1.Suppress the identifier attributes
      2.For each membership function (Triangular, Z-shaped, S-
shaped, Gaussian)
            For each sensitive element in S do
            {
              Convert the element using selected
      fuzzy membership function.
            }

    3. Release the all distorted datasets for training by SOM
function
}
End
```

Method 1 transforms the data set which consists of some sensitive attributes. As discussed before, this method is not enough, but in combining with other methods in different parts of the model sounds useful to implement. However, the idea of using fuzzy set in this method returns to one of the SOM breaches that mentioned in the previous section. This breach is the high number of communication in the network. Fuzzy function in this method transform the sensitive data before training and prevent the bad guys to reach the sensitive data. Protecting in this point of the model (before training) is crucial at this stage data is raw and with a high rate of probability bad guy can take advantage of reaching them because of the high number of distributed interactions. It makes a lot of sense if in this point of model raw data transform to a shape in which bad guy cannot recognize them anymore.

Here, the sensitive data that defined by the user is the input of method and, during an iteration function try to fuzzify each item using a membership function like Triangular, Z-shaped, S-shaped, Gaussian.

The output of the method would be a distorted dataset that is ready for training by SOM function. In this way, the worry of the high number of interaction in the network reduced and, this because of applying a method that distorted the data. As stated before, this method for protection against privacy attacks is not sufficient, but by applying other methods like method 3 and the idea of method 2, protection will be even stronger.

Method 2: Implementing Offline SOM clustering.

The high amount of communication in SOM is a considerable risk factor which exposes it to a variety of security and privacy attacks. Decreasing the amount of communication may diminish these risks.

In our proposed method which is based on (Gorgonio and Costa, 2008), we apply extra add-ons in order to decrease the risk of privacy attacks, and also as a way to try to keep the performance of the processes in the system.

The general plan is shown in figure 5 including different offline SOM parties which train separately and are then integrated with least steps.

Method 2. Offline SOM clustering

INPUT :
Raw Data to be trained and mapped including sensitive information
OUTPUT:
Clusters consist of protected information

```
//Network  Administrator decide about applying M1,M2,... based
on the level protection that is required
Begin
{
1. Repeat
        {
         each offline SOM train and map the data independently
         //Arbitrary applying Method 1
        add a new record to index vector for this offline SOM
        }
        Until the termination criterion is not satisfied.
         //defined by Admin with number of offline SOM servers

2. integrating trained and mapped data into central SOM
3. arbitrary applying Method 2
4. training and mapping by central SOM
}
End
```

This method defines the overall process of the SPE model. Its aim is reducing the number of communications in a distributed network. So, administrator divides the data in different local point based on the e-business strategy. Each point acts independently and, run the protecting methods like method 1 to enrich itself against privacy breach. Next, it runs SOM clustering and gets the clusters as the results of the process. In this step, the local administrator could decide about applying other protecting methods like method 3 to be even more protected. In this way, each local point has a unique result which is the local clustering. Finally, the main administrator asks local points to integrate trained and mapped data into a central SOM. This integration by itself runs on a secure channel and protected against possible attacks.

Method 3: Using Data Hiding with GA to Improve Privacy Protection.

In a privacy-preserving concept, GAs (Genetic Algorithms) can be applied as a complementary tool in optimizing the results of primary algorithms. Most of the approaches maximize the results of other conventional methods of PPDM. According to our research, when optimization is compulsory, soft computing techniques, especially GAs, could yield better results.

Moreover, in GAs, a chromosome resembles a possible, feasible solution. Since the goal here is to hide at most m appropriate transactions from a database such that the fitness value can be optimal, a chromosome with m genes is used, with each gene representing a possible operation to hide.

Consideration of this method in our desired hybrid model guarantees the accomplishment of all PAE factors.

Method 3. Hiding sensitive data using GAs

INPUT:
A dataset S, consist of a set of sensitive itemsets, and the maximum number m of transactions to be hidden, and a population size n.
OUTPUT:
An appropriate set of transactions to be hidden S'
//suitable fitness function to apply: Confidence based, Support based and hybrid function
Begin
{
1. Derive the lower support threshold and Scan the database to find the sensitive itemsets
2. Randomly generate a population of n individuals, with each gene being the ID number of the transaction to be hidden
3. Repeat
{
1.Calculate the fitness value of each chromosome Ci in the population // according to the selected fitness function
2. Execute the crossover and the mutation operations on the population
3. Probabilistically choose individuals for the next generation based on the selection scheme
}
Until the termination criterion is not satisfied.
4. create the output hidden transaction numbers in the best chromosome to users
}
End

In this method, sensitive data defined by the user is the input of methods. Firstly, a lower support threshold set constantly and then the sensitive itemsets derive from the dataset. This threshold indicates the limit of deleting sensitive items set in following steps. To start the operation of a GA method usually a random population generate, but in this method, this random population set with the ID number of the sensitive transactions that defined in the previous step.

The most important part of the method is the fitness function that directly influences the usefulness of the method. We are going to define this fitness using some evaluation formulas used for checking a privacy protection method. In the loop of GA method, at step 2 the crossover and mutation operation execute on the population. Population means the extracted sensitive itemset. The last step of the loop chooses the next generation based on a selection scheme. This loop executes as long as the termination criteria met which may be reaching an optimal set of sensitive items or some extra factors.

5 DISCUSSION

Privacy-preserving in e-business comprises the issue of big data and still is an open problem. Using a new methodology like soft computing may bring much needed improvements.

Neural Networks like SOM clustering is a relatively newly suggested approach to privacy-preservation and still needs to be worked on for two central factors, optimizing the time for training the network and privacy-preserving of its different parts. Hybrid methods based on a Neural Network in combination with other soft computing methods like fuzzifying before training the neurons to close the privacy gaps may yield better results.

The designed methods in the proposed model in this paper have such a structure. The SPE model described in Figure 5 provides the practitioner with the ability to keep the balance between privacy protection and performance. As discussed for Method 2, administrators decide about applying privacy methods defined in different parts of the system. In this way, whenever they want to improve the level of protection, they should just enable the methods in those parts of the system that may be confronted with possible attacks.

First evaluations of simple and small datasets show positive results on both sides, privacy protection as well as performance. However, to precisely evaluate the usefulness of this model, further simulations with huge datasets are required to prove the practical feasibility of the concepts. Thus, in future work based on this model, independent simulations using real data sets like the UCI machine learning and the Movielens datasets will be used. Moreover, extra soft computing methods such as Deep Learning Neural Networks and Back Propagation will be implemented, because of the similarity in structure with the SPE models and they will be compared with the currently proposed model.

In case of positive results of the simulations, a further comprehensive case study will implement the model. Recommender systems that currently are a topic of discussion in e-business applications from both sides, privacy and utility, seem to be the most suitable case study for evaluating the defined model.

All in all, the expected results of applying this model, including various methods, would be sensitive data better protected against privacy attacks in different parts of SOM clustering when run in a distributed environment. It should be noted that this improvement of privacy protection would not sacrifice the efficiency of regular SOM clustering. Also, the amount of information loss would not be considerable. In other words, the result of SOM clustering before and after applying the proposed methods in SPE model should be relatively similar in terms of performance and accuracy.

6 CONCLUSION

The idea presented in this paper focuses on the issue of privacy of individuals in e-business applications, which involves big data and therefore data mining techniques. Data mining on big data in combination with privacy-preservation is still an open problem. Among lots of methods proposed to improve privacy, the lack of a strong method that could protect privacy and also keeps efficiency and accuracy of the data mining tasks at hand, still exists. Therefore, newer methodologies like soft computing, also known as machine learning, seem to be more useful for closing these gaps. The proposed model in this paper contributes to solving the defined problem in e-business environments. The SPE model is flexible and helps system administrators to keep a balance between performance and privacy protection. Two privacy-preserving methods have been introduced for the SPE model which are independent and arbitrary to implement. First results prove the usefulness of the model and the methods, respectively. However more simulations with huge datasets are still required to check the utility of the SPE model in general. The result of the proposed model in this paper is sensitive data being protected against privacy attacks in SOM clustering without significantly jeopardizing the efficiency and accuracy of the general process.

ACKNOWLEDGEMENTS

This paper is partially funded by SBA Research, Vienna, Austria.

REFERENCES

Aggarwal, C. C. & Philip, S. Y. 2008. A general survey of privacy-preserving data mining models and algorithms. *Privacy-preserving data mining*. Springer.

Agrawal, R. & Srikant, R. Privacy-preserving data mining. ACM Sigmod Record, 2000. ACM, 439-450.

Amiri, F. & Quirchmayr, G. 2017. A comparative study on innovative approaches for privacy-preserving in knowledge discovery. *ICIME ,ACM*.

Blocki, J. & Williams, R. Resolving The Complexity of some data privacy problems. International Colloquium on Automata, Languages, and Programming, 2010. Springer, 393-404.

Bonizzoni, P., Della Vedova, G. & Dondi, R. The k-anonymity problem is hard. International Symposium on Fundamentals of Computation Theory, 2009. Springer, 26-37.

Chen, H. 2006. Intelligence and security informatics: information systems perspective. *Decision Support Systems*, 41**,** 555-559.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. & Uthurusamy, R. 1996. *Advances in knowledge discovery and data mining*, AAAI press Menlo Park.

Gorgonio, F. L. & Costa, J. A. F. Combining parallel self-organizing maps and k-means to cluster distributed data. Computational Science and Engineering Workshops, 2008. CSEWORKSHOPS'08. *11th IEEE International Conference on, 2008. IEEE, 53-58*.

Han, S. & Ng, W. K. Privacy-preserving self-organizing map. DaWaK, 2007. *Springer, 428-437*.

Jahan, T., Narasimha, G. & Rao, C. G. 2014. A Comparative Study of Data Perturbation Using Fuzzy Logic to Preserve Privacy. *Networks and Communications (NetCom2013)*. Springer.

Kohonen, T. 1982. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43**,** 59-69.

Kohonen, T. & Maps, S.-O. 1995. Springer series in information sciences. *Self-organizing maps*, 30.

Li, P., Chen, Z., Yang, L. T., Zhao, L. & Zhang, Q. 2017. A privacy-preserving high-order neuro-fuzzy c-means algorithm with cloud computing. *Neurocomputing*, 256**,** 82-89.

Maimon, O. & Rokach, L. 2007. *Soft Computing for knowledge discovery and data mining*, Springer Science & Business Media.

Malik, M. B., Ghazi, M. A. & Ali, R. Privacy preserving data mining techniques: current scenario and future prospects. Computer and Communication Technology (ICCCT), 2012 Third International Conference on, 2012. IEEE, 26-32.

Mingoti, S. A. & Lima, J. O. 2006. Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms. *European Journal of Operational Research*, 174**,** 1742-1759.

Mitra, S., Pal, S. K. & Mitra, P. 2002. Data Mining in soft computing framework: a survey. *IEEE transactions on neural networks*, 13**,** 3-14.

Narwaria, M. & Arya, S. Privacy Preserving Data Mining—'A state of the art'. Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on, 2016. IEEE, 2108-2112.

Roh, T. H., Oh, K. J. & Han, I. 2003. The Collaborative filtering recommendation based on SOM cluster-indexing CBR. *Expert systems with applications,* 25**,** 413-423.

Samet, S. & Miri, A. 2012. Privacy-Preserving Back-propagation and extreme learning machine algorithms. *Data & Knowledge Engineering,* 79**,** 40-61.

Saxena, V. & Pushkar, S. Fuzzy-Based Privacy Preserving Approach in Centralized Database Environment. Advances in Computational Intelligence: Proceedings of International Conference on Computational Intelligence 2015, 2017. Springer, 299-307.

Verykios, V. S., Bertino, E., Fovino, I. N., Provenza, L. P., Saygin, Y. & Theodoridis, Y. 2004a. State-of-the-art in privacy preserving data mining. *ACM Sigmod Record,* 33**,** 50-57.

Verykios, V. S., Elmagarmid, A. K., Bertino, E., Saygin, Y. & Dasseni, E. 2004b. Association rule hiding. *IEEE Transactions on knowledge and data engineering,* 16**,** 434-447.

Von Alan, R. H., March, S. T., Park, J. & Ram, S. 2004. Design science in information systems research. *MIS quarterly,* 28**,** 75-105.

Westin, A. 1999. Freebies and privacy: What net users think.

Wu, X., Chu, C.-H., Wang, Y., Liu, F. & Yue, D. 2007. Privacy preserving data mining research: current status and key issues. *Computational Science–ICCS 2007***,** 762-772.

Zadeh, L. A. 1994. Fuzzy logic, neural networks, and soft computing. *Communications of the ACM,* 37**,** 77-85.

Zhang, G. P. 2009. Neural networks for data mining. *Data mining and knowledge discovery handbook.* Springer.