universität wien

SBA Research

Towards Data Anonymization in Data Mining via Meta-Heuristic Approaches

Fatemeh Amiri, Gerald Quirchmayr, Peter Kieseberg,
Edgar Weippl, and Alessio Bertone

University of Vienna, Faculty of Computer Science, Vienna, Austria
SBA Research GmbH, Vienna, Austria

DPM 2019, 26th September, Luxembourg.

Agenda

- Introduction
- Background
- Problem Definition
- A Meta-heuristic Approaches for Anonymization
- Experiments
- Results
- Conclusion

Introduction...

- Privacy-preserving in big data - PPDM

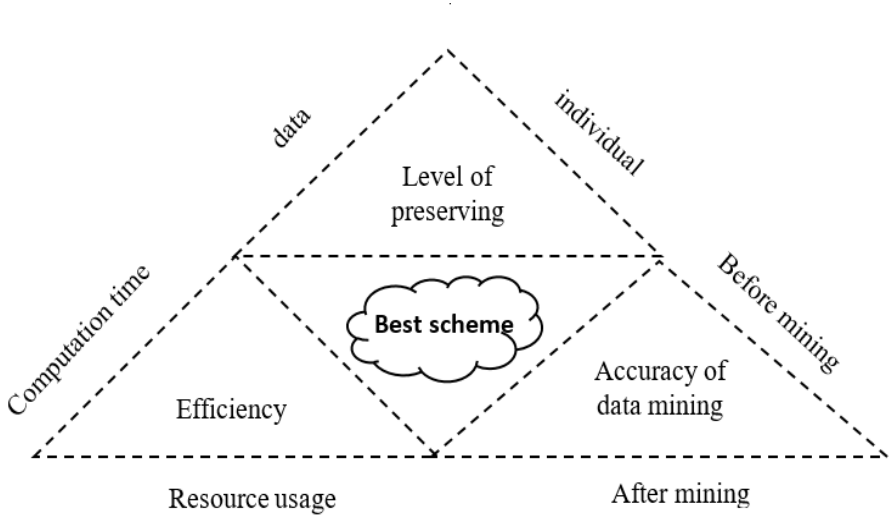
is one of the crucial factors in adopting online transactions by users and known as an NP-hard problem

- In this paper

a meta-heuristics model proposed to protect the confidentiality of data through anonymization.

- The aim is

to minimize information loss as well as the maximization of privacy protection using Genetic algorithms and fuzzy sets.



Background...

- The problem to be solved, PPDM, introduced by
  - Agrawal, R., & Srikant, R. (2000). Privacy-preserving data mining. Paper presented at the ACM Sigmod Record.
- Most of the existing PPDM approaches use conventional techniques
  - like perturbation, generalization, suppression and k-anonymity and new versions (like l-diversity and t-closeness)
    - Xu,L.,Yung,J.,Ren,Y.:Informationsecurityinbigdata:privacyanddatamining. IEEE Access, vol: 2, pp: 1149–1176(2014).
- We compared the conventional vs. soft methods in PPDM in:
  - Amiri, F., Quirchmayr, G.: A comparative study on innovative approaches for privacy-preserving in knowledge discovery, ICIME ,ACM (2017).
- The ideal aim is to minimize the selection of sensitive data from the database:
  - Sivanandam, S., Deepa, S. : Genetic algorithm optimization problems: in Introduction to Genetic Algorithms. , Springer. pp:165–209(2008).
- This work is based on the method, **GASOM** introduced in:
  - Amiri,F.,G.Quirchmayr:SensitiveDataAnonymizationUsingGeneticAlgorithms for SOM-based Clustering: Secureware (2018).
  - Here, a new model based on meta-heuristics is introduced that uses GAs and Fuzzy sets to anonymize selective sensitive items without compromising the utility of a SOM clustering. So that, it is shortened as **GAFSOM**.

Problem Definition

- This paper introduces a meta heuristics model in a specific use case:

  “Instead of anonymizing everything in the database, in case of knowing the sensitive items, we find them in database and we just apply the anonymization just on this portion of data, not all of them!”.

- As a case study, we focus on unsupervised clustering (SOM) tasks because of their wide application and limited privacy-preserving methods.

- A structural anonymization approach based on meta-heuristics approaches applied.

- First, a subset is extracted with a specially designed Genetic Algorithm (GA). The aim is to minimize the selection of sensitive data from the database. The output of the GA algorithm is used as the input of a fuzzy membership function to anonymize the content of the sensitive subset. Finally, the result of this step is appended to the primary database to be imported for the usual clustering data mining task.

Problem Definition

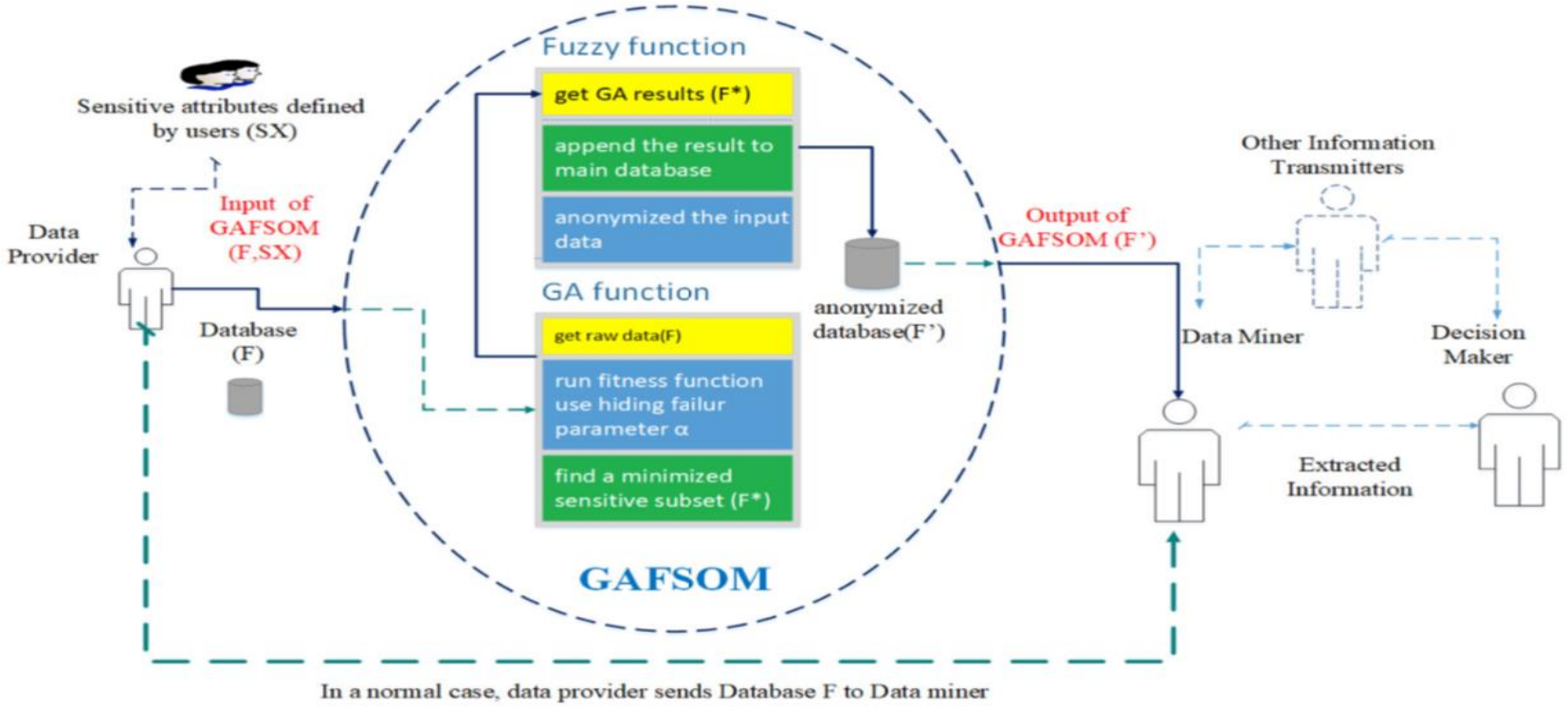


Fig. 1. A simple illustration of the application scenario with the proposed model (GAFSOM). Data provider sends the (F, SX) as the input of GAFSOM (SX is defined by user). The process begins from GA function. Then the process continues by sending the F* to fuzzy function. Anonymized Database F is the output of GAFSOM which will be sent to data miner.

A Meta-heuristic Approaches for Anonymization

- **Definition 1. (Input and Output).**

F is the original database with a set of sensitive data to be hidden SX = S1, SX2, ..., SXn. SXi is a field in the current record that should fulfil the user-defined criteria in order to be selected as a sensitive item. F' be the output of GAFSOM that is an anonymized database.

- **Definition 2. (Optimization Problem).**

The problem is to find the optimum partition set F^* to achieve an almost minimize Information Loss-IL (F,F^*) value for the given graph F. As an optimization problem we expect:

$$SOM(F) \approx SOM(F^*)$$

A Meta-heuristic Approaches for Anonymization

- **Definition 3. (Fitness Function).**

Fitness Function is the heart of a GA method and it is used ***to evaluate the hiding failures*** of each processed transaction; we assess the hiding failures of each processed transaction in the anonymization process using:

$$\alpha^j(S_x) = \frac{MAX_s x - freq(S_x) + 1}{MAX_s x - \lceil |F| \times MST \rceil + 1}$$

Where MAXsx is the maximum number of sensitive data of record with $F(S_x)$, $|F|$ is the number of records and $freq(S_x)$ is the frequency of SX in the current record. MST (Minimum Support Threshold) is a pre-defined parameter that limits the number of records to be selected as sensitive records. We use MST as a condition of termination of the GA function as an influencing parameter on the speed and quality of GA function.

The overall amount of α per record defined by:

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

A Meta-heuristic Approaches for Anonymization

- **Definition 4. (Termination of Algorithm).**

Threshold of termination defines a minimum support threshold ratio MST, as the percentage of the minimum support threshold used in the GA algorithm and plays a crucial role.

✓ MST, Minimum Support Threshold

Definition 5. (GA Operation Parameters).

Table 1. Parameters used in designed genetic algorithm of GAFSOM.

Parameter	value
Creation method	Pop Function
Population Size	50
Generations	100
Population Type	bit string
Selection function selection	tournament
Mutation function mutation	uniform, 0.1
Crossover function crossover	arithmetic,0.8
Elite Count	2
StallGen Limit	100

A Meta-heuristic Approaches for Anonymization

- **Definition 6. (Fuzzifying Process).**

In GAFSOM, the output of the GA method is defined as the input of a fuzzy function. In this paper, the *Triangular Membership* Function is used to anonymized the sensitive content.

- **As a Case Study**

Kohonen Maps put in practice through Self Organizing Map (SOM) applied to test the validity of the proposed model. SOM suffers from some privacy gaps and also demands a computationally, highly complex task. The experimental results show an improvement of protection of sensitive data without compromising cluster quality and optimality.

Experiment Data

For experiments, we use two real-world datasets :

- The Adult dataset, which is released by the UCI Machine Learning repository for research purpose. There are 14 attributes and 48, 842 records in total.

- Bank Marketing: This dataset is generated through direct marketing campaigns (phone calls) of a Portuguese banking institution. It contains 45,212 records and 17 attributes.

Experiment Test Cases

Two different test cases to evaluate the proposed methods presented:

- **Test case 1 (aim: random execution):** In Adult dataset when the sensitive criteria defined by data miner as black female who are post graduated and work more than 20 hours per week and younger than 30 years old.

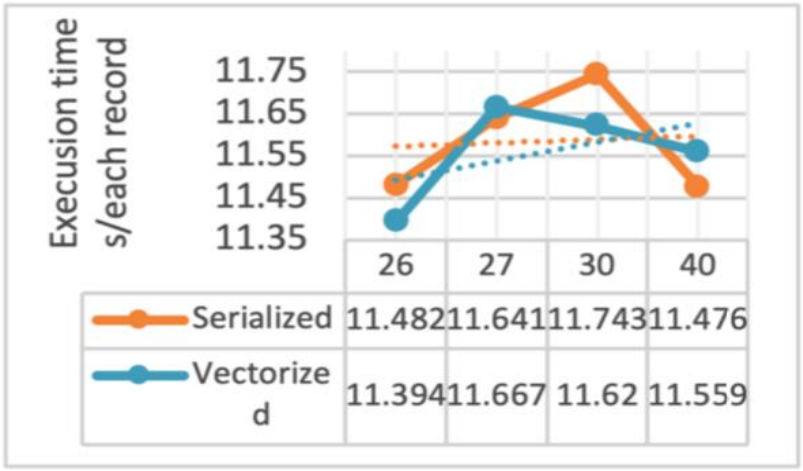
Design: a test case including 1000 records/tuples, which includes age, work-class, gender, education, and race as sensitive attributes.

- **Test case 2 (aim: worst case-too many items defined as sensitive**). In Bank dataset when the sensitive criteria defined by data miner as any young employee who is married and work at high ranked position like manager.

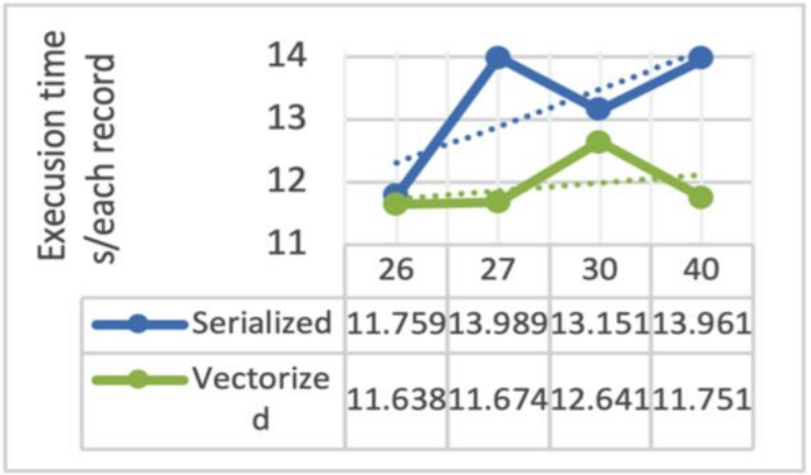
Design: those tuples with more sensitive data selected for test case. For Bank dataset a bigger test case including 3000 tuples is selected. The selected sensitive attributes are: age work-class and marital case.

Execution Time

The algorithms were implemented in MATLAB, and executed on a VM/ Linux Ubuntu platform with four vCPU in Intel(R) Xeon (R) E5-2650 v4 processors and 4 GB memory.



a) Serialized & Vectorized GA in Adult dataset



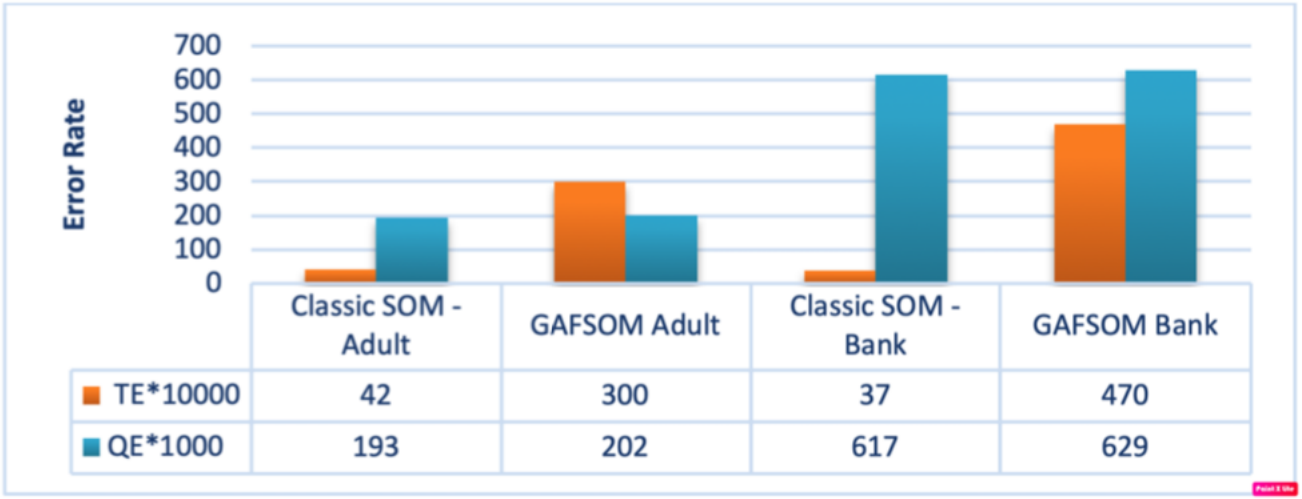
b) Serialized & Vectorized GA in Bank dataset

Fig. 2. Execution time for Adult and Bank dataset with different values for minimum support threshold, MST. Execution time is tested in both serialized and vectorized cases of GA function. Results show that vectorization is faster. Also, the impact of MST parameter on execution times is evaluated. As you can see, with the increasing of MST (horizontal axis) in both datasets, execution time also increases.

Analysis of the Accuracy and IL of GAFSOM

For measuring the results of SOM clustering two factors tested:

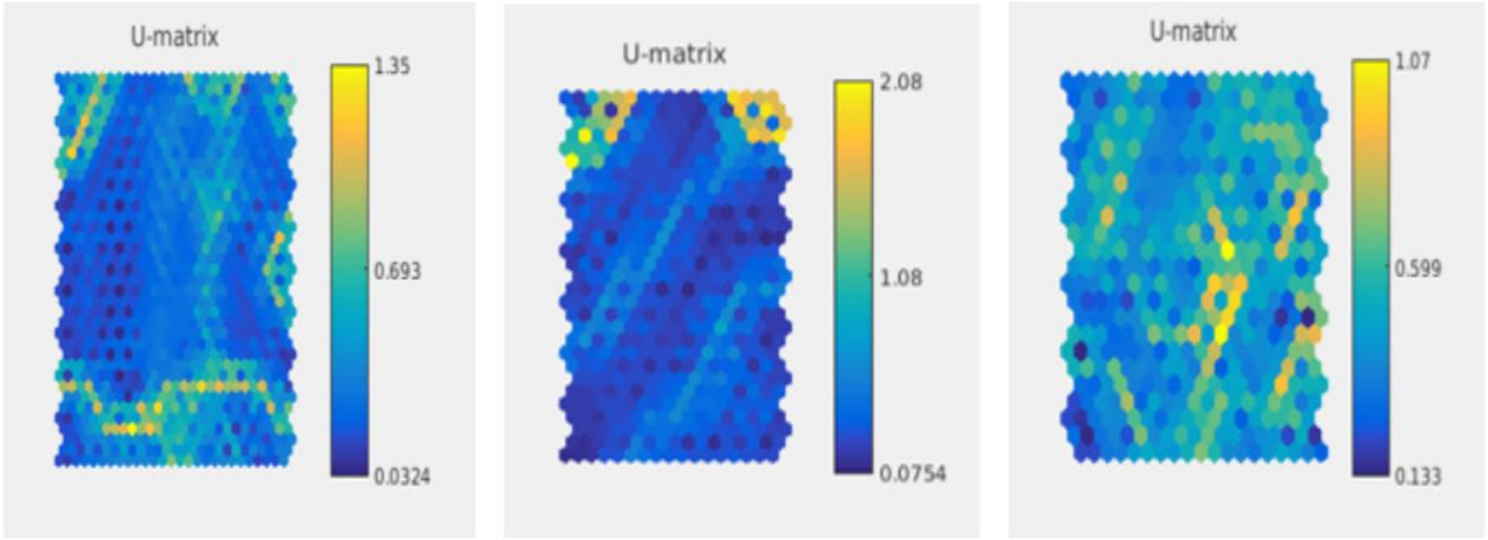
- Quantization Error (QE): the average distance between current BMU and each data vector :
- Topographic Error (TE): describes how well the SOM preserves the topology of the studied data set:



	Classic SOM - Adult	GAFSOM Adult	Classic SOM - Bank	GAFSOM Bank
TE*10000	42	300	37	470
QE*1000	193	202	617	629

Fig. 3. Accuracy measure results

Analysis of the Accuracy and IL of GAFSOM



(a) Comparing the U-matrix of classic SOM

(b) Comparing the U-matrix of GASOM

(c) Comparing the U-matrix of GAFSOMt

Fig. 5. Comparison of U-Matrixes for Bank dataset: (a) presents the clustering result of classic SOM for original dataset. (b) shows clustering applied on the anonymized dataset that is generated by GASOM. As you can see, the trend of clusters is similar to (a).(c) illustrates the experiments of clustering applied to the database retrieved by GAFSOM. This time the similarity between clusters is distorted. Though the accuracy experiments demonstrate the usefulness of GAGSOM, the information loss still needs to be decreased.

Conclusion

- PPDM using meta-heuristic techniques brings smarter solutions not only to protect against privacy breaches but also to increase accuracy in the final results of data mining.
- We introduced GAFSOM method, which uses a combination of genetic algorithm and fuzzy sets for a trade-off between privacy and utility.
- The overhead of GAFSOM in negligible and using the topological error formulas of clustering its correctness proved.
- Experiment results show that selective deletion of valuable data items is less destructive than general anonymization done by fuzzification, so that complying with other similar techniques especially differential privacy is still preferable to taking preemptive steps to de-identify personal information in databases.
- In future work
  - Differential privacy will apply to perturb the selected sensitive items by GA in order to compare the validity with current work.
  - Privacy-as-a-service

Thank you for attention!