A Novel Model for Privacy Preserving in Data Mining using Meta Heuristic Techniques[⋆]

Fatemeh Amiri[1,2] and Gerald Qurichmayr[1,2]

[1] Department of Computer Science,University of Vienna,Vienna, Austria
[2] SBA Research Institute, Vienna, Austria
{amirif86,gerald.quirchmayr}@univie.ac.at

Abstract. To mitigate privacy breaches in Big Data processes, several approaches have been proposed. However, they typically not only sacrifice the accuracy of data mining results but also the utility. Utility indicates both accuracy and efficiency (time/memory) of the general process. In this project, we propose a model that protects the confidentiality of data through anonymization. We use a structural anonymization approach based on Machine Learning and meta-heuristics. The model encompasses three phases. 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. In our proposed function, we try to minimize a parameter called hiding failure. So, the result of this GA function is a subset from the main dataset that is sensitive and should be anonymized. The second phase of the process is anonymization to hide sensitive data. 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 second step is appended to the primary database to be imported for the usual clustering data mining task. Kohonen Map clustering is used then as the case study, which was selected due to its popularity and its currently existing privacy gaps. It might perceive that the proposed model is similar to some traditional PPDM methods like the concept of differential privacy and k-anonymity. Despite the similarity of the concepts, we use a completely different approach, which is based on machine learning and heuristic techniques. As the state of the art of subject shows, the need for such a method to have a durable power of privacy protection and also a lower level of information loss is still highlighted. Our experimental results show an improvement of protection of sensitive data without considerably jeopardizing the clustering process.

[⋆] Supported by SBA Research Institute, Vienna, Austria