# Chapter 26
# Privacy Preserving Classification of Biomedical Data With Secure Removing of Duplicate Records

**Boudheb Tarik**
*EEDIS Laboratory, Djillali Liabes University, Sidi Bel-Abbes, Algeria*

**Elberrichi Zakaria**
https://orcid.org/0000-0002-3391-6280
*EEDIS Laboratory, Djillali Liabes University, Sidi Bel-Abbes, Algeria*

## ABSTRACT

*Classifying data is to automatically assign predefined classes to data. It is one of the main applications of data mining. Having complete access to all data is critical for building accurate models. Data can be highly sensitive, such as biomedical data, which cannot be disclosed or shared with third party, because it can harm individuals and organizations. The challenge is how to preserve privacy and usefulness of data. Privacy preserving classification addresses this problem. Collaborative models are constructed over networks without violating the data owners' privacy. In this article, the authors address two problems: privacy records deduplication of the same records and privacy-preserving classification. They propose a randomized hash technic for deduplication and an enhanced privacy preserving classification of biomedical data over horizontally distributed data based on two homomorphic encryptions. No private, intermediate or final results are disclosed. Experimentations show that their solution is efficient and secure without loss of accuracy.*

## 1. INTRODUCTION

The recent advances in storage and network technology have led to an explosion of the data. In healthcare domain, many institutions such as hospitals, clinics and laboratories are collecting patients' data (medical analyzes results, healthcare diagnosis, etc.) with the aim to extract a hidden and a valuable knowledge using powerful data mining tools. Data can be stored in different ways, centralized or distributed. In

DOI: 10.4018/978-1-7998-8954-0.ch026

horizontal distribution, different sites have different sets of records about the same topic with the same attributes. In vertical distribution, different sites might have different attributes of same records.

In order to improve patients' medical care, laboratories or hospitals want to perform collaborative computations to build more accurate models. But in contrast, due to the privacy laws or other concerns, they are reluctant to share their local data. However, data mining tools need a whole access to data. Face these constraints; privacy preserving data mining is becoming an important issue. Models are built without seeing the private data. Several solutions are proposed. Nowadays, privacy is included in many fields such as IoT, cloud, big data and distributed computing. In healthcare domain, it is becoming increasingly essential due to the confidentiality of patients' data (Yang et al., 2018).

The secure preprocessing step is crucial to data mining process, e.g., outliers and duplicate records can decrease performances. The problem of duplicate records has been largely studied. The researchers (Kołcz et al., 2003) have examined the impact of duplicate records and the rate of duplication on performances. The results obtained indicate that the presence of duplicate records in the learning sample pose a problem. The loss of classification accuracy can be strongly correlated with the contamination level. They affirm that, classifiers such as Naive Bayes and PAM can be robust, only if the presence of duplicate records is not extreme. The challenge of secure deduplication records, over distributed data, is how to detect duplicate records without having access to the private data. Many solutions were proposed based on hash functions, clustering data mining, computing distance between two records, etc.

Data mining classification over distributed data with respect of privacy has been considered too. The challenge is how to construct collaborative model without having access to the private data. The intermediate or final results can be sensitive too. They can leak some private information such as sum of values, number of instances, mean, variance, etc.

In this paper, the researchers propose an enhanced protocol to detect securely duplicate records and build securely a Naïve Bayes model without loss of performances. Local biomedical data and intermediate results are kept secret. The authors combine a customized hash technic, pallier and RSA cryptosystems. The solution is suitable for semi-honest model. The researchers imagine a real more complex environment where: (a) the network is not secured such as public network or internet. (b) Internal parties such as master site or external hackers can analyze network traffic, in real time, between two sites and can get some statistical private information.

The remainder of this paper is organized as follows: In section 2, the authors introduce the related work. In section 3, they present a background. In section 4, they explain their contributions. In Section 5, they present the proposed approach. In section 6, they perform experimentations. In section 7, they discuss the security of the solution. Finally, section 8 presents the conclusion and future works.

## 2. STATE OF THE ART

Data preprocessing is a vital task for data mining. It is mainly important for making data appropriate, e.g., avoiding duplicate records, estimating missing data, selecting best attributes, etc. Deduplication is a field of the preprocessing step. It can be local (centralized database), or over distributed data (multiple databases). Removal and prevention of duplication is an essential part of the security (Chang and Ramachandran, 2016). According to the authors (Yigzaw et al., 2017), duplicate records may lead to incorrect statistical results. Therefore, to increase the accuracy of analysis, deduplication is important.