Chapter 5
Machine Learning-Based Data Analytics With Privacy: Privacy-Preserving Data Analytics

Rupali Tajanpure
GITAM University, India

Akkalakshmi Muddana
GITAM University, India

ABSTRACT

Data analytics is a very common word today. Data is collected from various sources and analyzed for decision making. The decisions help for growing business, for healthcare support, as well as to keep track of some useful information on communication media. For the same data may be shared, stored, and analyzed. Each of these three processes involves threat of data leakage to hacker. To prevent this, privacy preservation algorithms are used. This chapter discusses about privacy preserving techniques right from data collection to analytics through data storage. The data classification techniques are also discussed for understanding of machine learning data analytics. At the end open issues in privacy preserving are also discussed.

INTRODUCTION

In today's world we use to talk about data excessively. Rather we think of data usefulness, its analysis and the space requirement of data. Also, data secrecy is very important. The data is nothing but the raw form of information, which needs to be processed and organized to gain information from it. The knowledge is the outcome of processed information with some insight and interpretation. Knowledge is useful to reach some decisions. The decisions can be the recommendation to user on online movements, decisions regarding the sales and production of different products, the healthcare decisions, decisions based on the input from sensors different sensors in the industry etc. User also get many recommendations once he start browsing different social media websites. Even the decision can be the carrier choice of the

DOI: 10.4018/978-1-6684-6519-6.ch005

students based on his hobbies, area of interests. For all this proper data analysis and its accuracy is very important. There are various parameters involved in this analysis. Also, there is need for easy and fast data analysis. Mostly the data generated through internet is multidimensional and big in size. Here big data analysis techniques like Hadoop, hive is in demand. Different tasks carried out for the data analysis are pattern mining, classification, prediction, characterization, discrimination and clustering. These all are called as data mining functionalities. The interesting patterns mined from information is said to be the knowledge of data. The main aim of data analysis is to find the useful patterns in the data which increases the productivity of data in many dimensions like classification, predictions and diagnosis etc. Due to use of Internet of things in every sector, the sensor-based applications are increased and hence the applications of machine learning and its analysis. Business intelligence is one of the applications where data analysis (Han & Kamber, 2006) plays a crucial role. The parameters like customer need and demand base on customer feedback, supply, resources, market strategies, strength and weaknesses of competitors are handled by decisions based on data analytics.

BACKGROUND

Privacy preservation of collected data is important in respect of decision making of analyzed data. This is true in every domain like healthcare, IoT based systems, business intelligence, decision support system etc. In literature data perturbation is done using randomization, K-anonymity, data suppression and data generalization. This is applicable to all types of data as stated in the work (Ge et al., 2005). Machine learning, deep learning, distributed machine learning are areas to work on privacy (Zhou et al., 2021; Niu et al., 2020; Mohassel & Zhang, 2017). K-anonymity works on the principle of hiding individuality of data. But it becomes difficult while dealing with high dimensional data as it becomes difficult to preserve privacy (Shokri & Shmatikova, 2015; Sweeney, 2002). Datafly, µ-Argus and k-Similar are some k-anonymities based modified algorithms presented in literature (Aggarwal, 2005). Some group-based privacy preserving algorithms are also proposed which works on different privacy aspects (Majid Rafiei, 2021). Author discussed methods of privacy preservation, metrics to assess privacy and application areas of privacy preservation in detail in (Mendes & Vilela, 2017). A review of differential privacy preserving in machine learning for balancing privacy and utility of data is put forth by author (Gong et al., 2020).

For effective mining, open government data can be merged into one data set. This helps the user to balance well between data utility and data disclosure risk (Lee & Jun, 2021). Author introduced review of privacy-preserving mechanisms on heterogeneous data types with systematic analysis (Cunha, Mendes, & Vilela, 2021). Matatov, Rokach, and Maimon (2020) proposed genetic algorithm approach for privacy preserving. k-anonymity based method is used to evaluate classification of ten datasets.

PRIVACY AND DATA MINING

Knowledge discovery from databases (KDD) is the process of extraction of knowledge from the collected raw data after data preprocessing and data mining. Data collection is the first step of KDD process.