Chapter 18
A Privacy-Preserving Feature Extraction Method for Big Data Analytics Based on Data-Independent Reusable Projection

Siddharth Ravindran
National Institute of Technology Puducherry, India

Aghila G.
National Institute of Technology Puducherry, India

ABSTRACT

Big data analytics is one of the key research areas ever since the advancement of internet technologies, social media, mobile networks, and internet of things (IoT). The volume of big data creates a major challenge to the data scientist while interpreting the information from raw data. The privacy of user data is an important issue faced by the users who utilize the computing resources from third party (i.e., cloud environment). This chapter proposed a data independent reusable projection (DIRP) technique for reducing the dimension of the original high dimensional data and also preserves the privacy of the data in analysis phase. The proposed method projects the high dimensional input data into the random low dimensional space. The data independent and distance preserving property helps the proposed method to reduce the computational complexity of the machine learning algorithm. The randomness of data masks the original input data which helps to solve the privacy issue during data analysis. The proposed algorithm has been tested with the MNIST hand written digit recognition dataset.

DOI: 10.4018/978-1-7998-8954-0.ch018

INTRODUCTION

In the era of ICE age (Information, Communication and Entertainment), the growth of the data is at an exponential rate. The statistics from IBM (Quick Facts and Stats on Big Data, n.d.) states that, there are approximately 294 billion of email sent and 230 million of tweets in a day and there are trillions of sensors populating the Internet of Things (IoT) with the real time data. The process of data generation has become a lot easier, thanks to the advancement in Smartphone, Internet and other sensor applications. In the current scenario, Data is not only a piece of information it is also an asset for the industries and organizations in order to reduce their operational cost and to improve the profit, apart from highly influencing working environment for betterment. The popular magazine forbes predicted that the revenue of the worldwide big data market for software and services are expected to increase from 42 Billion USD (2018) to 103 Billion USD (2027) with the compound annual growth rate of 10.48% (Columbus, L, 2018). The real problem of big data analytics arise if the system is not capable of addressing 5 V's (Volume, Velocity, Variety, Value and Veracity) to harvest the best yield from the huge amount of data (Tsai, C.et al., 2015). The sophisticated hardware is not only the solution to address the complexities in big data but also requires significant contributions from the software. This chapter highlights the research issues especially in software related data analysis and applications.

Issues in Big Data Analytics

Big data analytics needs significant research concern for solving the large, complex and unstructured data collected from various independent sources. The applications of big data analytics has been widely expanded to almost all the engineering and science domain (Ouf, S., & Nasr, M. (2015)). The main constraint for big data analytics is the various levels of information which could be obtained or inferred from the available large chunk of data. Among the 5 V's of big data, Volume is one of the key research areas since the voluminous of data frustrates the data scientist and programmers to take insights from the data. The more we collect the data, the more we get the information which is not always true in the case of big data analytics (Chen, C. P., & Zhang, C., 2014). All the collected data may not be useful for the data analysis because of the huge number of uninformative features in the data set. The humungous amount of data increases both the storage and computational complexity in big data analytics. This problem is commonly termed as Curse of Dimensionality (Xie et al., 2016). The traditional machine learning algorithm does not suit well for handling big data analytics due to the complex characteristics of big data like huge volume, different variety of data, etc. The dimension reduction techniques come into the scenario in order to reduce the complexity of existing machine learning algorithm. For example, consider a scenario where the data scientist wants to train a model to find the activity performed by the user using smart phone sensors using k- Nearest Neighbor (k-NN) classifier. k-NN is one of the most influential machine learning algorithm and it follows lazy leaning approach. If the user performs any activity, then the sensors transfer the test data to the k-NN for classification. k-NN computes distance between the test data with all the training data to find the k- nearest neighbors. The complex computation during testing makes the k-NN algorithm not suitable for big data analytics. Dimension reduction techniques reduce the complexity of machine learning algorithm either by projecting the data into lower dimensional space or by removing the unwanted features.