Name: Adeel Shiraz

Supervisor: Prof. Tanvir Ahmad

Department: Computer Engineering

Title of Thesis: Big Data Mining through various tools and Techniques

Abstract

Big data mining refers to mining of data without any constraints of size, variety or speed. The traditional machine learning algorithms are slow and also not scalable, whereas an algorithm for mining of big data should be scalable as well as fast. The big data algorithms employ strategies like sampling, incremental or distributed processing to enable them handle large datasets. Distributed machine learning algorithms are provided by libraries like H2O.ai which can run distributed jobs over a cluster of computers to handle large datasets. Existing machine learning libraries for mining of big data, like Apache Mahout and Spark MLlib focus on recommender systems, classification, and clustering. Anomaly detection is an area of data mining which finds its applications in many areas; however, none of the existing big data tools provide algorithms for anomaly detection. Anomaly detection in big data is a more difficult task compared to classification/clustering as the percentage of the anomalous points is extremely low (below 5% in most of the cases). There are many techniques for anomaly detection; however distance-based unsupervised learning techniques are most common. The existing distance-based solutions are not scalable and also require setting of few parameters which is generally done by hit-and-trial. In the present work, an optimized distance-based algorithm for anomaly detection has been proposed. The algorithm is implemented in Apache Spark for scalability, the parameters have

been optimized by swarm intelligence meta-heuristics, and an optimal function to generate the distance matrix has also been identified. Apart from this distance-based approach, a neural network based approach is also proposed. Replicator Neural Network (RNN) is a neural network which regenerates the input at the output layer, and can be utilized for anomaly detection. In the present work, RNN has been optimized using Extreme Learning Machine (ELM) learning mechanism, and Garson's algorithm. The proposed algorithm is implemented on TensorFlow to enable its execution on GPUs for high speed. Both of the proposed algorithms are compared with state-of-the-art algorithms, and the results show that the proposed algorithms outperform the existing algorithms. Among both the proposed algorithms, the neural network based algorithm is faster than the distance-based algorithm.