ABSTRACT

Computers are used extensively in business and scientific data processing. Organisations have archived tremendous amount of data in few decades. Normally, data is generated through MIS system and is utilised by decision support system for taking strategic decisions. Decision support system makes use of data mining techniques that is applied upon large pool of data. A number of applications require data mining techniques for extracting interesting and potentially useful information. These techniques may be exploited for analysis of data related to finance, e-commerce, medical field, retail industry, telecommunication industry, World Wide Web, biological sciences and many more. Data mining is also extensively applied for scientific data analysis, intrusion detection, information security and fraud detection.

Data mining techniques are based on algorithms. These algorithms are both compute and data intensive. The main issues with data mining algorithms are to handle the large size data having distributed location. For timely analysis and mining of such data, affordable and efficient data mining solutions are required.

Distributed sources of voluminous data have raised the need of distributed data mining. Conventional data mining techniques works well on structured data which is clean, pre-processed and properly arranged either in the form of structured files, databases or data warehouse. These techniques are based upon centralized data store. However, the size of data keeps growing day by day and several limitations arise in mining large distributed data. Parallel and distributed computation in such scenario can help to improve the efficiency of existing data mining techniques.
Different challenges in the field of data mining are thoroughly examined and latest trends in centralized, parallel and distributed data mining are investigated. Solutions are proposed to evaluate the performance of different data mining algorithms.

The initial investigation is related to application of different data mining algorithms on different application domains like ATM machine maintenance and World Health Organisation. The experimentation indicated the significance of data mining in these domains.

Parallel processing techniques are investigated to enhance the performance of data mining algorithms. Parallel data mining techniques were analysed using multi core CPU and highly parallel GPUs (Graphical Processing Unit). Data compaction strategy and vertical data layout is leveraged for performance improvement. A parallel frequent item set mining algorithm is proposed for execution on GPUs. Experimental results have demonstrated multi fold performance gain in frequent item set mining when applied using a combination of CPU and GPU.

Further, the mining techniques are modified so as to make them suitable for distributed computation. Cloud technology support distributed computations. Cloud based tools and technique provides an infrastructure for distributed data mining. Therefore these techniques are exploited for evaluating data mining algorithms. A solution for distributed data mining is designed by taking the advantages of cloud technology. This solution is intended to use private cloud and is created on Hadoop platform. A distributed version of K-Means algorithm is
designed and a common interface is also designed from where any user can execute data mining algorithm on specified number of nodes. Hadoop multinode setup requires many configuration steps but the proposed solution successfully incorporates distributed computation without any user level configuration. Data loading process is synchronised for different nodes. Data can be loaded and mined from any node on the network through a user friendly interface. It is observed that the proposed framework is helpful in mining large size data efficiently through a cluster of nodes.

Next, Map Reduce programming model is thoroughly investigated and its performance bottlenecks are identified. Map Reduce is popular for its simplicity and is generally used for parallel processing. But for an algorithm that requires iterative and parallel computation both, it shows several performance penalties. An extended Map Reduce framework is proposed that works better on iterative algorithms. A comparison of Hadoop and the proposed framework is carried out and the results indicate that novel framework executes iterative data mining algorithm efficiently.