Data mining is a challenging task these days because of issues like increase in data size and distributed nature of data. It is observed that data mining techniques can be speeded up by proper combination of parallel and distributed approaches. Data mining algorithms are thoroughly reviewed and then applied and analysed in centralized, parallel and distributed environment during this research.

During the application and evaluation of parallel data mining technique, it was identified that GPU can be utilised for performance enhancement in compute intensive data mining tasks. CUDA a programming language designed for GPU can be used to design faster algorithms. Parallel algorithms can run faster on CPU+GPU combination. The performance analysis of Frequent Item set Mining algorithm on GPU is presented in this research. A new parallel version of Frequent Itemset Mining algorithm is proposed. It is observed that combination of data compaction technique, vertical data format and parallel execution resulted in overall improvement in performance of FIM algorithm. At present, low end 48 core single GPU, is used for experimentation, but in future, same approach can be implemented using cluster of multiple GPUs.

After carefully examining the current trends in distributed data mining, it is identified that distributed data mining techniques in near future will be oriented towards Hadoop and G-Hadoop. Hadoop is based on cloud and is used for big data analytics. Limited work is carried out on data mining based on cloud. Thus, the research targets to exploit cloud based tools and techniques for data mining.

The applicability of DM techniques under cloud environment is explored further. A framework is developed to facilitate easy porting of DM applications to cloud environment. The advantage is that working on single desktop, user can virtually utilize the computing power of the desktop nodes under private cloud environment. Files to be mined are automatically stored on ownCloud since they are synchronized. Currently, K- Means algorithm can be accessed through this
solution. Distributed Map Reduce version of K-Means algorithm is designed for the framework. Files can be selected and can be provided to K-Means algorithm through an interface. Algorithm runs in a distributed fashion on Hadoop platform. The framework is tested to compare the mining performance for different large size inputs. A generalized solution that supports Distributed Data Mining and Storage as a service on private cloud has come out to be an outcome of our study. Currently the framework is limited to K-Means algorithm only. The framework can be extended further for other data mining algorithms.

Further, a framework is proposed that is capable of executing iterative algorithms efficiently. Hadoop that is designed for parallel applications is not suitable for iterative algorithms. If iterative algorithms are run using Hadoop MapReduce platform, it creates independent jobs in every iteration. Because of this overhead, the performance of Hadoop degrades. The framework proposed in this research is compared with typical Hadoop and the experimental results indicate that new framework is efficient for iterative algorithms.

Future Enhancements:

- Using GPUs in clusters of computers, efficient solutions of data mining can be developed.
- Map-reduce-like models can be developed for programming in heterogeneous CPU-GPU clusters.
- Combining Mahout, Sqoop, Flume and Mongo-Hadoop Connector, mine NoSQL Big databases can be mined.
- CUDA a parallel programming language that is designed for GPU can run within MapReduce for further improving efficiency of mining compute-intensive tasks over petabytes of data.
- Use of G-Hadoop with G-Farm file system, a MapReduce framework can be used for large-scale distributed computing on distributed data.