|  |  |
| --- | --- |
| Group Members  Siddhant Kulkarni, Ritesh Sangurmath |  |

Exploration of parallel implementation of Machine Learning algorithms for shared memory architecture

Project Proposal for CSC5551

|  |  |
| --- | --- |
| Date | Project Performed By: |
| October 20, 2016 | Group Members  Siddhant Kulkarni, Ritesh Sangurmath |

# Motivation & Goals

Over the past couple of decades, machine learning has seen an exponential growth in research interest from scholars in all domains. As the value of information becomes increasingly evident, this growth does not show signs of slowing down. The fact that machine learning can be applied to problems in almost every domain has opened doors that we could have never imagined. Today, it plays a role in everything from research on particle accelerators to simply finding out if a person who bought diapers will buy beer. This appeal has also introduced several new challenges in the field of machine learning. Introduction of World Wide Web and Business Analytics shifted the focus from “how do we deal with this lack of data?” to “how do we deal with all this data?!”. In order to deal with the large volumes and extreme velocity of the data being produced from billions of data sources has become a major area of focus for several researchers. Specific applications require these huge amounts of data to be processed in near-real time (for example, self-driving cars need to take split second decisions).

In order to address this problems, researchers have started looking at parallel and distributed computing [1] as an alternative to traditional single processors single node computing. The reason being that the traditional systems are simply no longer capable of handling such computations. As part of this project we intend to implement and analyze machine learning algorithms and the applicability of parallel computing to improve the performance of the same.

Primary objective of this project is to develop serial and parallel implementations of two machine learning algorithms and analyze the challenges and benefits of the process.

# Scope of the Project

Following list defines the scope of all the tasks that will be undertaken as a part of this project:

1. Review literature on parallel implementations of machine learning algorithms.
2. Serial implementation of the following machine learning algorithms:
   1. Naïve Bayesian classifier
   2. K-means clustering
3. Extend the understanding of C++ and OpenMP constructs along with architecture of DOZER.
4. Parallel implementation of the following machine learning algorithms:
   1. Naïve Bayesian classifier
   2. K-means clustering
5. Analysis of empirical results and detailed observations.

# Environments

This project will build the implementation using C++ and OpenMP constructs which will be executed and evaluated using DOZER.

# Proposal Presentation Considerations

1. What is Machine Learning?
2. Importance of Machine Learning
3. Key challenges in Machine Learning
4. Introduction to Naïve Bayesian classifier
   1. Learning Model
   2. Classification Model
5. How to implement Naïve Bayesian serially?
6. Introduction to K-means clustering
7. How to implement K-means serially?
8. Proposed work
9. Conclusion

# Completion Presentation Considerations

1. Introduction of Machine Learning
2. Description of Challenges
3. Status of Work
4. Implemented Algorithms
5. Empirical Results
6. Observations
7. Conclusion
8. Future Work

# Work Distribution

|  |  |  |
| --- | --- | --- |
| **Work Distribution** | | |
| Task | Siddhant | Ritesh |
| Review literature | √ | √ |
| Serial implementation of Naïve Bayesian | √ |  |
| Serial implementation of K-means |  | √ |
| Parallel implementation of Naïve Bayesian | √ |  |
| Parallel implementation of K-means |  | √ |
| Analysis of Empirical Results | √ | √ |

# Bibliography

1. Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu. 2002. An Efficient k-Means Clustering Algorithm: Analysis and Implementation. *IEEE Trans. Pattern Anal. Mach. Intell*. 24, 7 (July 2002), 881-892.