CHAPTER - 6

CONCLUSION AND FUTURE DIRECTIONS

6.1 INTRODUCTION

This chapter visualizes all the main contributions and the noteworthy accomplishments done through this research to benefit the Big Data researchers. It summarizes the detailed work that envisages the significant achievements made through this research work. The conclusion followed by the summary highlights the salient feature contributions made in the area of Apache Spark using Machine Learning algorithms. Finally, future directions for the developed model ‘Framework model for prediction on datasets using Machine Learning techniques’ gives an insight to the researchers with enriched possibilities to explore and extend the work.

6.2 CONCLUSION

While executing data analysis, jobs using parameters related to time and space are observed using Machine Learning techniques. The proposed model optimizes the Machine Learning techniques on a distributed environment using Spark framework to minimize the total execution time and space complexities for future predictions. By implementing Spark framework on top of Machine Learning techniques the efficiency of the model is improved by 70% when compared to the MapReduce paradigm. The study gives a clear picture of each learning algorithm with their specialized features in which the presented model is scrutinized with certain measures. Some measures of model are listed below:

- **Prediction:** This developed model analyses the Indian temperature data set collected between the years 1901 to 2015. It predicts the temperature for the subsequent years from 2016 to 2020. This model is applicable for any prediction dataset.
- **Time:** This model evaluates the time of Machine Learning algorithms and traces out the utilized time by predicting the values on various algorithms.
- **Space:** This developed model is implemented on Machine Learning methods which require space as a factor. It also evaluates the space required for each algorithm.
- **Spark Framework:** This model can replace the existing MapReduce model of Hadoop system with Spark framework for reducing complexities.
- **Job Allocation:** This model distributes the job in a cluster and computes the time, space utilization and analytical performance.
The architectural system of this model consists of Spark framework, which is of crucial relevance with the context. The application of PySpark on this model assists in enabling the development of large-scale Machine Learning algorithms. Machine Learning algorithms are expressed and compared with highly optimized execution plans over existing techniques.

Systematic empirical results present the benefit of optimization strategies such as blocking, local aggregation and the applicability scale up on diverse set of Machine Learning algorithms. Development of additional constructs to support Machine Learning meta-tasks such as model selection enables a large class of algorithms to be probed at an unprecedented data scale.

6.3 FUTURE DIRECTIONS

The objective of this thesis is to provide a deep understanding on model-based performance analysis for prediction based on the performance results and to generate alternatives to astound the performance flaws.

This model for prediction is simple and effective that motivates for further advancement by incorporating additional features such as humidity, moisture, fog and pollution consistency to identify better estimate of temperature for future analysis. These models are highly used for the exponential growth in computing power and figuring out best performance results. Application of such strategy can increase the accessibility of the model for future predictions.

Weather prediction substantially benefits a country economy by informing farmers about the expectations and apprise emergency services when adverse weather condition likely to take action.

Table 6.1 tabulates the various applications in time series domain that supports prediction.

- The rainfall datasets available in the existing data repository for all the states of India from 2004-2010 is used. Additional information with satellite images can also be included.
- More factors in rainfall prediction like humidity, pressure, temperature, wind, soil erosion, ground water can be utilized for assessing the natural disasters.
Additional toolboxes which are complementary to the problem are available and can be utilized in MATLAB as an extension of the work.

Table – 6.1: Applications in Time Series Domain for Prediction

<table>
<thead>
<tr>
<th>Domain</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meteorology</td>
<td>Temperature, Pressure, Wind</td>
</tr>
<tr>
<td>Finance</td>
<td>Exchange Rate, Speed of Import and Export</td>
</tr>
<tr>
<td>Economy</td>
<td>Gross National Product, Indexes of Economy</td>
</tr>
<tr>
<td>Marketing</td>
<td>Promotional Advertisements, Activities of Business, Sales</td>
</tr>
<tr>
<td>Industry</td>
<td>Electric Load, Power Consumption, Voltage, Sensors</td>
</tr>
<tr>
<td>Biomedicine</td>
<td>Physiological Signals, Heart Rate, Patient Temperature</td>
</tr>
<tr>
<td>Web</td>
<td>Clicks, Logs</td>
</tr>
<tr>
<td>Genomics</td>
<td>Time series of gene expression during cell cycle</td>
</tr>
</tbody>
</table>

Further, the research for short term can be extended to hourly and the day-to-day predictions with geographical locations can be implemented to impart regionalization.

6.4 OPEN PROBLEMS

The following are some of the open problems, which need to be addressed:

i. PARKING AVAILABILITY ON PEAK DAYS OF A WEEK

"When Solomon said there was a time and a place for everything he had not encountered the problem of parking his automobile."

-Bob Edwards, Broadcast Journalist.

Prediction models are exceedingly used to determine the traffic and parking availability based on various factors such as time, day of the week, weather, special events and holidays and so on. This model should be designed to incorporate some additional features such as finding the legal parking slots, knowing the count of parking areas and reasonably payable
parking slots. These models can be depicted by implementing the graphical layouts in multivariable graphs.

The complex problems like addressing the free parking spots in real time are detectable. Planar graphs that generates the structure of traffic flows are complex in different layouts. These models are in out of focus mode when the supply and demand are in constant flux.

To address this issue Machine Learning algorithms are applied to determine the parking allotment position. They are difficulties in finding relation between public transit and crowd sourcing. To solve this problem technical possibility can be built by interpreting parking difficulty features. This can be constructed using a truth data from crowd sourcing and appropriate Machine Learning model with a set of features.

ii. **BREAST CANCER RECURRENCE**

Patterns are the significant structures for detecting hidden relations and faulty composition learned through advanced methods. Models have been developed for determining such patterns through data sources like medical databases by tracking the information of the patients every two years. The prediction model presents a way in early detection of the disease for the best cure. The efficiency is based on the sensitivity, specificity and accuracy and by understanding the recurrence rate with least error rate on each model.

iii. **PROTEIN PREDICTION STRUCTURE**

Machine Learning methods are widely used in bioinformatics and computational biological systems. The development of Machine Learning methods for protein structure prediction is one of the most fundamental problems in structural biology and bioinformatics.

Protein structure prediction is such a complex problem that it is often decomposed and attacked at four different levels: 1-D prediction of structural features along the primary sequence of amino acids; 2-D prediction of spatial relationships between amino acids; 3-D prediction of the tertiary structure of a protein and 4-D prediction of the quaternary structure of a multi-protein complex.

A diverse set of both supervised and unsupervised Machine Learning methods has been applied over the years to tackle these problems. This has significantly contributed advanced state-of-the-art of protein structure prediction. Development and application of hidden Markov models, neural networks, support vector machines, Bayesian methods and clustering methods in 1-D, 2-D, 3-D and 4-D protein structure predictions are few mentionable works.
6.5 SUMMARY

This chapter gives a depiction on the overall functionality of the model. It is followed by focusing the future work with prevailing open issues. Some of the open problems discussed in this thesis can be addressed using the developed model, which can apparently benefit the future researchers.