Chapter – 2: Problem Formulation

In the last few decades in medical diagnosis, many researchers have contributed for breast cancer recurrence using machine learning techniques. A critical survey of the existing literature reveals that a good amount of work has been done in the application of data mining technique to develop predictive models for breast cancer recurrence in patients who were had been observed for more than two years.

Breast cancer recurrence system has been analyzed by various machine learning techniques such as C5.0, CART, QUEST, Artificial Neural Networks, and so on. In recent days, different data mining techniques have also been used to discover hidden patterns and relationships.

The existing machine learning techniques have been adopted to improve the accuracy and objectivity of the breast cancer diagnosis and prognosis. However, these approaches are tedious, time-consuming and not economical to handle large datasets. In most of this scenario, the experiments are done on a local dataset with a limited number of attributes for prediction of breast cancer recurrence.

Hence, there is a definite need for the predictive approach in handling breast cancer cases for both recurrence and non-recurrence for a specific period of time. The model should predict recurrence patients without a surgical biopsy which gives the physician and the patient better information for treatment.

The present work aims at the following:

- To investigate the role of Association rule learning for the successful detection of breast cancer recurrence in SEER dataset.
- To develop a machine learning model on high-performance computing platform to analyze breast cancer recurrence using MapReduce approach.
- To develop a parallel MapReduce algorithm that encourages concurrent participation of various computing hubs to develop a classifier model for breast cancer recurrence.
To investigate the Hadoop MapReduce data positioning system on data locality and develop a model to enhance the execution speedup for big data positioning to find the recurrence of breast cancer.