ABSTRACT

The past two decades have witnessed rapid advances in the area of genomics and proteomics. This has resulted in explosive growth of biological data. It is not easy to manually organize and extract useful information from such large amount of data. The necessity of automated biological data analysis methods has led to the emergence of a new field called bioinformatics. Bioinformatics is an interdisciplinary field that applies computational and statistical techniques to extract useful information from biological data. The objectives of bioinformatics are threefold. The first objective is to organize data in a way that is easy to access existing information and to update new information as it is produced. The second objective is to develop algorithms and tools that aid in the analysis of data. The third objective is to use these algorithms and tools to analyze the data and interpret the results in a biologically meaningful manner.

Data Mining is the process of discovering useful information from large amount of data. Parallel data mining aims at developing parallel algorithms, methods, and tools for the extraction of useful patterns from massive data thereby decreasing the processing time. Biological data like sequence data, gene expression data, microarray images are voluminous and unstructured. Processing such data is time consuming. An active area of research in bioinformatics is the application and development of parallel data mining techniques to effectively extract useful information from biological data which helps in drug discovery, personal medicine, crop improvement, insect resistance and so on.

Biological data available in the bioinformatics databases are raw and unstructured. Hence it should be cleaned and transformed into structured data to mine useful information efficiently. The process of converting unstructured data into structured data is called data pre-processing. Steps involved in data pre-processing are data cleaning, integration, transformation and discretization. The pre-processed data is then analyzed to extract useful and meaningful information. Association analysis, classification, clustering, prediction are some of the data mining techniques used in data analysis. There arises the need for pre-processing and analyzing the biological data efficiently.

This research proposes feature extraction and feature selection for sequence data using MapReduce programming model. It also proposes clustering and classification algorithms for mining sequence data and gene expression data. DNA sequences are made up of four nucleotides A, C, G and T. This research extracts triplets from DNA sequences using MapReduce programming model. Similarly protein sequences are made up of twenty amino acids. N-gram approach is commonly used for feature representation. As the number of N-grams required to represent a protein sequence is large, feature selection is used to select the most important features. This research selects features from protein sequences using Apriori property and Mutual Correlation based Feature Selection (AMFS).

Clustering of biological data is an unsupervised process of grouping similar data into clusters. Partition based clustering is the most widely used clustering approach. This research proposes MapReduce based parallel clustering approach using K-means approach hybridized with Differential Evolution (DE) and Ant Colony Optimization (ACO). Classification of biological data is a supervised data mining task that predicts the class label of the unknown sample from the available data. In order to improve the accuracy of classification this research proposes MapReduce based parallel classification using K-nearest neighbour (K-NN) approach hybridized with Particle Swarm Optimization (PSO).

The effectiveness of clustering methods is measured using various metrics like cluster compactness, cluster separation and overall cluster quality. The effectiveness of classification methods is measured using various metrics like precision, recall and accuracy. The performance of the proposed techniques are tested with five datasets. It includes DNA sequences dataset of Escherichia coli, Homo sapiens, Mus musculus, Rattus norvegicus, Saccharomyces cerevisiae and Pneumocystis carinii, BRENDA enzyme dataset, SCOP1.79 protein family dataset, breast cancer gene expression dataset (GSE26304) and rice plant disease gene expression dataset (GSE16793).

Experimental results show that AMFS based feature selection reduces the number of features of protein sequences by 33%. The overall cluster quality of the proposed K-means clustering with DE and ACO is improved by 0.3064 compared to the existing K-means clustering and by 0.1003 compared to K-means clustering with DE. Hence that parallel K-means clustering with DE and ACO generates high quality clusters compared to simple parallel K-means clustering and parallel K-means clustering with DE. The proposed classification using K-NN with PSO has accuracy of 92.83% and yields 4.85% increase in the accuracy compared to the existing simple K-NN classifier. It shows that the parallel K-NN with PSO performs well, compared to simple parallel K-NN. For processing 1.4GB of data on 8 processors, the average speedup obtained is 7.17. This shows that, the parallel implementation of the data mining tasks using MapReduce programming paradigm exhibits good scalability.