ABSTRACT

Recently, there has been a common augment in the amount of data publicly obtainable in wide-reaching manner predominantly in the field of Bioinformatics, where massive amounts of data have been collected in the form of Deoxyribo Nucleic Acid (DNA) sequences, protein sequences and structures, information on biological pathways, etc. With the above mentioned biological data, clustering or grouping of data typically carried out by labor-intensive experimental work or in a semi-automatic manner by making use of sequence homology.

Clustering is an unsupervised learning method and it is used in the exploration of similarities among a collection of patterns, by organizing them into homogenous clusters. It is expected that, the state of the art of clustering algorithms will help the interested researchers to put forward in proposing more robust against accuracy and computation time. The proposed work consists of three major research contributions. First Robust Ensemble Co-Clustering Algorithm (RECCA) is proposed for clustering of data. Second Improved Ant Colony Optimization with Robust Ensemble Co-Clustering Algorithm (IACO-RECCA) is proposed for clustering the data. Finally the combination of Improved Cuckoo Search and Robust Ensemble Co-Clustering Algorithm (ICS - RECCA) is proposed for clustering.

The first stage of the work aims to propose an Enhanced Principal Component Analysis (EPCA) based preprocessing technique for solving missing value problem of data clustering. A robust ensemble mechanism is proposed in this stage. It deals with the EPCA where absolute expression value is related to the particular topic. At last, in enhanced PCA, a correlation coefficient is used to provide higher weights to the significant observed data. The Pearson's correlation coefficient of the ranked data is thus obtained. Then the objective function for the co-clustering ensemble towards application to

enzyme clustering is presented. The statistical information shared between two co-clustering is measured by using the mutual information as an important parameter in this method. In bipartite graph, the last ensemble step can be used as a separation problem. The proposed algorithm is capable enough to perform co-clustering with the objective function as the primary component. The final consensus co-clustering result is obtained by using the classical k-means algorithm. However, RECCA requires more time for searching best co-clusters so, to increase the searching speed, in the next stage IACO is introduced for data clustering.

The second stage of the work intends to propose a system with IACO based on enhanced preprocessing method for data clustering. The first stage in an IACO algorithm starts with designing a problem search space in which the ants conduct the search in order to find the candidate solutions. The search space for IACO- RECCA is defined with the help of input datasets. The proposed IACO-RECCA algorithm is capable enough to perform co-clustering with an objective function as the primary component. Simulation results proved that the proposed mechanism IACO-RECCA performs better in terms of accuracy and computation time.

The final stage of the work intends to propose a system with ICS - RECCA for data clustering. The Cuckoo Search Algorithm (CSA) has been inspired by the obligate brood parasitism of some cuckoo species by laying their eggs in the nests of other host birds. ICS provides better convergence rates compared to the standard CSA, while applying to standard optimization benchmark objective functions. Next, quantum-inspired cuckoo search is proposed by adding some more improvement in CSA and provides good results. Simulation results proved that the proposed mechanism ICS-RECCA performs better in terms of accuracy and computation time.