Gene Clustering Using Metaheuristic Optimization Algorithms

P. K. Nizar Banu, Department of Computer Applications, B.S. Abdur Rahman University, Chennai, India

S. Andrews, Department of Information Technology, Mahendra Engineering College, Mallasamudram, India

ABSTRACT

Gene clustering is a familiar step in the exploratory analysis of high dimensional biological data. It is the process of grouping genes of similar patterns in the same cluster and aims at analyzing the functions of gene that leads to the development of drugs and early diagnosis of diseases. In the recent years, much research has been proposed using nature inspired meta-heuristic algorithms. Cuckoo Search is one such optimization algorithm inspired from nature by breeding strategy of parasitic bird, the cuckoo. This paper proposes cuckoo search clustering and clustering using levy flight cuckoo search for grouping brain tumor gene expression dataset. A comparative study is made with genetic algorithm, PSO clustering, cuckoo search clustering and clustering using levy flight cuckoo search. Levy flight is an important property of levy distribution which covers the entire search space. Breeding pattern of cuckoo is associated with the genes that cause tumor to grow and affect other organs gradually. Clusters generated by these algorithms are validated to find the closeness among the genes in a cluster and separation of genes between clusters. Experimental results carried out in this paper show that cuckoo search clustering outperforms other clustering methods used for experimentation.

Keywords: Brain Tumor, Cuckoo Search, Gene Clustering, Gene Expression Data, Genetic Algorithm, Levy Flight, PSO

INTRODUCTION

DNA microarray technology is a fundamental tool in the study of gene expression data analysis. The accumulation of datasets from this technology that measures the relative abundance of mRNA of thousands of genes across tens or hundreds of samples has underscored the need for quantitative analytical tools to examine such data. Due to the large number of genes and complex gene regulation networks, clustering is a useful exploratory technique for analyzing these data. Clustering divides the data of interest into a small number of relatively homogeneous groups or clusters. There are two ways of applying cluster analysis to microarray data. One way is to cluster genes according to their expression patterns across different conditions. The other way is to cluster samples from different tissues, cells at different time points of a biological process or...
under different treatments (Chen et al., 2002). Gene expression profiles can be built by measuring transcription levels of genes in an organism under various conditions, at different developmental stages and in different tissues that characterizes the dynamic functioning of each gene in genome (Alvis & Vilo, 2000). These gene expression data in microarray are presented in $M \times N$ matrix where $M$ is the number of microarray experiments and $N$ being the number of genes (Tuzhilin & Adomavicius, 2002). Certain analysis needs to be performed on this gene expression data to retrieve useful biological information. Cluster analysis is one such technique which discovers useful biological information by detecting genes that have identical expression profile (Kotala et al., 2001). A wide variety of clustering algorithms are available for clustering gene expression data (Bezdek, 1981). Researchers introduced a number of clustering algorithms, based on the characteristics of the clustering procedure; clustering algorithms are classified into two broad categories namely partitional and hierarchical clustering. Grid-based clustering (Liao et al., 2004), projection based clustering (Bouguessa & Wang, 2009), subspace clustering (Agrawal et al., 1998), density based clustering (Ester et al., 1996), model based methods, graph theoretic methods and soft computing methods are the other clustering algorithms that are presented in the literature.

In the recent years, optimization algorithms are also introduced for clustering process. In optimization based clustering, minimum sum of squared error is considered as the objective and the researchers have used optimization procedure defined in their algorithm for solving clustering objective (Binu et al., 2013). Based on the similar procedure, Genetic Algorithm (Mualik & Bandyopadhyay, 2002), Particle Swarm Optimization (Premalatha & Natrajan, 2008), bacterial foraging optimization (Wan et al., 2012), simulated annealing (Selim & Alsultan, 1991), artificial bee colony (Zhang et al., 2010; Karaboga & Ozturk, 2011), flyfire algorithm (Senthilnath et al., 2011) and cuckoo search (Goel et al., 2011) algorithms were applied for clustering. This paper focuses on the application of cuckoo search based clustering for brain tumor gene expression dataset.

Brain tumor is highly fatal because, it cannot be diagnosed easily until it reaches its final stage. This may also cause other organs of the body to dysfunction. The breeding pattern of parasitic cuckoos inspired the detection of rapid growth of tumor and for the identification of major organs that may be affected due to brain tumor. In the recent research carried out by (Senthilnath et al., 2013), it is proved that heavy-tail property plays an important role and works efficiently in cuckoo search clustering. We have also tested it for brain tumor dataset.

In this paper, cuckoo search clustering proposed by (Goel et al., 2011), cuckoo search via levy flight introduced by (Yang & Deb, 2009), levy flight cuckoo search clustering, PSO clustering and genetic algorithm presented in (Senthilnath et al., 2013) is applied for brain tumor gene expression dataset. A comparative study is carried out by taking different objective functions that finds minimum sum of squared error for every clustering algorithm and it is observed that objective function plays a vital role in swarm clustering.

This paper is organized as follows. Next section discusses gene clustering background followed by the motivation towards this research. A brief introduction to GA, PSO and cuckoo based clustering algorithms is presented and methodology adopted in this paper is discussed. Results are presented and discussed and finally this paper is concluded.

**GENE CLUSTERING BACKGROUND**

Clustering is a fundamental technique in exploratory data analysis and pattern discovery, aiming at extracting and creating groups where there is no prior information on underlying cluster structures. Clustering of microarray gene expression data helps to understand the gene functions,
Exploring Type-and-Identity-Based Proxy Re-Encryption Scheme to Securely Manage Personal Health Records
Luan Ibraimi, Qiang Tang, Pieter Hartel and Willem Jonker (2012). Innovations in Data Methodologies and Computational Algorithms for Medical Applications (pp. 73-93).
www.igi-global.com/chapter/exploring-type-identity-based-proxy/65153?camid=4v1a

Gene Clustering Using Metaheuristic Optimization Algorithms
www.igi-global.com/article/gene-clustering-using-metaheuristic-optimization-algorithms/132522?camid=4v1a