Engineering Applications of Artificial Intelligence 25 (2012) 376–391

Contents lists available at SciVerse ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Gene expression data analysis with the clustering method based on an improved quantum-behaved Particle Swarm Optimization

Jun Sun*, Wei Chen, Wei Fang, Xiaojun Wun, Wenbo Xu

Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), Department of Computer Science and Technology, Jiangnan University, No. 1800, Lihu Avenue, Wuxi, Jiangsu 214122, PR China

ARTICLE INFO

Article history:
Received 22 August 2010
Received in revised form
24 August 2011
Accepted 18 September 2011
Available online 5 October 2011

Keywords:
Gene expression data
Clustering
Particle Swarm Optimization (PSO)
Quantum-behaved Particle Swarm Optimization (QPSO)

ABSTRACT

Microarray technology has been widely applied in study of measuring gene expression levels for thousands of genes simultaneously. In this technology, gene cluster analysis is useful for discovering the function of gene because co-expressed genes are likely to share the same biological function. Many clustering algorithms have been used in the field of gene clustering. This paper proposes a new scheme for clustering gene expression datasets based on a modified version of Quantum-behaved Particle Swarm Optimization (QPSO) algorithm, known as the Multi-Elitist QPSO (MEQPSO) model. The proposed clustering method also employs a one-step K-means operator to effectively accelerate the convergence speed of the algorithm. The MEQPSO algorithm is tested and compared with some other recently proposed PSO and QPSO variants on a suite of benchmark functions. Based on the computer simulations, some empirical guidelines have been provided for selecting the suitable parameters of MEQPSO clustering. The performance of MEQPSO clustering algorithm has been extensively compared with several optimization-based algorithms and classical clustering algorithms over several artificial and real gene expression datasets. Our results indicate that MEQPSO clustering algorithm is a promising technique and can be widely used for gene clustering.

1. Introduction

In an attempt to understand complicated biological systems, many researchers have generated large amounts of gene expression data under complex conditions of experiments so that clustering has been effectively applied in molecular biology for gene expression data analysis (Shamir and Sharan, 2001). Total of genes in a dataset are assigned into different clusters of similar expression patterns according to a dissimilarity measure (usually correlation-based or distance-based) between any two genes by clustering algorithms. The goal of the clustering process is thus to identify the genes with the same functions or the same regulatory mechanisms.

Many clustering algorithms have been proposed for gene expression data analysis. The hierarchical clustering is one of the earlier methods applied to clustering of gene expression data. Eisen et al. (1998) used a variant of the hierarchical average-link clustering algorithms to identify groups of co-regulated yeast genes. However, the hierarchy of the algorithm is greatly affected by the minor change of the given data, which makes the clustering show the lack of robustness and nonuniqueness. K-means is one of another popular methods used in gene expression data analysis due to its high computational performances (Tavazoie et al., 1999). But it might converge to a local optimum, and its results is quite subject to the random initialization process, which means that different runs of K-means on the same data set might produce different clusters (Kao et al., 2008). As one kind of neural network, self-organizing map (SOM) which presents high-dimensional data by the low dimensional data has also been used for gene expression data clustering (Tamayo et al., 1999). However it always produces an unbalanced solution and it is difficult to find clear clustering boundaries from results of the SOM. Other common clustering methods include CAST algorithm (Ben-Dor and Yakini, 1999), model-based clustering (Yeung et al., 2001a) and tight clustering (Tseng and Wong, 2005). Especially, the last two algorithms were proposed to allow a noise set of genes (or so-called scattered genes) without being clustered. It is in view of the fact that very often a significant number of genes in an expression profile do not play any role in the disease or perturbed conditions under investigation.

Recently, some novel methods have been proposed for gene clustering technology. Qin (2006) devised an improved model-based Bayesian approach, known as the weighted Chinese restaurant process (CRP), to cluster microarray gene expression data. Cluster assignment of CRP is carried out by an iterative weighted Chinese restaurant seating scheme such that the optimal number of clusters can be determined simultaneously with cluster assignment. Affinity Propagation (AP), a new powerful algorithm based

* Corresponding author. Tel./fax: +86 510 85916500.
E-mail address: sunjun_wx@hotmail.com (J. Sun).

0952-1976/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.engappai.2011.09.017

on message-passing techniques, was proposed by Frey and Dueck (2007). In AP, each cluster is identified by a common exemplar that all other data points of the same cluster refer to, and the exemplars have to refer to themselves. Leone (2007) improved the original AP algorithm by relaxing its hard constraints and the resulting soft-constraint affinity propagation (SCAP) became more informative, accurate and led to more stable clustering.

Many researchers have employed genetic algorithms (GA) for clustering (Hall et al., 1999). The fundamental strategy of such clustering approaches is to imitate the evolution process of nature and evolve the solutions of clustering from one generation to the next. In contrast to K-means algorithm, clustering algorithms based on GA are insensitive to the initialization process and always converge to the global optimum eventually. However, these algorithms are usually computationally expensive, which impedes the wide application of them in the field of gene expression data analysis. To overcome this limitation and accelerate the convergence, some hybrid methods have been proposed. For an instance, Krishna and Murty (1999) proposed a new clustering method called Genetic K-means Algorithm (GKA), which hybridizes a genetic algorithm with the K-means algorithm. This hybrid approach combines the robust nature of the genetic algorithm with the high performance of the K-means algorithm. Based on the GKA, Lu et al. (2003, 2004) proposed Fast Genetic K-means Algorithm (FGKA) and Incremental Genetic K-means Algorithm (IGKA) for analyzing gene expression data. Moreover, for up to research of GA clustering in gene expression data analysis, Bandyopadhyay et al. (2007) devised a two-stage clustering algorithm, which employs a recently proposed variable string length genetic scheme and a multiobjective genetic clustering algorithm. It is based on the novel concept of points having significant membership to multiple classes. An iterated version of the well-known Fuzzy C-means is also utilized for clustering.

Particle Swarm Optimization (PSO) (Kennedy and Eberhart, 1995), a population-based random search technique motivated by the behavior of organisms such as fishing schooling and bird flock, has been applied to data clustering (Van der Merwe and Engelbrecht, 2003; Tzay-Farn 2006; Yang et al., 2009; Du et al., 2008). As a method of swarm intelligence, PSO is easier to implement for clustering than GA, since it does not need any complex operation, such as selection, crossing and mutation in GA.

More recently, a new variant of PSO, called Quantum-behaved Particle Swarm Optimization (QPSO), has been proposed in order to improve the global search ability of the original PSO (Sun et al., 2004a, 2004b, 2005). The iterative equation of QPSO is far different from that of PSO in that it is needs no velocity vectors for particles, has fewer parameters to adjust and can be implemented more easily. It has been proved that this iterative equation leads QPSO to be global convergent (Fang et al., 2010). The QPSO algorithm has been aroused the interests of many researchers from different communities.

Although empirical studies have verified that the QPSO algorithm works better than PSO in solving a wide range of continuous optimization problems (Coelho, 2010; Omkar et al., 2009; Sabat et al., 2009; Shayeghi et al., 2010; Sun and Lu, 2010; Zhang, 2010); researchers have showed that further improvement of QPSO is possible and many efficient strategies have been proposed to improve this algorithm (Coelho, 2008; Xi et al., 2008; Huang et al., 2009). The main reason for developing these improved versions of QPSO is that like other evolutionary algorithm including PSO, this algorithm may also encounter the problem of premature convergence, particularly for the problems with high dimensionality and multiple local optima. Bearing this in mind, we have always been devoting ourselves to enhance the performance of the QPSO algorithm. Thus, in this paper, we proposed a novel variant of QPSO, called Muti-Elitist Quantum-behaved Particle Swarm Optimization (MEQPSO), in which a Multi-Elitist strategy for searching the global best position is employed to enhance the global search ability of the QPSO algorithm so as to avoid premature convergence efficiently. In the original QPSO, the particle is guided by the global best position as well as its personal best position. If the global best position traps into a local optimum, all the particles will be pulled toward this local optimal point and will have higher possibility to failing in search of the better region where the global optimal solution may be located, which can lead to premature convergence of the algorithm. On the other hand, in the proposed MEQPSO, the particle's search is influenced by the position, which may not be the global position but may lie in a promising search region, so that the particles have much chance to search this region and find out the global optimal solution. As a result, MEQPSO may have better overall performance than the original QPSO, particularly for the hard optimization problems.

In this paper, we also show how to apply the MEQPSO algorithm to gene expression data clustering, which can be reduced to an optimization problem. In our work, like other related works (Krishna and Murty, 1999; Lu et al., 2003, 2004), the task of the optimization problem is to minimize the internal dissimilarity of each cluster, i.e., Total Within-Cluster Variation (TWCV), although other objectives can be used, such as external dissimilarities among clusters. Furthermore, in order to accelerate the convergence rate of MEQPSO for clustering, we incorporate K-means clustering into the MEQPSO-based clustering algorithm as in GKA. In this hybrid clustering method, a one-step K-means algorithm, called K-means operator (KMO), is executed after the updating of each particle in MEQPSO. The K-means algorithm used consists of two phases, one of which is calculating new cluster centers according to current particle, and the other of which is reassigning each data point to the cluster with the nearest cluster center to form the new partition. The hybrid clustering method takes the advantages of the strong global search ability of MEQPSO and fast clustering convergence speed, which is verified by testing on three gene expression datasets.

The rest of this paper is organized as follows. Section 2 provides a brief review for the QPSO algorithm. Section 3 describes the proposed MEQPSO and Section 4 presents how to use MEQPSO for gene expression data clustering. Section 5 gives the experimental results of the tested PSO and QPSO variants on some well known benchmark optimization functions. Section 6 provides the results of four groups of experiments on gene expression datasets. Finally, the paper is concluded in Section 7.

2. Quantum-behaved particle swarm optimization

2.1. A brief introduction of PSO algorithm

The proposal of PSO algorithm was put forward as an optimization technique by several scientists who developed computational simulations of the movement of organisms such as flocks of birds and schools of fish. Since its origin in 1995, there have been many works done on the PSO algorithm (Shi and Eberhart, 1998; Angeline, 1998; Clerc, 1999; Suganthan, 1999; Kennedy, 2003; Krohling, 2004; Liang and Suganthan, 2005; Wu, 2011). In the PSO with m individuals, each individual is treated as a volume-less particle in the D-dimensional space, with the position vector and velocity vector of particle i at the tth iteration represented as $X_i(t)=[x_{i,1}(t),x_{i,2}(t),\ldots,x_{i,D}(t)]$ and $V_i(t)=[v_{i,1}(t),v_{i,2}(t),\ldots,v_{i,D}(t)]$. The particle moves according to the equations:

$$v_{i,j}(t+1)=w\cdot v_{i,j}(t)+c_1\cdot R_{i,j}(t)\cdot[pbest_{i,j}(t)-x_{i,j}(t)]+c_2\cdot r_{i,j}(t)\cdot[gbest_j(t)-x_{i,j}(t)], \quad (1)$$