Expert Systems with Applications 41 (2014) 6009–6016

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A hybrid approach for data clustering based on modified cohort intelligence and K-means

Ganesh Krishnasamy [a,*], Anand J. Kulkarni [b], Raveendran Paramesran [a]

[a] Department of Electrical Engineering, Faculty of Engineering, Universiti Malaya, Lembah Pantai, 50603 Kuala Lumpur, Malaysia
[b] Odette School of Business, University of Windsor, 401 Sunset Ave, Windsor, ON N9B 3P4, Canada

ARTICLE INFO

Keywords:
Clustering
Cohort intelligence
Meta-heuristic algorithm

ABSTRACT

Clustering is an important and popular technique in data mining. It partitions a set of objects in such a manner that objects in the same clusters are more similar to each another than objects in the different cluster according to certain predefined criteria. K-means is simple yet an efficient method used in data clustering. However, K-means has a tendency to converge to local optima and depends on initial value of cluster centers. In the past, many heuristic algorithms have been introduced to overcome this local optima problem. Nevertheless, these algorithms too suffer several short-comings. In this paper, we present an efficient hybrid evolutionary data clustering algorithm referred to as K-MCI, whereby, we combine K-means with modified cohort intelligence. Our proposed algorithm is tested on several standard data sets from UCI Machine Learning Repository and its performance is compared with other well-known algorithms such as K-means, K-means++, cohort intelligence (CI), modified cohort intelligence (MCI), genetic algorithm (GA), simulated annealing (SA), tabu search (TS), ant colony optimization (ACO), honey bee mating optimization (HBMO) and particle swarm optimization (PSO). The simulation results are very promising in the terms of quality of solution and convergence speed of algorithm.

1. Introduction

Clustering is an unsupervised classification technique which partitions a set of objects in such a way that objects in the same clusters are more similar to one another than the objects in different clusters according to certain predefined criterion (Jain, Murty, & Flynn, 1999; Kaufman & Rousseeuw, 2005). The term unsupervised means that grouping is establish based on the intrinsic structure of the data without any need to supply the process with training items.

Clustering has been applied across many applications, i.e., machine learning (Anaya & Boticario, 2011; Fan, Chen, & Lee, 2008), image processing (Das & Konar, 2009; Portela, Cavalcanti, & Ren, 2014; SiangTan & MatIsa, 2011; Zhao, Fan, & Liu, 2014), data mining (Carmona et al., 2012; Ci, Guizani, & Sharif, 2007), pattern recognition (Bassiou & Kotropoulos, 2011; Yuan & Kuo, 2008), bioinformatics (Bhattacharya & De, 2010; Macintyre, Bailey, Gustafsson, Haviv, & Kowalczyk, 2010; Zheng, Yoon, & Lam, 2014), construction management (Cheng & Leu, 2009), marketing (Kim & Ahn, 2008; Kuo, An, Wang, & Chung, 2006), document clustering (Jun, Park, & Jang, 2014), intrusion detection (Jun et al., 2014), healthcare (Gunes, Polat, & Sebnem, 2010; Hung, Chen, Yang, & Deng, 2013) and information retrieval (Chan, 2008; Dhanapal, 2008).

Clustering algorithms can generally be divided into two categories; hierarchical clustering and partitional clustering (Han, 2005). Hierarchical clustering groups objects into tree-like structure using bottom-up or top-down approaches. Our research however focuses on partition clustering, which decomposes the data set into a several disjoint clusters that are optimal in terms of some predefined criteria.

There many algorithms have been proposed in literature to solve the clustering problems. The K-means algorithm is the most popular and widely used algorithm in partitional clustering. Although, K-means is very fast and simple algorithm, it suffers from two major drawbacks. Firstly, the performance of K-means algorithm is highly dependent on the initial values of cluster centers. Secondly, the objective function of the K-means is non-convex and it may contain many local minima. Therefore, in the process of minimizing the objective function, the solution might easily converge to a local minimum rather than a global minimum (Selim & Ismail, 1984). K-means++ algorithm was proposed by Arthur and Vassilvitskii (2007), which introduces a cluster centers initialization procedure to tackle the initial centers sensitivity problem of

* Corresponding author. Tel.: +60 124225549.
E-mail addresses: krishnasamy.ganesh@gmail.com (G. Krishnasamy), kulk0003@uwindsor.ca (A.J. Kulkarni), ravee@um.edu.my (R. Paramesran).

http://dx.doi.org/10.1016/j.eswa.2014.03.021
0957-4174/

a standard K-means. However, it too suffers from a premature convergence to a local optimum.

In order to alleviate the local minima problem, many heuristic clustering approaches have been proposed over the years. For instance, Selim and Alsultan (1991) proposed a simulated annealing approach for solving clustering problems. A tabu search method which combines new procedures called packing and releasing was employed to avoid local optima in clustering problems (Sung & Jin, 2000). Genetic algorithm based clustering method was introduced by Maulik and Bandyopadhyay (2000) to improve the global searching capability. In Fathian, Amiri, and Maroosi (2007), a honey-bee mating optimization was applied for solving clustering problems. Shelokar, Jayaraman, and Kulkarni (2004) proposed an ant colony optimization (ACO) for clustering problems. A particle swarm optimization based approach (PSO) for clustering was introduced by Chen and Ye (2004) and Cura (2012). A hybrid technique for clustering called K-NM-PSO, which combines the K-means, Nedler–Mead simplex and PSO was proposed by Kao, Zahara, and Kao (2008). Zhang, Ouyang, and Ning (2010) proposed an artificial bee colony approach for clustering. More recently, black hole (BH) optimization algorithm (Hatamlou, 2013) was introduced to solve clustering problems. Although these heuristic algorithms address the flaws of K-means but they still suffer several drawbacks. For example, most of these heuristic algorithms are typically very slow to find optimum solution. Furthermore, these algorithms are computationally expensive for large problems.

Cohort intelligence (CI) is a novel optimization algorithm proposed recently by Kulkarni, Durugkar, and Kumar (2013). This algorithm was inspired from natural and society tendency of cohort individuals/candidates of learning from one another. The learning refers to a cohort candidate's effort to self-supervise its behavior and further adapt to the behavior of other candidate which it tends to follow. This makes every candidate to improve/evolve its own and eventually the entire cohort behavior. CI was tested with several standard problems and compared with other optimization algorithms such as sequential quadratic programming (SQP), chaos-PSO (CPSO), robust hybrid PSO (RHPSO) and linearly decreasing weight PSO (LDWPSO). CI has been proven to be computationally comparable and even better performed in terms of quality of solution and computational efficiency when compared with these algorithms. These comparisons can be found in the seminal paper on CI (Kulkarni et al., 2013). However, for clustering problems, as the number of clusters and dimensionality of data increase, CI might converge very slowly and trapped in local optima. Recently, many researchers have incorporated mutation operator into their algorithm to solve combinatorial optimizing problems. Several new variants of ACO algorithms have been proposed by introducing mutation to the traditional ACO algorithms and achieve much better performance (Lee, Su, Chuang, & Liu, 2008; Zhao, Wu, Zhao, & Quan, 2010). Stacey, Jancic, and Grundy (2003) and Zhao et al. (2010) also have integrated mutation into the standard PSO scheme, or modifications of it. In order to mitigate the short-comings of CI, we present a modified cohort intelligence (MCI) by incorporating mutation operator into CI to enlarge the searching range and avoid early convergence. Finally, to utilize the benefits of both K-means and MCI, we propose a new hybrid K-MCI algorithm for clustering. In this algorithm, K-means is applied to improve the candidates' behavior that generated by MCI at each iteration before going through the mutation process of MCI. The new proposed hybrid K-MCI is not only able to produce a better quality solutions but it also converges more quickly than other heuristic algorithms including CI and MCI. In summary, our contribution in this paper is twofold:

1. Present a modified cohort intelligence (MCI).
2. Present a new hybrid K-MCI algorithm for data clustering.

This paper is organized as follows: Section 2 contains the description of the clustering problem and K-means algorithm. In Sections 3 and 4, the details of cohort intelligence and the modified cohort intelligence are explained. In Section 5, we discussed the hybrid K-MCI algorithm and its application to clustering problems. Section 6 presents the experimental results that prove our proposed method outperforms other methods. Finally, we conclude and summarize the paper in Section 7.

2. The clustering problem and K-means algorithm

Let $R = [Y_1, Y_2, \ldots, Y_N]$, where $Y_i \in \Re^D$, be a set of N data objects to be clustered and $S = [X_1, X_2, \ldots, X_K]$ be a set of K clusters. In clustering, each data in set R will be allocated in one of the K clusters in such a way that it will minimize the objective function. The objective function, intra-cluster variance is defined as the sum of squared Euclidean distance between each object Y_i and the center of the cluster X_j which it belongs. This objective function is given by:

$$F(X,Y) = \sum_{i=1}^{N} Min\left\{\|Y_i - X_j\|^2\right\}, \quad j = 1,2,\ldots,K \tag{1}$$

Also,

- $X_j \neq \emptyset, \ \forall j\{1,2,\ldots,K\}$
- $X_i \cap X_j = \emptyset, \ \forall i \neq j$ and $i,j \in \{1,2,\ldots,K\}$
- $\cup_{j=1}^{K} X_j = R$

In partitional clustering, the main goal of K-means algorithm is to determine centers of K clusters. In this research, we assume that the number of clusters K is known prior to solving the clustering problem. The following are the main steps of K-means algorithm:

- Randomly choose K cluster centers of $X_1, X_2, \ldots, X_K$ from data set $R = [Y_1, Y_2, \ldots, Y_N]$ as the initial centers.
- Assign each object in set R to the closest centers.
- When all objects have been assigned, recalculate the positions of the K centers.
- Repeat Steps 2 and 3 until a termination criterion is met (the maximum number of iterations reached or the means are fixed).

Arthur and Vassilvitskii (2007) introduced a specific way of choosing the initial centers for K-means algorithm. The procedure of K-means++ algorithm is outlined below:

- Choose one center X_1, uniformly at random from R.
- For each data point Y_i, compute $D(Y_i)$, the distance between Y_i and the nearest center that has already been chosen.
- Take new center X_j, choosing $Y \in R$ with probability $\frac{D(Y)^2}{\sum_{Y \in R} D(Y)^2}$.
- Repeat Steps 2 and 3 until K centers have been chosen.
- Now that the initial centers have been chosen, proceed using standard K-means clustering.

3. Cohort intelligence

Cohort intelligence (CI) is a new emerging optimization algorithm, which is inspired from natural and society tendency of cohort candidates of learning from one another. The term cohort refers to a group of candidates competing and interacting with one another to achieve some individual goal which is inherently common to all the candidates. Each candidate tries to improve its own behavior by observing every other candidates in a cohort. Every candidate might follow certain behavior in the cohort which