ACADEMIC
World Academic Union

ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 9, No. 3, 2014, *pp*.196-209

A Hybrid Data Clustering Approach Based on Cat Swarm Optimization and K- Harmonic Mean Algorithm

Yugal Kumar and G. Sahoo
Department of Information Technology, Birla Institute of Technology, Mesra, Ranchi, Jharkhand, India.
yugalkumar@bitmesra.ac.in
Department of Information Technology, Birla Institute of Technology, Mesra, Ranchi, Jharkhand, India.
gsahoo@bitmesra.ac.in
(*Received November 5, 2013, accepted May 21, 2014*)

Abstract. Clustering is an important task that is used to find subsets of similar objects from a set of objects such that the objects in the same subsets are more similar than other subsets. Large number of algorithms has been developed to solve the clustering problem. K-Harmonic Mean (KHM) is one of the popular technique that has been applied in clustering as a substitute of K-Means algorithm because it is insensitive to initialization issues due to built in boosting function. But, this method is also trapped in local optima. On the other hand, Cat Swarm Optimization (CSO) is the latest population based optimization method used for global optimization. In this paper a hybrid data clustering method is proposed based on CSO and KHM which includes the advantage of both algorithms and named as CSOKHM. The hybrid CSOKHM not only improved the convergence speed of CSO but also escape the KHM method to run in local optima. The performance of the CSOKHM is evaluated using seven datasets and compared with KHM, PSO, PSOKHM, ACA, ACAKHM, GSAKHM, CSO methods. The experimental results show the applicability of CSOKHM method..

Keywords: Cat swarm optimization, Data clustering, K-harmonic means, Gravitational search algorithm, Particle swarm optimization

1. Introduction

Clustering is an essential tool in pattern recognition, data mining and machine learning domain. It is NP Complete problem to find out hidden patterns, knowledge and information from a dataset that is previously unknown using some criterion function [1]. In clustering, a dataset is divided into K number of groups. The elements in one group are more similar to another group. K-Means (KM) algorithm is the oldest algorithm that has been widely used in clustering domain to find optimal cluster centers in datasets [2]. This algorithm is simple, fast and efficient but suffered with initialization and local optima problem [3]. Hence, to overcome the problems of KM algorithm and improve the efficiency of the clustering, hybridize version of KM algorithms have been developed by various researchers [4]. Instead the hybridization of KM, Zhang et al. [5] has developed K-Harmonic Means (KHM) algorithm for data clustering. In KHM, the clustering objective is to minimize average harmonic means to all instances of dataset in lieu of average mean (KM) from all cluster centers. The KHM algorithm has provided better result in comparison of KM but this algorithm is also suffered with stuck in local optima problem. In recent years, numbers of algorithms based on swarms, insects and natural phenomena's have been developed by researchers to solve clustering problem such as ABC [6], ACO [7], GA [8], PSO [9], CSO [10], BH [11], GSA [12] and many more. These algorithms are categorized as swarm based algorithms, biological based algorithms and basic science based algorithms. The above mentioned algorithms have immense potentials over prevailing traditional methods but these methods have suffered with several problems, for instance GA suffers from population diversity problem and the quality of solutions in GA depends on mutation and crossover probability [13]. The convergence time of ACO method is uncertain and probability distribution function change in each iteration [14]. PSO algorithm has weak exploitation property and sometimes stuck in local optima [15]. The performance of ABC algorithm is depended on the dimension of problem as dimension of problem is increased the convergence speed of ABC is decreased [16]. The GSA algorithm is sometimes suffered with premature convergence and there is no recovery if premature convergence exists because GSA is memory less algorithm [17].

The cat swarm optimization (CSO) is the latest, state of art animal inspired algorithm developed by Chu et al. [18], observing the behavior of cats. CSO is the first algorithm based on the behavior of animals as reported in the literature. The animal inspired algorithms are the sub branch of swarm based algorithms. CSO algorithm has been applied in many areas and provides remarkable results [19, 20, 21]. The main advantage of the CSO algorithm is good exploration property. Hence in this paper, a hybrid data clustering algorithm is proposed based on the CSO and KHM, to escapes the KHM run in local optima problem and increases the convergence speed of CSO. The performance of proposed algorithm is tested on several benchmark datasets which are downloaded from UCI repository and the proposed algorithm is more accurate and precise than others. The rest of the paper is organized as follow. Section 2 introduces KHM algorithm. Section 3 describes CSO technique. Section 4 presents hybrid CSOKHM clustering algorithm. Section 5 illustrates investigational results. Finally, section 6 gives conclusions.

2. K Harmonic Algorithm

The KHM is partition based iterative algorithm that evaluates the cluster using K centers. KMH is unconcerned to the initialization issues and provides faster convergence than KM when the initial cluster points far from local optimal. In case of KM, quality of solution depends on the initial cluster centers [22]. In KHM, the distance among instances of dataset to cluster centers are calculated by harmonic means. The performance of KHM algorithm is evaluated using the following equation.

$$KHM(X,C) = \sum_{i=1}^{N} \frac{k}{\sum_{j=1}^{k} \frac{1}{\|x_i - c_j\|^p}} \tag{1}$$

$X = \{x_1, x_2, x_3, \ldots\ldots, x_n\}$: Data instance for clustering.

$C = \{c_1, c_2, \ldots\ldots, c_k\}$: Number of clusters.

$m(c_j/x_i)$: membership function to define the data point x_i belongs with center c_j.

$w(x_i)$: weight function to measure the influnce of data instances x_i to recompute the cluster centers.

The steps of KHM clustering algorithm can be given as

- Randomly initialize the cluster centers.
- Evaluate the value of objective function using equation 1.
- For each data instance x_i ,
  - Calculate its membership function $m(c_j/x_i)$ from all cluster centers using given equation

$$m(c_j/x_i) = \frac{\|x_i - c_j\|^{-p-2}}{\sum_{j=1}^{k} \|x_i - c_j\|^{-p-2}} \tag{2}$$

  - Calculate the weight $w(x_i)$ using following equation

$$w(x_i) = \frac{\sum_{j=1}^{k} \|x_i - c_j\|^{-p-2}}{\left(\sum_{j=1}^{k} \|x_i - c_j\|^{-p}\right)^2} \tag{3}$$