Applied Soft Computing 13 (2013) 2643–2657

Contents lists available at SciVerse ScienceDirect

Applied Soft Computing

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

A novel ant-based clustering algorithm using Renyi entropy

Lei Zhang[a,*], Qixin Cao[a], Jay Lee[b]

[a] Research Institute of Robotics, Shanghai Jiao Tong University, Shanghai 200240, China
[b] NSF Center for Intelligent Maintenance Systems, University of Cincinnati, OH 45221,USA

ARTICLE INFO

Article history:
Received 27 May 2011
Received in revised form 31 October 2012
Accepted 22 November 2012
Available online 6 December 2012

Keywords:
Swarm intelligence
Ant-based clustering
Renyi entropy
Kernel
The Friedman test

ABSTRACT

Ant-based clustering is a type of clustering algorithm that imitates the behavior of ants. To improve the efficiency, increase the adaptability to non-Gaussian datasets and simplify the parameters of the algorithm, a novel ant-based clustering algorithm using Renyi Entropy (NAC-RE) is proposed. There are two aspects to application of Renyi entropy. Firstly, Kernel Entropy Component Analysis (KECA) is applied to modify the random projection of objects when the algorithm is run initially. This projection can create rough clusters and improve the algorithm's efficiency. Secondly, a novel ant movement model governed by Renyi entropy is proposed. The model takes each object as an ant. When the object (ant) moves to a new region, the Renyi entropy in its local neighborhood will be changed. The differential value of entropy governs whether the object should move or be moveless. The new model avoids complex parameters that have influence on the clustering results. The theoretical analysis has been conducted by kernel method to show that Renyi entropy metric is feasible and superior to distance metric. The novel algorithm was compared with other classic ones by several well-known benchmark datasets. The Friedman test with the corresponding Nemenyi test are applied to compare and conclude the algorithms' performance The results indicate that NAC-RE can get better results for non-linearly separable datasets while its parameters are simple.

1. Introduction

Swarm intelligence is one kind of intelligent behavior shown by the cooperation of collective insects, such as ants and bees. Since 1990, several collective behavior inspired algorithms have been proposed. Particle Swarm Optimization (PSO) [1,2] and Ant Colony Optimization (ACO) [3,4] are the most popular in this domain. Recently, prey models [5] also show an increase popularity. PSO is designed to simulate the choreography of bird flocking. The birds are represented by a population of particles and each particle has a certain location and velocity within the search space [1]. Particles fly through the search space in search of high quality solutions. ACO is inspired by the behavior of ant colonies for food searching. The pheromone trails between the ants enables them to find the shortest path between their nest and food source [3,4]. In the prey model, a forager needs to decide whether to attack the prey or to continue searching. The foraging agent should maximize the energy intake with respect to the probability to attack [5].

The application areas of these algorithms include NP hard optimization problems (such as the traveling salesman problem), the quadratic assignment, the network routing, clustering and job scheduling. The general review of swarm intelligence in data mining such as rule induction, classification and clustering, can refer to [6,7]. In this paper, we mainly focus the algorithms to imitate the ants' behavior for clustering purpose.

Clustering is a method that divides a dataset into groups of similar objects, thereby minimizing the similarities between different clusters and maximizing the similarities between objects in the same cluster. Clustering is widely applied in data mining, such as in document clustering and Web analysis. Classic clustering approaches include partition-based methods, such as K-means, K-medoids, and K-prototypes [8,9]; hierarchy-based methods, such as BIRCH [10]; density-based methods such as LDBSCAN [11,12]; grid-based methods such as GGCA [13]; and model-based methods, such as neural networks and Self-Organizing Map (SOM) [14,15].

Recently, ant-based clustering, which is a type of clustering algorithm that imitates the behavior of ants, has earned researchers' attention. Ant-based clustering can be divided into two classes. The first class imitates the ant's foraging behavior, which involves finding the shortest route between a food source and the nest. This intelligent behavior is achieved by means of pheromone trails and information exchange between ants [16,17]. The algorithms treat clustering as an optimization task and utilize ACO methods to obtain optimal clusters. A variant of ACO, called the Aggregation Pheromone density-based Clustering algorithm (APC), was also suggested [18]. Similar to ACO, APC is based on the aggregation pheromones found in ants. The advantage of these methods is that the objective function is explicit. The key elements of these

* Corresponding author.
E-mail addresses: zhanglei@sjtu.edu.cn, zhanglei75@sina.com (L. Zhang).

1568-4946/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.asoc.2012.11.022

algorithms are the pheromone matrix updating rule and the heuristic function.

The second class imitates ants' behavior of clustering their corpses and forming cemeteries. Some ants can pick up dead bodies randomly distributed in the nest and group them into different sizes. The large group of bodies attracts the ants to deposit more dead bodies and becomes larger and larger. The essence of this phenomenon is positive feedback [19]. One of the first studies related to this domain is the work of Deneubourg [20], who came up with the Basic Model (BM) to explain the ants' movement. In the BM, the ants move randomly and pick up or drop objects according to the number of similar surrounding objects to cluster them. Lumer and Faieta [21] extended the model and applied it to data analysis (they called this the LF algorithm). In their analysis, an object with n attributes can be viewed as a point in the R^n space. The point is projected into a low-dimensional space (often a two-dimensional plane). The similarity of the object with those in the local neighborhood is calculated to determine whether the object should be picked up or dropped by ants. As a basic algorithm, LF was followed and improved by a number of modified algorithms in different applications. Wu et al. [22] further explained the idea of the similarity coefficient (this coefficient defines the scale for objects' similarity) and suggested a more simple probability conversion function. Ramos and Merelo [23] studied ant-based clustering with different ant speeds to cluster text documents. Yang et al. [24,25] suggested multiple ant colonies consisting of independent colonies and a queen ant agent. Each ant colony had a different moving speed and probability conversion function. The hypergraph model was used to combine the results of all parallel ant colonies.

In addition to the above-mentioned studies, a series of research by Handl deserves special attention. She came up with a set of strategies for increasing the robustness of the LF algorithm and applying it to document retrieval [26]. She performed a comparative study of ant-based clustering with K-means, average links, and 1d-SOM [27,28]. An improved version, ATTA, which incorporates adaptive and heterogeneous ants and time-dependent transporting activity, was proposed in her latest paper [29]. The main feature of this kind of algorithm is that the algorithm directly imitates the ant's behavior to cluster data and the clustering objective is implicitly defined [30].

Beyond these two classes of ant-based clustering, Tsang and Kwong [31] proposed Ant Colony Clustering for anomaly intrusion detection. This method integrates the characteristics of the two above-mentioned classes. Specifically, cluster formation and searching for an object are regarded as nest building and food foraging, respectively. The ant exhibits picking up and dropping behaviors while simultaneously depositing cluster-pheromones on the grid. Xu et al. [32] suggested a novel ant movement model wherein each object was viewed as an ant. The ant determines its behavior according to the fitness of its local neighborhood. Essentially, this model is similar to that in the second class of ant-based clustering.

Combinations of ant-based clustering with other clustering methods can also be found. For example, ant-based clustering has been combined with K-means [33] and with K-harmonic means [34]; ant colonies have been hybridized with fuzzy C-means [35]; fuzzy ants have been endowed with intelligence in the form of IF-THEN rules [36]; and the hybrid approach has been generated based on Particle Swarm Optimization (PSO), ACO, and K-means [37]. In these methods, the role of ant-based clustering is mainly to create initial clusters for other clustering algorithms.

A comprehensive overview of ant-based and swarm-based clustering can be found in [30]. Our particular interest is in the second kind of ant-based clustering discussed above. The process of this kind of algorithms can be generalized as five steps (detailed description is in Section 2):

(1) *Projection*: All objects and ants are randomly projected onto the toroidal grid.
(2) *Calculating the similarity*: Each ant calculates the object's similarity to others in the object's local neighborhood.
(3) Picking up or dropping objects.
(4) Ants move.
(5) Repeat (2)–(4).

Although this kind of ant-based clustering has been modified gradually, there are still some problems needed to be solved. The focus of our work is on the following three important problems.

- Improving the algorithm's efficiency
  It is not highly efficient because of the randomness in the algorithm. Because the objects are randomly projected onto the toroidal grid at the initial time of the algorithm, the similarities of the objects in a local neighborhood are very low. Therefore, the objects are easily picked up but not easily dropped by the ants. It takes a long time to go from the inception of the algorithm to the moment when the rough clusters are created. Commonly, tens of thousands of iterations are needed for ant-based clustering algorithms [17,29,39].
- Improving the adaptability of the algorithm to the datasets with special structures
  In the essence, ant-based clustering algorithms are distance-based because the similarity of the objects is computed by Euclidean distance or Cosine distance. Just like other distance-based clustering algorithms, it is effective for the datasets with ellipsoidal or Gaussian structure. If the separation boundaries between clusters are nonlinear, it will fail [38].
- Simplifying the parameters in the algorithm
  There are several parameters in ant-based clustering, such as the similarity coefficient, the constants in the probability conversion functions (which will be described in Section 2). Some parameters are difficult to set properly, while they have an important effect on the clustering results. For example, a too small choice of the similarity coefficient α prevents the formation of clusters; on the other hand, a too large choice results in the fusion of individual clusters [22,26–29]. As mentioned in [39], the complex parameter setting should be avoided to simplify the use of the algorithm.

To solve these problems, a novel ant-based clustering algorithm integrated with Renyi entropy (NAC-RE) is proposed. The applications of Renyi entropy in NAC-RE are shown in two aspects. First, Kernel Entropy Component Analysis (KECA) is used to modify the initial projection of all objects. Second, a novel ant movement model governed by Renyi entropy is created. These two applications are geared toward solving the problems mentioned above.

Various attempts have been made to utilize information theory in clustering [40,41]. Tsang and Kwong [31] first introduced the application of local regional entropy in ant-based clustering. Liu et al. [38] proposed entropy-based metrics in ant-based clustering. Entropy governs the ant's picking up and dropping behaviors. They pointed out that entropy-based ant clustering required fewer training parameters than density-based. However, they used traditional Shannon entropy. First, the attributes of the objects must be independent. Second, the computation of the entropy needs discretization of each attribute of the object. They did not indicate how to set the resolution of each attribute. Different from their work, we use Renyi entropy in our study. Renyi entropy lends itself nicely to non-parametric estimation and overcomes the difficulty in computing Shannon entropy [42]. In our proposed method, each object