Pattern Recognition 60 (2016) 334–347

Contents lists available at ScienceDirect

Pattern Recognition

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

A dynamic niching clustering algorithm based on individual-connectedness and its application to color image segmentation

Dongxia Chang [a,b,c,*], Yao Zhao [a,b,c], Lian Liu [a,b,c], Changwen Zheng [d]

[a] Institute of Information Science, Beijing jiaotong University, Beijing 100044, China
[b] School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China
[c] Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing 100044, China
[d] National Key Lab of Integrated Information System Technology, Institute of Software, Chinese Academy of Sciences, Beijing 100080, China

ARTICLE INFO

Article history:
Received 15 September 2014
Received in revised form
29 February 2016
Accepted 9 May 2016
Available online 21 May 2016

Keywords:
Clustering
Genetic algorithms
Niching
Connected individual
k-distance neighborhood
Image segmentation

ABSTRACT

In this paper, a dynamic niching clustering algorithm based on individual-connectedness (DNIC) is proposed for unsupervised classification with no prior knowledge. It aims to automatically evolve the optimal number of clusters as well as the cluster centers of the data set based on the proposed adaptive compact *k*-distance neighborhood algorithm. More specifically, with the adaptive selection of the number of the nearest neighbor and the individual-connectedness algorithm, DNIC often achieves several sets of connecting individuals and each set composes an independent niche. In practice, each set of connecting individuals corresponds to a homogeneous cluster and this ensures the separability of an arbitrary data set theoretically. An application of the DNIC clustering algorithm in color image segmentation is also provided. Experimental results demonstrate that the DNIC clustering algorithm has high performance and flexibility.

1. Introduction

Clustering analysis is a common technique for statistical multivariate analysis and has been used in a wide variety of engineering and scientific disciplines. In the past decades, a large family of clustering methods were proposed to partition data points into clusters such that the data points within the same group are similar to each other, while the data points in different groups are dissimilar [1–3]. Generally, these algorithms can be broadly divided into several classes: hierarchical [4], partitional [5,6], model-based [8] and density-based [7–9]. Among them, partitional methods attempt to directly decompose the data set into several disjointed clusters based on some criteria (i.e. validity functions). The most common criterion adopted by partitional methods is minimizing some measure of dissimilarity in the samples within each cluster and maximizing the dissimilarity of different clusters. However, the clustering algorithms based on some criteria become computationally expensive when the distribution of the data to be clustered is sophisticated.

Since the global optimum of the validity function would correspond to the most "valid" solution with respect to the functions, stochastic clustering algorithms based on evolutionary algorithms (EAs) have been reported to be able to optimize the validity functions to determine the number of clusters and partitioning of the data set simultaneously. In recent years, several clustering algorithms based on simple EA or its variants have been developed [10–24]. In fact, the original and many existing forms of EAs are usually designed for locating a single global solution as they typically converge to one final solution because of the global selection scheme used. So, the chromosome of these evolutionary clustering algorithms is described by a sequence of the cluster centers. When every cluster center is contained in the chromosome, then the fitness function reaches its global optimum. However, the clustering problem are "multimodal" by nature, that is, multiple clusters exist. Therefore, it might be desirable to locate all clusters that are considered as being satisfactory.

Numerous techniques have been developed in the past for locating multiple optima. These techniques are commonly referred to as "niching" methods. A niching method can be incorporated into a standard EA to promote and maintain formation of multiple stable subpopulations within a single population, with an aim to locate multiple optimal or suboptimal solutions. The basic idea of the niching methods is inspired by nature. In nature, an ecosystem is typically composed of different physical niches that exhibit different features and allow both the formation and the maintenance of different types of life (species). It is assumed that a species is made up of individuals with similar biological features

* Corresponding author at: Institute of Information Science, Beijing jiaotong University, Beijing 100044, China.
E-mail address: chang_dongxia@hotmail.com (D. Chang).

http://dx.doi.org/10.1016/j.patcog.2016.05.008
0031-3203/

capable of interbreeding among themselves, but unable to breed with individuals of other species [25]. By analogy, in artificial systems, a niche corresponds to a local optimum of the fitness function, and the individuals in one niche exhibit similar feature in terms of a given metric.

Many niching methods have been proposed in the EA literature [26–32]. Some representative examples include crowding [27], deterministic crowding [28], clearing [29], speciation [30], fitness sharing [31], implicit fitness sharing [32]. Crowding was initially designed to preserve population diversity [27]. In this method, an offspring is compared to a small random sample taken from the current population, and the most similar individual in the sample is replaced. Therefore, a parameter crowding factor (CF) is used to determine the size of the sample. In fact, the crowding was shown not be very effective at identifying multiple optima [28]. Then deterministic crowding was developed to improve the basic crowding method [28]. For replacement, instead of using crowding factor, deterministic crowding compares the new offspring directly to their parents, and replaces the parents only if the children have higher fitness. Another mechanism for maintaining population diversity is fitness sharing (FS) and implicit fitness sharing, which are probably the most widely used niching method [31–35]. In the fitness sharing method, the fitness represents the resource for which the individuals belonging to the same niche compete [31], while in the implicit methods [33,34], the sharing effects are achieved by means of a sample-and-match procedure. Fitness sharing was inspired by the "sharing" concept observed in nature, where an individual has only limited resources that must be shared with other individuals occupying the same niche in the environment. Moreover, niching methods have also been incorporated into the differential evolution [36], evolutionary computation [37], particle swarm optimization [38] to enhance their ability to handle multimodal optimization problems.

However, most existing niching methods suffer from a serious problem that their performance is subjected heavily to some niching parameters which are often difficult to set by a user. In FS, an appropriate niche radius should be defined which representing the maximal distance among individuals to be consider similar and therefore belonging to the same niche. In most circumstance, it is difficult to give an effective value for the niche radius without any a priori knowledge. In Ref. [33], a criterion for estimating the niche radius was proposed when the heights of the peaks and their distances is known a priori. Since in most of the real applications there is very little prior knowledge about fitness landscape, it is difficult to estimate the niche radius. In the implicit fitness sharing [34], sharing is accomplished by inducing competition for limited and explicit resources, and there is no specific limitation on the distance between peaks. This method avoids the difficult of appropriately choosing the niche radius. So, one of the most important limitations of FS seems to be removed. In fact, some other parameters, such as the size of the sample of individuals that compete, the number of competition cycles and the definition of a matching procedure, need to be set. To improve this situation, some adaptive niching methods have been developed [39,40]. In Ref. [40], the parameters can dynamically adjust according to the devised population diversity index.

In addition to the difficulty of setting effective values for the niche radius [41] and the lack of an explicit mechanism for identifying or providing any information about the location of the peaks in the fitness landscape [42], the definition of niches used by FS method is implicit. In order to ensure the subpopulations are steadily formed and maintained, only the individuals belonging to the same niche should share the resources of the niche. But for the FS method [43], each individual shares its fitness with all the individuals located at a distance smaller than the niche radius, no matter for the niche to which they belong. Therefore, individuals belonging to different niches will share their fitness, and this will make the nonperfect discrimination between niches. In order to overcome this drawback, several dynamic niching methods were proposed [42,44]. These methods are based upon a dynamic, explicit identification of niches discovered at each generation and the FS mechanism is restricted to individuals belonging to the same niche. However, the performance of these algorithms is dependent on the niche radius. When wrong value for the niche radius is selected, the algorithm did not find all the niches perfectly.

In this paper, a dynamic niching clustering algorithm based on individual-connectedness (DNIC) is proposed. Within the DNIC clustering algorithm, an adaptive compact k-distance neighborhood algorithm is developed to preserve the diversity of the population. A simpler representation is adopted, whereby each individual represents a single cluster center. All the niches presented in the population at each generation are automatically identified. In order to overcome the dependence of the parameter k (i.e. the size of the neighborhood), an adaptive selection of the number of the nearest neighbor is considered. This makes the algorithm work properly. After the adaptive selection of the nearest neighbor, an individual-connectedness algorithm is used to achieve several sets of connecting individuals and each set composes an independent niche.

The remainder of this paper is organized as follows. Section 2 provides some definitions necessary for our approach. Section 3 describes the compact k-distance neighborhood algorithm. Then a detail of our DNIC clustering algorithm is presented in Section 4. Extensive experimental comparisons on synthetic and real-world images are demonstrated in Section 5. Finally, the paper is concluded in Section 6.

2. Preliminaries

In this section, some definitions needed in the next section are given. In Ref. [45], a neighborhood based density factor (NDF) was proposed which uses the neighborhood relationship among data points. Here we refine its basic concepts into just four ones: neighbor-based density factor, local dense point, local sparse point and local even point. The four key definitions facilitate to design DNIC clustering algorithm. Let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be a finite subset of a N-dimensional vector space.

Definition 1. Let $\mathbf{x}$ be an arbitrary vector in $\mathbf{X}$. Then the neighbor-based density factor of $\mathbf{x}$, denoted by $NDF(\mathbf{x})$, is evaluated as follows [45]:

$$NDF(\mathbf{x}) = \frac{|R - kNB(\mathbf{x})|}{|kNB(\mathbf{x})|} \tag{1}$$

where $kNB(\mathbf{x})$ is the k nearest neighbors set of $\mathbf{x}$ and $R - kNB(\mathbf{x})$ is the reverse k nearest neighbors set of $\mathbf{x}$. More specifically, $R - kNB(\mathbf{x})$ is the set of vectors whose k neighborhood contain $\mathbf{x}$. In practice, $|kNB(\mathbf{x})|$ equals to k, and $|R - kNB(\mathbf{x})|$ is quite discrepant for different vectors. As a result, there are three situations for $NDF(\mathbf{x})$: larger than 1, equal to 1 and less than 1 [45].

Definition 2. Point $\mathbf{x}$ is a local dense point if its $NDF(\mathbf{x}) \geq 1$.

This means that $\mathbf{x}$ is surrounded by many points (i.e. points in its reverse k-neighborhood). In fact, these data points look more like the centroids of the data set.

Definition 3. Point $\mathbf{x}$ is a local sparse point if its $NDF(\mathbf{x}) \ll 1$.

This means that $\mathbf{x}$ almost has no reverse k-neighborhood and all the points are far from it.