Applied Mathematics and Computation 218 (2011) 1267–1279

Contents lists available at ScienceDirect

Applied Mathematics and Computation

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

Automatic clustering using genetic algorithms

Yongguo Liu [a,b,c,d,*], Xindong Wu [d], Yidong Shen [b]

[a] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, PR China
[b] State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing 100191, PR China
[c] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, PR China
[d] Department of Computer Science, University of Vermont, Burlington, VT 05405, USA

ARTICLE INFO

Keywords:
Clustering
Genetic algorithms
Noising method
Davies–Bouldin index
K-means algorithm

ABSTRACT

In face of the clustering problem, many clustering methods usually require the designer to provide the number of clusters as input. Unfortunately, the designer has no idea, in general, about this information beforehand. In this article, we develop a genetic algorithm based clustering method called automatic genetic clustering for unknown K (AGCUK). In the AGCUK algorithm, noising selection and division–absorption mutation are designed to keep a balance between selection pressure and population diversity. In addition, the Davies–Bouldin index is employed to measure the validity of clusters. Experimental results on artificial and real-life data sets are given to illustrate the effectiveness of the AGCUK algorithm in automatically evolving the number of clusters and providing the clustering partition.

© 2011 Elsevier Inc. All rights reserved.

1. Introduction

Clustering is a fundamental problem that frequently arises in a great variety of application fields such as pattern recognition, machine learning, statistics, etc. It is a formal study of algorithms and methods for grouping or classifying objects without category labels. The resulting partition should possess two properties: (1) homogeneity within the clusters, i.e. objects belonging to the same cluster should be as similar as possible, and (2) heterogeneity between the clusters, i.e. objects belonging to different clusters should be as different as possible. Many clustering techniques have been proposed [1,2]. Among them, the *K*-means algorithm is an important one. It is an iterative hill-climbing algorithm and the solution obtained depends on the initial clustering. Although the *K*-means algorithm had been applied to many practical clustering problems successfully, it may converge to a partition that is significantly inferior to the global optimum [3].

Recently, researchers solved the clustering problem by stochastic optimization methods such as genetic algorithms, tabu search, simulated annealing, etc. Liu et al. [4] integrated a tabu list into the genetic algorithm based clustering algorithm to prevent several fitter individuals from occupying the population and to maintain population diversity. In addition, an aspiration criterion is adopted to keep selection pressure. Bandyopadhyay and Maulik [5] designed a genetic clustering approach. They used the *K*-means algorithm to provide the domain knowledge and improve the search capability of genetic algorithms. Laszlo and Mukherjee [6] presented a genetic algorithm for evolving the cluster centers in the *K*-means algorithm. The set of the cluster centers is represented using a hyper-quadtree constructed on the data. Liu et al. [7] combined the *K*-means algorithm and the tabu search approach to accelerate the convergence speed of the tabu search based clustering algorithm. Ng and Wong [8] proposed a tabu search based fuzzy *K*-modes algorithm for clustering categorical objects. Bandyopadhyay et al. [9] integrated the *K*-means algorithm into the simulated annealing based clustering method to modify the cluster centroids.

* Corresponding author at: School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, PR China.
E-mail address: liuyg@uestc.edu.cn (Y. Liu).

0096-3003/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.amc.2011.06.007

By redistributing objects among clusters probabilistically, the presented method obtains better results than the *K*-means algorithm. Güngör and Ünler [10] combined the *K*-harmonic means algorithm and the simulated annealing method to deal with the clustering problem. The simulated annealing method is used to generate non-local moves for the cluster centers and to select the best solution. Liu et al. [11] adopted the noising method, a metaheuristic technique reported by Charon and Hudry [12], to solve the clustering problem. With lower computational cost than Bandyopadhyay et al.'s method [9], the proposed method is inferior to the latter in terms of solution quality. By modeling the clustering problem as an optimization problem, Mahdavi et al. [13] proposed a harmony search based clustering algorithm for grouping the web documents. They hybridized the *K*-means algorithm and the harmony search method in two ways and designed two hybrid algorithms. Pacheco [14] adopted the scatter search approach to deal with the clustering problem under the criterion of minimum sum-of-squares clustering. Within the framework of the proposed method, greedy randomized adaptive search procedure (GRASP) based constructions, *H*-means+ algorithm, and tabu search are integrated. Jarhoui et al. [15] designed a clustering approach based on the combinatorial particle swarm optimization (CPSO) algorithm. In the CPSO method, each particle is represented as a string of length n (where n is the number of objects) and the ith element of the string denotes the group number assigned to object i. The CPSO algorithm obtains better results than a genetic algorithm based clustering method in some cases. Shelokar et al. [16] proposed an ant colony optimization method for grouping N objects into K clusters. The presented method employs distributed agents which mimic the way real ants find the shortest path from their nest to food source and back. Fathian et al. [17] presented the application of honeybee mating optimization in clustering (HBMK-means). By experimental simulations, the HBMK-means method is proved to be better than other heuristic algorithms in clustering, such as genetic algorithm, simulated annealing, tabu search, and ant colony optimization.

The aforementioned clustering techniques [3–11,13–17] require the designer to provide the number of clusters as input. Unfortunately, in many real-life cases the number of clusters in a data set is not known *a priori*. How to automatically find a proper value of the number of clusters and provide the appropriate clustering under this condition becomes a challenge. In this paper, our aim is to develop a genetic algorithm based clustering method called automatic genetic clustering for unknown K (AGCUK) to automatically find the number of clusters and provide the proper clustering partition. We design two new operators, noising selection and division–absorption mutation, to keep the balance between selection pressure and population diversity. The Davies–Bouldin index is employed as a measure of the validity of clusters. Experimental results on artificial and real-life data sets are given to illustrate the superiority of the AGCUK algorithm over four known genetic clustering methods.

The remaining part of this paper is organized as follows. The related work on the automatic clustering method based on genetic algorithms is reviewed in Section 2. In Section 3, we propose the AGCUK algorithm and give detailed descriptions. In Section 4, the choice of the original noise rate $r_{\max}$ and the terminal noise rate $r_{\min}$ is discussed, how to estimate selection pressure and population diversity is given, and performance comparison between AGCUK and four known genetic algorithm based clustering methods is conducted for experimental data sets. Finally, some conclusions are drawn in Section 5.

2. Related work

In this study, we focus on how to solve the automatic clustering problem using genetic algorithms. In this regard, some attempts have been made to use genetic algorithms for automatically clustering the data. Bandyopadhyay and Maulik [18] applied the variable string length genetic algorithm, with real encoding of the coordinates of the cluster centers in the chromosome, to the clustering problem. Experimental results on artificial and real-life data sets show that their algorithm is able to evolve the number of clusters as well as provide the proper clustering. Tseng and Yang [19] proposed a genetic algorithm based approach for the clustering problem. Their method consists of two stages, nearest neighbor clustering and genetic optimization. Equipped with a heuristic strategy, the proposed method can search for a proper number of clusters and classify nonoverlapping objects into these clusters. Bandyopadhyay and Maulik [20] exploited the searching capability of genetic algorithms for finding the number of clusters as well as the proper clustering of a given data set. A string representation, comprising both real numbers and the do not care symbol, is used to encode a variable number of clusters. Effectiveness of their technique is demonstrated for both artificial and real-life data sets. Lai [21] adopted the hierarchical genetic algorithm to solve the clustering problem. In the proposed method, the chromosome consists of two types of genes, control genes and parametric genes. The control genes are coded as binary digits. The total number of "1" represents the number of clusters. The parametric genes are coded as real numbers to represent the coordinates of the cluster centers. The relationship between the control genes and the parametric genes is that the activation of the latter is governed by the value of the former. If the value of a control gene is "1", then the associated parametric genes due to that particular active control gene are activated; otherwise the associated parametric genes are disabled. Experimental results on artificial and real-life data sets show Lai's method can search for the number of clusters and provide the proper clustering. Lin et al. [22] presented a genetic clustering algorithm based on the use of a binary chromosome representation. The proposed method selects the cluster centers directly from the data set. With the aid of a look-up table, the distances between all pairs of objects are saved in advance and evaluated only once throughout the evolution process. By experimental simulations, the superiority of their algorithm over Bandyopadhyay and Maulik's method [20] is shown. Lai and Chang [23] presented a clustering based approach using a hierarchical evolutionary algorithm (HEA) for medical image segmentation. By means of a hierarchical structure in the chromosome, the proposed approach can automatically classify the image into appropriate classes and avoid the difficulty of

Download English Version:

https://daneshyari.com/en/article/4630415

Download Persian Version:

https://daneshyari.com/article/4630415

Daneshyari.com