Available online at www.sciencedirect.com
ScienceDirect
Pattern Recognition Letters 29 (2008) 515–524
Pattern Recognition Letters
www.elsevier.com/locate/patrec

An integer-coded evolutionary approach for mixture maximum likelihood clustering

Mohamad M. Tawfick [a], Hazem M. Abbas [b,*], Hussein I. Shahein [b]

[a] Mentor Graphics Inc., 51 Beirut Street, Helipolis, Cairo 11341, Egypt
[b] Ain Shams University, Department of Computer and Systems Engineering, Abbasia, Cairo 11571, Egypt

Received 2 July 2006; received in revised form 28 May 2007
Available online 19 November 2007

Communicated by T.K. Ho

Abstract

This paper outlines an algorithm for solving the maximum mixture likelihood clustering problem using an integer-coded genetic algorithm (IGA-ML) where a fixed length chromosome encodes the object-to-cluster assignment. The main advantage of the outlined algorithm (IGA-ML) compared with other known algorithms, such as the k-means technique, is that it can successfully discover the correct number of clusters, in addition to carrying out the partitioning process. The algorithm implements a post-fixing sorting mechanism that drastically reduces the searched solution space by eliminating duplicate solutions that appear after applying the genetic operations. Simulation results show the effectiveness of the algorithm especially with the case of overlapping clusters.

Keywords: Clustering; Mixture maximum likelihood; Evolutionary algorithms; Genetic algorithms

1. Introduction

Clustering (Jain et al., 1999; Hartigan, 1975) is an unsupervised classification process that aims at dividing a given set $\{x_1, ..., x_n\}$ of n data points into several non-overlapping homogeneous groups. Each such group (or cluster) should only contain similar data items that should not be scattered in different groups. This definition requires a measure of similarity/dissimilarity, based on which clustering becomes a problem of grouping data objects such that the within-group similarity and the between-groups dissimilarity are maximized.

Among many important issues pertained to the clustering problem is the large solution space to be explored. Another critical issue is the definition of the clustering quality criterion, which is represented in this work through the maximum mixture likelihood. Sorting n objects into k groups can be done in several ways. The number of such possible ways, $N(n,k)$, is given by Lui's formula (Lui, 1968),

$$N(n,k) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^i\binom{k}{i}(k-i)^n.$$

The problem gets more complicated when the number of clusters is not known a priori. Traditional clustering algorithms search a relatively small subset of the solution space, due to constraints applied on the number of clusters, clustering criteria, and clustering method; and therefore, finding the optimal solution is not guaranteed.

Many different approaches (Verbeek, 2004) to the clustering problem have been developed and applied in different fields. Traditional clustering algorithms can be categorized into either hierarchal or optimization based algorithms. Hierarchal algorithms can be further categorized into either agglomerative or divisive. The most famous agglomerative algorithms are single-linkage (nearest neighbour), complete linkage (furthest neighbour),

* Corresponding author. Tel.: +2 02 413 5457; fax: +2 02 418 6945.
E-mail address: h.abbas@ieee.org (H.M. Abbas).

0167-8655/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2007.11.003

average linkage (Sokal and Sneath, 1963), centroid clustering (Lance and Williams, 1967), and Ward's minimum variance method (Ward, 1963). Further details on these clustering techniques can be found in (Anderberg, 1973; Everitt, 1993; Vogt and Nagel, 1992). Optimization techniques, such as hill-climbing and k-means (Anderberg, 1973; Everitt, 1993), optimize a predefined criterion producing a single clustering. Jain et al. (1999) presents an extensive study covering a wide range of clustering algorithms, including fuzzy algorithms, neural network algorithms, hierarchal clustering, and evolutionary approaches for clustering.

Since genetic algorithms (GA) (Goldberg, 1989) have been proved to be an efficient and highly parallel way of optimizing complex, multi-modal and multi-variable functions, an integer-coded genetic algorithm is proposed here to address the clustering problem. Some adaptations to the classical GA are needed to deal with the problem at hand. We used the group-number for chromosome representation and the likelihood function as the fitness function to maximize. Suitable crossover and mutation operators are implemented, and a sorting algorithm is introduced to provide faster convergence.

The remainder of this paper is structured as follows. Section 2 reviews some published algorithms that implement evolutionary concepts in the clustering problem. Maximum likelihood estimation is presented in Section 3. The proposed genetic algorithm, together with the modifications necessary to adapt the GA for the clustering problem, are presented in Section 4. Section 5 describes the experimental setup employed for evaluating the algorithm on different datasets. Detailed discussion on the results is provided.

2. Evolutionary clustering algorithms

Static clusterings techniques require that the number of clusters be chosen a priori. For static clustering approaches to find the optimal number of clusters requires repeated runs with different cluster numbers, and this is often extremely time consuming. A review of some evolutionary implementations for static clustering is presented in (Hall et al., 1999). A real-coded GA is proposed in (Maulik and Bandyopadhyay, 2000) that finds the optimal centers of a fixed number of clusters, and shows the superiority of the proposed GA algorithm over the commonly used k-means algorithm.

Dynamic partitioning, as opposed to the static one, does not require the a priori specification of the number of clusters. Ghozeil and Fogel (1996) implemented a GA adopting variable length genomes to address the problem of dynamic partitioning clustering, where insertion and deletion operators are used to modify the number of clusters. The algorithm lacks the crossover operator and operates on small ($\leqslant 5$) number of clusters. Tseng and Yang (2001) proposes a genetic algorithm for the clustering problem that is suitable for clustering the data with compact spherical clusters. The algorithm obtains several possible clusterings, then, applies a heuristic strategy to choose a good clustering. A genetic-based expectation-maximization (GA-EM) algorithm, for learning Gaussian mixture models from multivariate data, is proposed in (Pernkopf and Bouchaffra, 2005). The algorithm takes advantage of the GA and the EM algorithms by combining them into a single procedure. In (Randy, 2003), GA and k-means clustering method were incorporated within a multi-resolution structure to approach the texture segmentation problem. Babu and Murty (1994) explore evolutionary strategies for solving the clustering problem. Cowgill et al. (1999) proposed a GA clustering technique to maximize a variance-ratio (VR) based goodness-of-fit criterion defined in terms of external cluster isolation and internal cluster homogeneity. Sharman and McClurkin (1989) outlines a GA for maximum likelihood (ML) parameter estimation from noisy data measurements, the ML estimator requires some a priori knowledge in the form of the exact joint probability density functions (PDF's) of the signal and noise sequences.

This paper presents another evolutionary clustering algorithm (IGA) that adapts an integer-coded chromosome representation and uses a maximum likelihood (ML) fitness function. The IGA algorithm avoids expensive floating-point computations. It also introduces a post-fixing sorting algorithm that provides faster convergence. It implements a problem-specific mutation operators that widen the search space. Some modifications are made to the conventional ML function that proved to produce better clustering output.

3. The maximum mixture likelihood

Maximum likelihood (ML) is one of the most widely used methods for statistical estimation. The usage of the term "maximum likelihood" was pioneered by Fisher (1925), who exerted great effort to spread its use and derived the optimality properties of the resulting estimates. The maximum likelihood estimate (mle) of Θ is that value of Θ that maximizes $\ell(\Theta)$: it is the value that makes the observed data the "most probable". For a Gaussian mixture, the random variables $X_1, X_2, \ldots X_n$ are iid $N(\mu, \Sigma)$, their joint density is given by

$$f(x_1, \ldots, x_n|\mu, \Sigma) = \prod_{i=1} n \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x_i-\mu)^{\mathrm{T}}\Sigma^{-1}(x_i-\mu)}$$

where Σ is the covariance matrix of the observed data, $|\Sigma|$ is its determinant, μ is the mean vector and T indicates transposition.

Regarded as a function of the two parameters, μ and Σ, the likelihood is given by

$$L(\mu, \Sigma) = -\frac{n}{2}\log|\Sigma| - \frac{n^2}{2}\log 2\pi - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu_i)^{\mathrm{T}}\Sigma^{-1}(x_i-\mu_i)$$

The standard method used to fit finite mixture models to observed data is the expectation-maximization (EM) algorithm, which converges to a maximum likelihood (ML) estimate of the mixture parameters (Dempster et al.,