Determining the Optimal Number of Clusters Using a New Evolutionary Algorithm

Wei Lu and Issa Traore
Department of Electrical and Computer Engineering
University of Victoria
{wlu, itraore}@ece.uvic.ca

Abstract

Estimating the optimal number of clusters for a dataset is one of the most essential issues in cluster analysis. An improper pre-selection for the number of clusters might easily lead to bad clustering outcome. In this paper, we propose a new evolutionary algorithm to address this issue. Specifically, the proposed evolutionary algorithm defines a new entropy-based fitness function, and three new genetic operators for splitting, merging, and removing clusters. Empirical evaluations using the synthetic dataset and an existing benchmark show that the proposed evolutionary algorithm can exactly estimate the optimal number of clusters for a set of data.

1. Introduction

Identifying the optimal number of clusters for a set of data is essential for effective and efficient data clustering. For instance, a clustering algorithm such as k-means may generate a bad clustering result if initial partitions are not properly chosen, a situation that occurs often and is not an obvious task. Another popular clustering approach sensitive to this problem is based on Gaussian mixture model (GMM). GMM is based on the assumption that the data to be clustered are drawn from one of several Gaussian distributions and it was suggested that Gaussian mixture distribution could approximate any distribution up to arbitrary accuracy, as long as a sufficient number of components are used [1]. A common approach for estimating the parameters of GMM is Expectation-Maximization (EM) algorithm [2].

Previous attempts for estimating the number of mixing components for the GMM are mainly based on statistical techniques. Some examples of these previous works were suggested in the literature [3] and [4]. Although the previous approaches based on statistical techniques have proved their ability to estimate the optimal number of clusters, they are prone to converge into local optima since they usually stop to perform further search when the corresponding criterions reach certain thresholds. In contrast, the evolutionary computation schemes have the inherent potential capability to escape from local maximum since their search space for optimal solutions can be extended by genetic operations and optimization selection. Based on this, in this paper we tackle the issue of number of clusters estimation using a novel evolutionary approach, which combines the Gaussian mixture and the EM algorithm.

The rest of the paper is structured as follows. Section 2 illustrates the evolutionary algorithm for determining the optimal number of clusters for a set of data. Section 3 presents and discusses the results obtained in the empirical validation of the algorithm.

2. Evolutionary algorithm

2.1. Evolutionary entities

2.1.1. Representation of evolutionary individuals. EM algorithm generates an estimate for the set of parameters $\{\alpha_i, \mu_i, \sigma_i\}$ and posterior probabilities $p(i|x_n)$. The posterior probability describes the likelihood that the data pattern x_n approximates to a specified Gaussian component i. Each data x_n is assigned to the corresponding Gaussian component i according to $p(i|x_n)$ and final clustering results are statistically represented by the set of parameters $\{\alpha_i, \mu_i, \sigma_i\}$, also evolutionary individuals.

2.1.2. Evolutionary operators. During the evolution, we need sometimes to split components, as well as to merge or delete them. Therefore we propose three new evolutionary operators called *splitting*, *merging* and *deletion* operators, used as their names indicate for splitting, merging and removing the components. The detailed definitions of the operators can be found in [5].

Proceedings of the 17th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'05)
1082-3409/05 $20.00 © 2005 IEEE

2.1.3. Fitness function. We define an entropy-based fitness function as the criterion for measuring how well the mixture model approximates the unknown density underlying the data samples: *the higher the fitness function value, the better the approximation is*. The entropy-based fitness function EF_T is defined as:

$$EF_T = \sum_{i=1}^{k} \alpha_i \left| ef_i \right|$$

Where ef_i denotes the individual fitness for each of the mixture component, which is defined as follows:

$$ef_i = \frac{H(C_i)}{H_{\max}(C_i)}$$

Where C_i stands for the set of data whose probabilities of belonging to the i^{th} component are the highest compared to other components, and $H(C_i)$ represents the entropy of these data. $H_{max}(C_i)$ is the theoretical maximum entropy for an individual component.

2.2. Proposed algorithm

Table 1. The proposed evolutionary algorithm

Function: GA_Mixture_Model (population)
returns a *population with optimal number of clusters*
Inputs: *population*
Initialization: $j \leftarrow 0$; $population_j \leftarrow population$;
$population_{optimal} \leftarrow EM(population_j)$;
$EF_{Toptimal} \leftarrow EF_T^j$;
Repeat: $j \leftarrow j+1$;
Apply genetic operators to $population_{optimal}$
yielding $population_j$;
$population_j \leftarrow EM(population_j)$;
Compute EF_T^j;
If $\left(EF_T^j > EF_{Toptimal}\right)$
do $population_{optimal} \leftarrow population_j$,
$EF_{Toptimal} \leftarrow EF_T^j$;
Until: $EF_{Toptimal} > th_1$ or $j \geq th_2$ or $k = 0$
Return $population_{optimal}$

The proposed evolutionary algorithm is described in Table 1. The algorithm starts with an initial number of components, which is empirically selected based on the size of the data. A population denoted by $population_j$ corresponds to a set of new individuals created during the j^{th} generation; we denote by k_j the size of $population_j$. An individual is denoted by $<\alpha_i^j, \mu_i^j, \Sigma_i^j>$, where i represents the i^{th} component generated; j stands for the number of generations. We use $population_{optimal}$ and $k_{optimal}$ to denote the optimal individuals and the optimal number of clusters, respectively. New individuals are composed by $population_j$ and $population_{optimal}$. The entropy-based fitness value associated with $population_j$ is denoted by EF_T^j, while $EF_{Toptimal}$ refers to the current optimal fitness value during evolution.

3. Empirical validation

We conducted an empirical validation of our evolutionary algorithm, in which, we used two different datasets. The first dataset, generated synthetically, contains 300 data items. The data items were independently generated using three bi-dimensional normal distributions. The second dataset is the famous *Iris* dataset. The maximum number of iterations was set to 500 and the fitness threshold was set to 0.98. During the evolution, the initial number of clusters was empirically set to 30.

The evaluation results show that our algorithm has the capability to converge into the global optimal solution with a high merging probability and a low splitting probability. It estimates exactly the optimal number of clusters for the synthetic dataset and the real dataset.

Interested readers are referred to [5] for more details about the algorithm, and the validation process and results.

4. References

[1] B.D. Ripley, Pattern Recognition and Neural Networks. Cambridge, U.K., Cambridge University Press, 1996.

[2] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion)", *Journal of the Royal Statistical Society B*, Vol. 39, Pages: 1-38, 1997.

[3] S. Richardson and P. J. Green, "On Bayesian Analysis of Mixtures with an Unknown Number of Components (with discussion)", *Journal of the Royal Statistical Society: Series B*, Vol. 59, No. 4, Pages: 731-792, 1997.

[4] N. Vlassis and A. Likas, "A Kurtosis-based Dynamic Approach to Gaussian Mixture Modeling", *IEEE Transaction on Systems, Man, and Cybernetics, Part A*, 29(4), Pages: 393-399, 1999.

[5] W. Lu, "An Unsupervised Anomaly Detection Framework for Multiple-Connection based Network Intrusions", PhD Thesis, November 2005, University of Victoria, ECE Department, Victoria, BC, V8W 3P6, Canada.