

Genetic algorithms for subset selection in model-based clustering

Luca Scrucca

Abstract Model-based clustering assumes that the observed data can be represented by a finite mixture model, where each cluster is represented by a parametric distribution. In the multivariate continuous case the Gaussian distribution is often employed. Identifying the subset of relevant clustering variables allows to achieve parsimony of unknown parameters, thus yielding more efficient estimation, clearer interpretation, and, often, better clustering partitions. This paper discusses variable or feature selection for model-based clustering. The problem of subset selection is recast as a model comparison problem, and BIC is used to approximate Bayes factors. Searching over the potentially vast solution space is performed through genetic algorithms, which are stochastic search algorithms that use techniques and concepts inspired by evolutionary biology and natural selection.

Key words: Model selection, model-based clustering, genetic algorithms.

1 Introduction

In the model-based approach to clustering each cluster is represented by a parametric distribution, and a finite mixture model is used to model the observed data. Parameters are then estimated by optimizing the fit, expressed by the likelihood (eventually penalized), between the data and the model.

In the multivariate case, selecting the relevant clustering variables allows to achieve parsimony of unknown parameters, yielding more efficient estimation, clearer interpretation of the parameters, and, often, better clustering partitions.

However, the problem of selecting the best subset of clustering variables is a non-trivial exercise, particularly when a large number of features are available. Therefore, search strategies for model selection are needed to explore the vast solution space. A classical way to search for a feasible solution is to adopt a forward/backward stepwise selection strategy. Major criticism is that stepwise searching rarely finds the overall best model or even the best subset of a particular size.

Dipartimento di Economia, Finanza e Statistica, Università degli Studi di Perugia, e-mail: luca@stat.unipg.it

This paper discusses a computationally feasible search based on genetic algorithms that addresses the potentially daunting statistical and combinatorial problems presented by subset selection in finite mixture modeling.

2 Model-based clustering

Model-based clustering assumes that the observed data are generated from a mixture of G components, each representing the probability distribution for a different group or cluster (McLachlan and Peel, 2000). The general form of a finite mixture model is $f(\mathbf{x}) = \sum_{g=1}^G \pi_g f_g(\mathbf{x}|\boldsymbol{\theta}_g)$, where the π_g 's are the mixing probabilities such that $\pi_g \geq 0$ and $\sum \pi_g = 1$, $f_g(\cdot)$ and $\boldsymbol{\theta}_g$ are, respectively, the density and the parameters of the g -th component ($g = 1, \dots, G$). With continuous data, we often take the density for each mixture component to be the multivariate Gaussian $\phi(\mathbf{x}|\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ with parameters $\boldsymbol{\theta}_g = (\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$. Thus, clusters are ellipsoidal, centered at the means $\boldsymbol{\mu}_g$, and with other geometric features, such as volume, shape and orientation, determined by $\boldsymbol{\Sigma}_g$.

Parsimonious parametrization of covariance matrices can be adopted through eigenvalue decomposition in the form $\boldsymbol{\Sigma}_g = \lambda_g \mathbf{D}_g \mathbf{A}_g \mathbf{D}_g^\top$, where λ_g is a scalar controlling the volume of the ellipsoid, $\mathbf{A}_g$ is a diagonal matrix specifying the shape of the density contours, and $\mathbf{D}_g$ is an orthogonal matrix which determines the orientation of the corresponding ellipsoid (Banfield and Raftery, 1993; Celeux and Govaert, 1995). Fraley and Raftery (2006, Table 1) report some parametrizations of within-group covariance matrices available in the MCLUST software, and the corresponding geometric characteristics. Maximum likelihood estimates for this type of mixture models can be computed via the EM algorithm (Dempster et al., 1977; Fraley and Raftery, 2002), while model selection could be based on the Bayesian information criterion (BIC) (Fraley and Raftery, 1998) or the integrated complete-data likelihood (ICL) criterion (Biernacki et al., 2000).

3 Subset selection in model-based clustering

Raftery and Dean (2006) discussed the problem of variable selection for model-based clustering by recasting the problem as a model selection procedure. Their proposal is based on the use of BIC (Schwartz, 1978) to approximate Bayes factors (Kass and Raftery, 1995) for comparing mixture models fitted on nested subsets of variables. A generalization of this approach has been recently discussed by Maugis et al. (2009).

Suppose the set of available variables is partitioned into three parts: the set of already selected variables, $\mathbf{X}_1$, the set of variables being considered for inclusion or exclusion from the active set, $\mathbf{X}_2$, and the set of the remaining variables, $\mathbf{X}_3$. Raftery and Dean (2006) showed that the inclusion (or exclusion) of variables can

be assessed by using the Bayes factor

$$B_{12} = \frac{p(\mathbf{X}_2|\mathbf{X}_1, M_1)p(\mathbf{X}_1|M_1)}{p(\mathbf{X}_1, \mathbf{X}_2|M_2)}, \quad (1)$$

where $p(\cdot|M_k)$ is the integrated likelihood of model M_k ($k = 1, 2$). Model M_1 specifies that $\mathbf{X}_2$ is conditionally independent of the cluster membership given $\mathbf{X}_1$, whereas model M_2 specifies that $\mathbf{X}_2$ is relevant for clustering once $\mathbf{X}_1$ has been included in the model. An important aspect of this formulation is that the set $\mathbf{X}_3$ of other variables plays no role. Minus twice the logarithm of the Bayes factor (1) can be approximated by the following BIC difference:

$$\text{BIC}_{\text{diff}} = \text{BIC}_{\text{clust}}(\mathbf{X}_1, \mathbf{X}_2) - \text{BIC}_{\text{not clust}}(\mathbf{X}_1, \mathbf{X}_2), \quad (2)$$

where $\text{BIC}_{\text{clust}}(\mathbf{X}_1, \mathbf{X}_2)$ is the BIC value for the “best” clustering mixture model fitted using both $\mathbf{X}_1$ and $\mathbf{X}_2$ features, whereas $\text{BIC}_{\text{not clust}}(\mathbf{X}_1, \mathbf{X}_2)$ is the BIC value for no clustering for the same set of variables. Assuming that $\mathbf{X}_2$ is made up of a single variable, the latter can be written as

$$\text{BIC}_{\text{not clust}}(\mathbf{X}_1, \mathbf{X}_2) = \text{BIC}_{\text{clust}}(\mathbf{X}_1) + \text{BIC}_{\text{reg}}(\mathbf{X}_2|\mathbf{X}_1),$$

i.e. the BIC value for the “best” clustering model fitted using $\mathbf{X}_1$ plus the BIC value for the regression of the $\mathbf{X}_2$ variable on the $\mathbf{X}_1$ variables. In all cases the “best” clustering model is identified with respect to both the number of mixture components and model parametrization. A positive value of BIC_{diff} indicates that variable $\mathbf{X}_2$ is relevant for clustering.

Equation (2) has been used by Raftery and Dean (2006) to propose a stepwise greedy search algorithm for variables selection. However, the proposed algorithm is known to be sub-optimal, in the sense that there exists the risk of finding only a local optimum in the model space. To overcome this drawback, we propose using genetic algorithms for searching over the whole model space. However, we need to recast the problem from the stepwise perspective adopted by the above mentioned authors.

Suppose the set of variables already included for clustering is empty and we want to evaluate the clustering model obtained from a candidate subset of k variables ($0 < k \leq p$). The Bayes factor for comparing the no clustering model (M_0) against a candidate clustering model (M_k) fitted using a subset of k variables ($\mathbf{X}_k$) is given by

$$B_{0k} = \frac{p(\mathbf{X}_k|M_0)}{p(\mathbf{X}_k|M_k)}.$$

Again, we can approximate the above equation using the following BIC difference:

$$\text{BIC}_k = \text{BIC}_{\text{clust}}(\mathbf{X}_k) - \text{BIC}_{\text{not clust}}(\mathbf{X}_k). \quad (3)$$

Thus, the criterion becomes to maximize the difference between the maximum BIC value from the finite mixture model for clustering, i.e. assuming $G \geq 2$, and the

maximum BIC value for no clustering, i.e. assuming $G = 1$, with both models estimated on the candidate subset of variables $\mathbf{X}_k$. By evaluating the criterion in (3) on a large number of possible subsets of different sizes, we may choose the “best” subset as the one which provides the largest BIC_k .

However, the number of possible subsets of k variables from a total of p is given by $\binom{p}{k}$. Thus, the space of all possible subsets of size k ranging from 1 to p has number of elements equal to $\sum_{k=1}^p \binom{p}{k} = 2^p - 1$. An exhaustive search becomes unfeasible even for moderate values of p . Genetic algorithms are a natural candidate for searching over this potentially very large model space.

4 Genetic algorithms

Genetic algorithms (GAs) are stochastic search algorithms based on concepts of biological evolution and natural selection (Holland, 1992), which have been applied to find exact or approximate solutions to optimization and search problems (Goldberg, 1989; Haupt and Haupt, 2004).

A GA is started with a set of randomly generated individuals or solutions called population. Solutions are traditionally represented as binary strings of 0s and 1s, but other encodings are also possible. The fitness of every individual in the population is evaluated, and a new population is formed by applying genetic operators, such as *selection*, *crossover* and *mutation*, to individuals.

The choice of the fitness function depends on the problem. In optimization problems, it is usually the objective function which needs to be minimized (or maximized). Individuals are selected to form new offsprings according to their fitness value. Genetic operators, such as crossover (exchanging substrings of two individuals to obtain a new offspring) and mutation (randomly mutate individual bits), are applied probabilistically to the selected offsprings to produce a new population of individuals. The new population is then used in the next iteration of the algorithm.

A GA is thus based on a sequence of cycles of evaluation and genetic operations, which are iterated for many generations. Generally the overall fitness of the population improves, and better solutions are likely to be selected as the search continues. Usually, the algorithm terminates when either a maximum number of generations has been produced, a satisfactory fitness level has been reached for the population, or the algorithm has achieved a steady state.

4.1 A genetic algorithm for subset selection in model-based clustering

In the following we describe specific points of our implementation of GAs to model-based clustering.

Genetic coding scheme Each variables subset is encoded as a string, where each locus in the string is a binary code indicating the presence (1) or absence (0) of a given variable. For example, in a model-based clustering with $p = 5$ variables the string 11001 represents a model where variables 1, 2, and 5 are included, whereas variables 3 and 4 are excluded from the model.

Generation of a population of models Population size N (i.e., the number of models to be fitted at each generation) is an important parameter of GA search. A sufficient large set of models ensures that a large portion of the search space is explored. On the other hand, if there are too many individuals (models) in the population, GA slows down. It is known that after some limit (which depends on the encoding and the problem) increasing population size is not useful. Because for p variables there are $2^p - 1$ possible subsets, we used $N = \min(2^p - 1, 50)$.

A fitness function to evaluate the clustering of any model The BIC criterion in (3) is used to assign a fitness value to each model in the GA population.

Genetic operators

- *Selection*: this step involves selecting models for mating based on their fitness. A rank selection scheme is used to assign a probability to each model, and a new population is randomly extracted with such computed probabilities. Thus, better models have higher chance of being included in the next generation.
- *Crossover*: with probability p_c , two random strings (parents) are selected from the mating pool to produce a child. In single point crossover, one crossover point is selected, binary string from beginning of string to the crossover point is copied from one parent, the rest is copied from the second parent. By default, we set $p_c = 0.8$.
- *Mutation*: this step introduces random mutations in the population to ensure that the searching process can move to other areas of the search space. With a probability p_m a randomly selected locus of a string can change from 0 to 1 or from 1 to 0. Thus, a randomly selected variable is either added to or removed from the model. Usually the mutation probability is small, and by default we set $p_m = 0.1$.
- *Elitism*: to improve performance of GA some individuals having particular good fitness may survive through generations. Here, the top model is retained at each iteration.

Computational issues The computational effort needed by GAs depends on the dimension of the search space, on its complexity, and the effort needed to associate a fitness value to each individual in the population. In our case, the latter step involves selecting the “best” clustering model for a candidate subset of variables, which is time consuming. However, since at each iteration some portion of the search space is likely to have already been visited, we can avoid to re-estimate the visited models by saving new string solutions and the corresponding fitness values during the iterations. This allows to save a large amount of computing time.

Table 1 Parameter settings for the scenarios used to generated synthetic data: β defines the correlation of irrelevant variables on clustering variables, whereas Ω is the covariance structure of the noise component. $\mathbf{0}_p$ indicates the $2 \times p$ matrix of zeroes, and $\mathbf{I}_p$ the $p \times p$ identity matrix.

Model 1	$\beta = \mathbf{0}_8$	$\Omega = \mathbf{I}_8$
Model 2	$\beta = \mathbf{0}_8$	$\Omega = 0.5\mathbf{I}_8$
Model 3	$\beta = \begin{pmatrix} 2 & \mathbf{0}_7 \\ 0 & \end{pmatrix}$	$\Omega = \mathbf{I}_8$
Model 4	$\beta = \begin{pmatrix} 0.5 & 0 \\ 0 & 1 & \mathbf{0}_6 \end{pmatrix}$	$\Omega = \mathbf{I}_8$
Model 5	$\beta = \begin{pmatrix} 0.5 & 0 & 2 & 0 \\ 0 & 1 & 0 & 3 & \mathbf{0}_4 \end{pmatrix}$	$\Omega = \text{diag}(\mathbf{I}_2, 0.5\mathbf{I}_2, \mathbf{I}_4)$
Model 6	$\beta = \begin{pmatrix} 0.5 & 0 & 2 & 0 & 2 & 0.5 \\ 0 & 1 & 0 & 3 & 0.5 & 1 & \mathbf{0}_2 \end{pmatrix}$	$\Omega = \text{diag}(\mathbf{I}_2, 0.5\mathbf{I}_4, \mathbf{I}_2)$
Model 7	$\beta = \begin{pmatrix} 0.5 & 0 & 2 & 0 & 2 & 0.5 & 2 & 0 \\ 0 & 1 & 0 & 3 & 0.5 & 1 & 0 & 3 \end{pmatrix}$	$\Omega = \text{diag}(\mathbf{I}_2, 0.5\mathbf{I}_4, \mathbf{I}_2)$

5 Synthetic data examples

In this section we present some empirical results based on the synthetic data examples described in Maugis et al. (2009). Samples of size $n = 400$ are simulated for a 10-dimensional feature vector. The first two variables are generated from a mixture of four Gaussian distributions $\mathbf{X}_{[1:2]} \sim N(\boldsymbol{\mu}_k, \mathbf{I}_2)$ with $\boldsymbol{\mu}_1 = (-2, -2)$, $\boldsymbol{\mu}_2 = (-2, 2)$, $\boldsymbol{\mu}_3 = -\boldsymbol{\mu}_2$, $\boldsymbol{\mu}_4 = -\boldsymbol{\mu}_1$, and mixing probabilities $\pi = (0.3, 0.2, 0.3, 0.2)$. The remaining eight variables are simulated according to the model $\mathbf{X}_{[3:10]} = \mathbf{X}_{[1:2]}\beta + \epsilon$, where $\epsilon \sim N(\mathbf{0}, \Omega)$. Different settings for β and Ω define seven different scenarios, from independence of clustering variables on the other features, to the case of dependence of the irrelevant variables on the clustering ones (see Table 1).

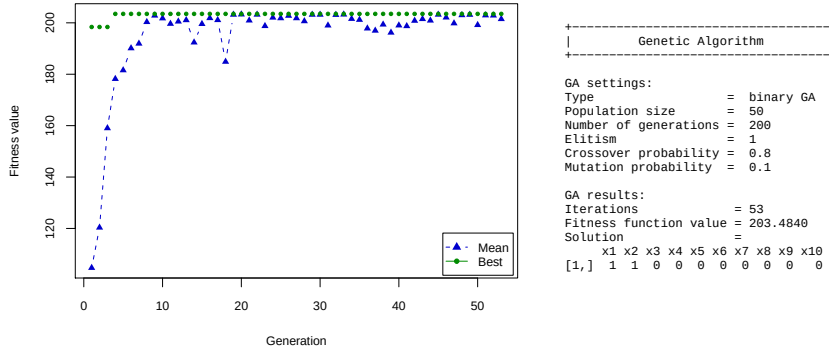


Fig. 1 Summary of a GA run for the simulated data according to model 1.

Figure 1 shows the summary of a typical application of GA search to a simulated dataset according to the first scenario. The green dotted points represent the

best fitness (i.e. BIC) values at each generation, whereas the blue triangles represent the corresponding fitness averages. Each generation is made up of 50 binary strings, each representing a candidate variables subset. Genetic operators and tuning parameters of GA search can be read from the output on the right of the graph. The algorithm soon achieves the optimal solution, and it terminates after there are no improvements in fifty consecutive iterations.

Results from a simulation study for the seven scenarios are shown in Table 2. The accuracy of the subset selection procedure is evaluated by computing both the true inclusion rate (TIR), i.e. the ratio of the number of correctly identified clustering variables to the number of truly clustering variables, and the false inclusion rate (FIR), i.e. the ratio of the number of falsely identified clustering variables to the total number of irrelevant variables. These measures are also known as sensitivity and 1-specificity, and, ideally, we wish to have TIR to be close to 1 and FIR to be close to 0 at the same time. The clustering obtained on the selected subset is evaluated by computing the classification error rate (Error) and the adjusted rand index (ARI; Rand, 1971). Results from Gaussian finite mixture models (GMM) on all the variables, and using the stepwise approach (GMM-Stepwise) of Raftery and Dean (2006) are also included for comparisons.

Table 2 Results obtained on 100 simulation runs for the models described in Table 1. Reported values are averages for the final selected variables subset.

Scenario	Method	Variables	TIR	FIR	Error	ARI
<i>Model 1</i>	GMM	10.00	1.0000	1.0000	0.0472	0.8791
	GMM-Stepwise	10.00	1.0000	1.0000	0.0472	0.8791
	GMM-GA	2.42	1.0000	0.0525	0.0468	0.8800
<i>Model 2</i>	GMM	10.00	1.0000	1.0000	0.0741	0.8572
	GMM-Stepwise	2.24	1.0000	0.0300	0.0427	0.8899
	GMM-GA	2.00	1.0000	0.0000	0.0427	0.8898
<i>Model 3</i>	GMM	10.00	1.0000	1.0000	0.1965	0.7092
	GMM-Stepwise	8.86	1.0000	0.8575	0.0467	0.8810
	GMM-GA	2.40	1.0000	0.0500	0.0463	0.8818
<i>Model 4</i>	GMM	10.00	1.0000	1.0000	0.0728	0.8499
	GMM-Stepwise	7.88	1.0000	0.7350	0.0458	0.8829
	GMM-GA	7.48	1.0000	0.6850	0.0541	0.8619
<i>Model 5</i>	GMM	10.00	1.000	1.0000	0.0848	0.8354
	GMM-Stepwise	5.61	0.995	0.4525	0.0458	0.8826
	GMM-GA	2.58	0.965	0.0812	0.0455	0.8834
<i>Model 6</i>	GMM	10.00	1.0000	1.0000	0.0710	0.8501
	GMM-Stepwise	3.42	0.9850	0.1812	0.0459	0.8819
	GMM-GA	2.20	0.9600	0.0350	0.0459	0.8818
<i>Model 7</i>	GMM	10.00	1.000	1.0000	0.0738	0.8460
	GMM-Stepwise	2.00	0.970	0.0075	0.0454	0.8842
	GMM-GA	2.05	0.965	0.0150	0.0459	0.8829

Overall, the GA search based on BIC criterion (3) is able to select the true clustering variables (on average TIR tends to be equal to 1), while minimizing the inclusion of irrelevant variables (on average FIR is close to 0), except for model 4 where a too

large number of variables is selected. Note that the stepwise method of Raftery and Dean (2006) tends to select a larger number of predictors than our GA search. This is reflected in higher average values for the FIR. Accuracy of the clustering partition is not affected by the inclusion of irrelevant clustering variables when these are independent from the clustering ones (see Model 1). On all the other cases, removing irrelevant variables allows to decrease the classification error.

References

- Banfield, J. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49:803–821.
- Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(7):719–725.
- Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, 28:781–793.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion). *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 39:1–38.
- Fraley, C. and Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, 41:578–588.
- Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631.
- Fraley, C. and Raftery, A. E. (2006). MCLUST version 3 for R: Normal mixture modeling and model-based clustering. Technical Report 504, Department of Statistics, University of Washington.
- Goldberg, D. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Professional.
- Haupt, R. L. and Haupt, S. E. (2004). *Practical genetic algorithms*. John Wiley & Sons, second edition.
- Holland, J. (1992). Genetic algorithms. *Scientific American*, 267(1):66–72.
- Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.
- Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2009). Variable selection for clustering with gaussian mixture models. *Biometrics*, 65(3):701–709.
- McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley, New York.
- Raftery, A. E. and Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178.
- Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, (66):846–850.
- Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:31–38.