# Evolutionary Cluster Analysis

Sandra Paterlini, Tommaso Minerva

Università di Modena e Reggio Emilia
Dipartimento di Economia Politica
Viale Berengario 51, 41100 Modena, Italy
paterlini@unimo.it, minerva@unimo.it

**Abstract.** The determination of the number of groups in a dataset, their composition and the most relevant measurements to be considered in clustering the data, is a high-demanding task, especially when the a priori information on the dataset is limited. Some different genetic approaches are proposed as tools for automatic data clustering and features selection. They differ in the adopted codification of the grouping problem, not in the evolutionary operator and parameters. Two of them deals with the grouping problem in a deterministic framework. The first directly approaches the grouping problem as a combinatorial one. The second wants to determine some relevant points in the data domain to be used in clustering data. A probabilistic framework is then introduced with the third one which wants to specify the statistical model from which data are assumed to be drawn. The evolutionary approaches are compared with respect to classical partitional clustering algorithms on simulated data and on Fisher's Iris dataset.

**Keywords:** clustering data, genetic algorithms, string code, genetic feature selection.

## 1. Introduction

The determination of the number of groups in a dataset, their composition and the most relevant measurements to be considered in clustering the data, is a high-demanding task, especially when the a priori information on the dataset is limited. In fact, to determine the number of the $g$ most representative groups and their composition of a dataset composed by $n$ items with $p$ measurements involves a time-unfeasible combinatorial effort when $n$ and g are not small numbers and the effort is also bigger if the number of groups is not known a priori (Liu G.L., 1968).

Cluster analysis deals with data pattern detection by forming homogeneous groups inside the whole dataset by determining the groups' composition, the number of groups in the dataset, the relevant features to be used in forming the groups and a measures of similarity / dissimilarity among the items and the groups. Different cluster algorithms have been developed and research is still ongoing. Among the partitional algorithms, the so called *k-means*, *EM* (*Expectation Maximization*) and *fuzzy c-means* algorithms are widely known. They start with a random choice of the

initial seeds with respect to which, computing a measure of similarity, they determine the belonging of each observation to a specific group. Hovever, starting from random seeds, make them not always converge to the global optimum. To overcome this shortcoming, we focus our attention on evolutionary clustering algorithms based on genetic algorithms (Holland 1975).

In this paper we introduce and discuss three different evolutionary approaches based on genetic algorithms. The capability of genetic algorithms to converge to the global optimum within an elitist schema (Rudolph 1994) supports their validity in identifying the best number of possible groups in a dataset and the more relevant measurements to be used in forming groups in an unsupervised learning mechanism. We assume, in fact, no a priori information is available. The grouping is determined through statistical criteria that aims at minimizing the dispersion *within* the groups and, contemporarily, maximizing the dispersion *between* the groups or maximising the likelihood of the statistical model underlying the data. Different genetic approaches to the clustering problem have been proposed in literature. V.V.Raghavan and K.Birchand (1979) were the first to propose to use the genetic algorithms to directly allocate the items in one of the $g$ groups, which have been supposed to be present in the dataset. A fitness function directed to minimize the squared error is used to determine the optimal composition of the groups in the dataset. Since then, different genetic codification and fitness functions have been tested to solve clustering and pattern recognition problems (see Bandyopadhyay S., Murthy C.A. 1998, Srikanth R. et all 1995; Baragona, Calzini, Battaglia, 1999). Moreover, genetic algorithms have been used not only to tackle directly the clustering problem but also through the development of hybrid algorithms, that is in conjunction with other standard localized clustering techniques, such as k-means, fuzzy c-means, artificial neural network in order to better their performance (Tseng 2001).

Pre-processing data could be essential to remove noise and outliers and to make the clustering algorithms determine more homogeneous groups. To consider the whole dataset could prevent to discover hidden patterns and underlying structure in data. Moreover, to deal with huge dataset increases the computational effort and the efficiency of clustering algorithms. Cluster analysis could be mislead by highly correlated measurements and the presence of noise and outliers. Genetic approaches have already shown to be capable to deal with the dimensionality reduction problem and the choice of the most relevant measurements in a promising way (Raymer et al. 1997, Kim Y., Street W.N., Menczer F. 2000). We tackle this problem modifying the genetic code to make them automatic selecting the most relevant measurements to be used in clustering data.

The three genetic approaches, we introduce, use different codification to tackle the clustering problem. The first, GAIE (*Genetic Algorithm for Items Evolution*), has a population of individuals which directly allocate each observation to a specific group considering the clustering problem from a combinatorial point of view. On the contrary, the second, GAME *(Genetic Algorithm for Medoids Evolution)*, and the third, GAPE (*Genetic Algorithm for Parameter Evolution*), not only exploit the genetic algorithm to determine the optimal partition, but also to get additional information about relevant grouping points in the data domain. GAME composes the groups after determining the medoids of the possible groups and the belonging of each observation to the group with minimum euclidean distance from the group's

medoid. GAPE tackles the clustering problem in a probabilistic framework, determining the parameters of the model from which it is assumed the data are drawn.

In section 2 we introduce the three algorithms while in section 3 the fitness functions we used to drive the evolutionary process. Finally, in section 4 , we discuss some empirical results and draw some conclusions.

## 2. Genetic Codes

Genetic algorithms are stochastic algorithm which have been widely used in different fields because of their capability of searching the whole solution domain and of dealing with complex optimization problems. A genetic algorithm is composed by a population of individuals where each individual represents the map of a possible solution of the problem the researcher is dealing with. The population is evolved by genetic operators driven by a fitness function able to measure the degree of optimality of the individuals through the generations. The best individual of the last generation represents the encoding of the best solution of the problem to be solved. The genetic operators (selection, crossover, mutation, elitism) are inspired to natural biological processes, driven by the Darwinian principle of the survival of the fittest individual through the generations. Genetic algorithms have numerous properties. Rudolph (1994) has shown how a genetic algorithm converges to the global optimum within an elitistic schema, that is when the best individual of a generation is re-inserted in the population of the following generation. Moreover their flexibility, their parallel and straight implementation and their capability of exploring the whole search space support their validity and efficacy in numerous applications, even if they are often criticized because of their sensitivity to the control parameters (for example: the number of individuals, the number of generations, the mutation rate, the crossover rate,…).

GAIE (Genetic Algorithm for Items Evolution) algorithm, which is the first proposal about using genetic algorithm to solve partitional problems (Raghavan V.V. and Birchand K. 1979) directly allocates each observation of a dataset *nxp* in one of the *g* groups. Each individual string has length equal to the number of observations in the dataset and each cell can contain an integer value in the interval [*1,g*]. But this codification is redundant in mapping the solution, increasing the computational time required for convergence (for example: 222111 string groups the data in the same ways 111222 string does, but they are different from the algorithmic point of view).

In the second alternative approach we use the genetic code in order to determine some relevant points, called medoids to be used in grouping data. The *GAME* algorithm assumes that each individual is formed by *p x g* cells, which represents the codification of the possible values of the medoids' measurements. Each group of *p* cells identifies the medoid coordinates in the $R^p$ space of the measurements. The *g* groups of cells represent the *g* medoids of the clusters. Each cell can assume a real value between the lower bound and upper bound of the whole series of the corresponding item measurements to which the medoid value of the cell is referred to. The algorithm is inspired to Forgy's approach of clustering (Forgy E.W. 1965). Once determined the possible values of the medoids of the *g* groups, the algorithm computes the euclidean distances between each measurement of each observation

with respect to the corresponding values of the *p x g* medoids and determines the belonging group. Each item belongs to the cluster with minimum euclidean distance with respect to the clusters' medoids among all the computed distances between the considered item's measurements and the medoids' values of all possible groups.

Tests, comparing the speed of convergence of GAIE and GAME algorithms with respect to the same fitness function, have shown that GA ME and GAIE converge to the same global optimum (if we use the same evolutionary schema and parameters), requiring a different amount of computational time. GAME converges more quickly (speed ratio 1 to 10) and moreover, exploring the whole search space give additional information through the identification of the medoids of the dataset, which can be considered as separation points between different groups.

A probabilistic and inferential framework is introduced with the third approach. GAPE (Genetic Algorithm for Parameters Evolution) uses the genetic codification to estimate the parameters of the statistical model which is supposed to be underlying the data and then it allocates each observation to the group with respect to which it has a higher probability of belonging. We considered the data as realization of random variables. We use the genetic algorithm to estimate the parameters in a fixed-classification model (Bock H.H. 1996). This model assumes, for a fixed number of groups *g,, (G1,G2,..Gg)*, a known parametric density family *f(.,è)* such that $X_k \sim f(.,\grave{e}_i)$ for all $k \in G_i$ , $i=1,..g$ where *g* is unknown and the parameters vector È=( è1,…,èg). It has been assumed that the parametric density function corresponds to the multivariate normal distribution and the genetic code is used to estimate the vector of the mean values and the diagonal elements of the variance-covariance matrix. Then, each string maps the parameters $\grave{e}_k=(\grave{i}_k, \ _k)$.

Banfield and Raftery (1993) propose to decompose the covariance matrix using eigenvalues and eigenvector and to express it as $\sum_k = l_k D_k A_k D_k^T$, where $D_k$ indicates the orthogonal matrix of the eigenvectors and determine the orientation of the principal components, $A_k$ is a diagonal matrix with elements proportional to the eigenvalues of $_k$ and determine the contours of the density functions and $\ddot{e}_k$ is a scalar that specific volume of the ellipsoids. We assume that $\sum_k = l_k \mathrm{I}$ and we introduce a genetic code such that three different structures of the matrix of variance are allowed. In the first case the matrix of covariance is constant among the groups and the measurements, in the second case it is constant just among the measurements and in the third case each group could have a different variance among the groups and the measurements.

The density value of each observation is computed with respect to the models determined by the genetic algorithm for each group. Each observation is attributed to the group with respect to which it has maximum density value. The fitness criteria to be used is directed to minimize the negative form of the log-likelihood of the fixed classification model:

$$F(g,\Theta) = -\sum_{i=1}^{n}\sum_{k=1}^{g} z_{ik} \log( f_{g_i}(x_i \mid q_k) \qquad \textbf{(1)}$$

where $z_{ik}$ is equal to one if observation *i* belongs to group *k* or zero otherwise and where $\tilde{\mathbf{a}}=(\tilde{a}_1,\ldots\tilde{a}_n)$ are labels such that $\tilde{a}_i=k$ if $x_i$ belong to groups *k*. After processing each individual, genetic operator are applied to evolve the population to determine the best individual with associated minimum fitness value.

To use the genetic algorithm allows us not just to determine the optimum partition with respect to the adopted fitness criteria, but also to determine the parameters of the structure underlying the data and a probabilistic measure of the belonging of each observation to a specific group.

To determine the partition of a big dataset not only requires a lot of computational effort, especially when hierarchical cluster algorithms are used, but also the analysis could be mislead by the presence of highly correlated variable and noise in the data. The genetic codification of the three algorithms have been modified to tackle this problem. *p* binary cells, one for each measurement, have been added to each individual in all the three algorithms. Then the data matrix to be processed will include just the measurements in correspondence of the cells which contain unitary values. The automatic data reduction has shown to reach a smaller number of misclassified items in the simulated dataset and in the Fisher's Iris data.

## 3. Fitness criteria

As previously described the GAPE algorithm uses as fitness criteria to drive the evolutionary process the negative form of the log-likelihood as reported in equation 2. It can be shown that in the case of constant variance matrix among the groups this is equivalent to minimize the trace of the matrix *within* the groups (Banfield J.D, Raftery A.E., 1993).

GAME and GAIE algorithms have been tested using different fitness criteria which aim to minimize the dispersion *within* the groups and/or to maximize the dispersion *between* the groups. In fact the total dispersion in the dataset could be decomposed as **T=B+W**, where **T** indicates the *total* scatter matrix of the n observation, **W** the pooled-*within* groups scatter matrix and **B** the *between* groups scatter matrix.

Different fitness functions (Calinski T., Harabasz J., 1974, Marriott F.H.C., 1982, Ricolfi L. 1992) have been used to drive the evolutionary operators towards the identification of the optimal partition of the dataset. If we suppose that the number of groups is known a priori, GAIE and GAME could be used to determine the optimum value of the following fitness criteria and the associated best partition of the dataset: *min(trace(W)),max( trace(B/W)), max(trace(B/T)), min(det(W)/det(T)).*

However, the number of groups is usually not known a priori. We wanted the algorithms to determine also the best number of natural groups in the dataset. Then, two different fitness functions , which include a penalisation factor depending on the number of groups , have been used to determine both the composition of the clusters and the best number of groups in the dataset.

They are respectively: $min(g^2 det(W)/det(T))$ (MC, *Marriott's criterion*, 1982) and $max(\{tr(B)/(g-1)\}/\{tr(W)/(n-g)\})$ (VRC, *Variance Ratio Criterion*, Calinski T., Harabasz J., 1974). The selection of the number of groups, *g*, has not been included within the evolutionary process. An iterative approach, limited to the maximum available number of groups has been proposed. The algorithm stops when increasing the number of groups, the fitness value of the best individual in the last generation is greater than the optimal fitness value of the previous iteration. This mixed iterative-evolutionary schema implies the exploration of disjoint solution subspaces where the number of groups is fixed.

## 4. Results and conclusions

GAIE, GAME and GAPE have been at first tested on simulated dataset composed by random observations drawn from multivariate normal distribution with different location and equivalent and not covariance structure. When data do not contain overlapping clusters, the three algorithm identify correctly the real classification. The number of misclassified items increases when the parameters used to specify the distribution from which to draw the data could lead to generate overlapping clusters. Table 1 shows the average error with respect to the real classification on simulated datasets. The average error refers to ten different simulated dataset of two hundred observations with four measurements. Fifty observations have been generated from each multivariate normal distribution respectively with parameters: $m_1=[1,1,1,1]$, $S_1=I$; $m_2=[5,5,5,5]$, $S_2=2I$; $m_3=[9,9,9,9]$, $S_3=3I$; $m_4=[13,13,13,13]$, $S_4=4I$. For each dataset 500 simulations have been performed. Overlapping clusters may be formed. We compared our evolutionary approaches with standard techniques. GAPE, GAME and GAIE outperforms the standard *k-means* algorithm, and the *Expectation Maximization (EM)* algorithm with diagonal and full specification of the covariance matrix (see table 1). Moreover, GAME and GAIE converges to the same optimal value while the other algorithms, except the fuzzy c-means and the EM with spherical structure, tend to fall in local minima leading to variability in the number of misclassified items. The usage of GAPE algorithm has been useful to determine the specification of the underlying model. In fact, the best individual reports values of the parameters to be estimated which are very near to the real values used to generate the data. The genetic approach not only allows to determine the optimal partition of the data, but also gives further inside about the structure of the model underlying the data.

| Clustering Algorithm | Average error. |
|---|---|
| GAPE | 2.3% |
| GAIE/GAME (with MC as Fitness Function) | 1.9% |
| GAIE/GAME (with VRC as Fitness Function) | 1.5% |
| K-means | 4.4% |
| Fuzzy c-means | 0.9% |
| EM diagonal | 4.6% |
| EM full | 5.9% |
| EM spherical | 0.7% |

**Table 1.** Cluster algorithms performance when data are generated from multivariate normal with parameters: $m_1=[1,1,1,1]$, $S_1=I$; $m_2=[5,5,5,5]$, $S_2=2I$; $m_3=[9,9,9,9]$, $S_3=3I$; $m_4=[13,13,13,13]$, $S_4=4I$. The average error has been evaluated on 10 different simulated datasets. For each dataset 500 runs were performed so the average is on 5000 different runs.

Fisher's Iris dataset is a well-known target dataset to be used in testing the validity of new clustering algorithms. Data are collected from three different species of iris flower, where observations from just one specie have clearly distinctive features. It is composed by 150 observation with four measurements each. Table 2 reports the number of misclassified items in correspondence of our evolutionary approaches compared with classical approaches and with results reported in literature. We also

report results related to different fitness criteria. The results show that the evolutionary algorithms are capable to identify the correct belonging of each observation but three items, which it is the best result up to now reported. Automatic data mining, as the third column shows, allow to reduce the number of misclassified items. Moreover, respectively in the fourth and fifth columns, the best fitness values reached by the GAIE/GAME algorithm and the one in correspondence of the real correct classification are reported.

| | Minimum number of misclassifed items | Average number of misclassified items on 500 runs | Misclassified Items with Automatic data-mining | Best FV | Real FV |
|---|---|---|---|---|---|
| GACE/GAIE *1/tr(B/W)* | 3 | 3 | 3 | 0.029 | 0.038 |
| GACE/GAIE *MC* | 3 | 3 | 3 | 0.198 | 0.210 |
| GACE/GAIE *VRC* | 16 | 16 | 8 | 0.0018 | 0.0021 |
| GACE/GAIE *trW* | 16 | 16 | 6 | 7885 | 8930 |
| GAPE Ók=ëI | 16 | 24.5 | 6 | 1635 | 1693 |
| K-means | 16 | 25.9 | --- | | |
| Fuzzy c-means | 16 | 16 | --- | | |
| EM-spherical | 16 | 16 | --- | | |
| EM-diag | 9 | 23.7 | --- | | |
| EM-full | 5 | 17.2 | --- | | |
| Friedman, Rubin, 1967 tr(W/B) | 3 | --- | --- | --- | --- |
| Friedman, Rubin, 1967 det(T)/det(W) | 3 | --- | --- | --- | --- |
| Fraley and Raftery, 1999 EM | 5 | --- | --- | --- | --- |

**Table 2.** Fisher's Iris data. Comparison among the evolutionary algorithms, the standard partitional algorithms and the literature reported results in grouping the Iris dataset. The number of misclassified items with automatic data mining and not, the best fitness values reached by GAIE/GAME/GAPE and the fitness values in correspondence of the real classification with no automatic data mining are reported. (FV=Fitness Value; the Fitness Functions we used are reported in italics)

It should be noticed that GAME/GAIE identify a partition with associated smaller fitness values. The choice of the fitness criteria to be used is therefore crucial. A deeper look inside the data shows that the three misclassified items are more homogeneous with respect to the group they are assigned by the clustering algorithms than to the group they actually belong. The possible presence of errors in collecting data and the existence of anomalous data should be considered when looking at the partition given by the algorithms. If we compare evolutionary with classical approaches (see Table 2) we note that standard partitional clustering algorithms, which start from the random choice of the initial seeds, as GAME and GAPE do, lead at most to misclassify 5 observations and how they, except the fuzzy c-means and the

EM with spherical covariance structure, can easily fall in local minima. The average number of misclassified items in 500 runs is well above the minimum number of misclassified items, showing that they are not stable in global convergence. On the contrary the evolutionary clustering algorithms show a very stable and robust convergence reaching the same state in all the 500 runs starting from different random seeds. The iterative evolutionary procedure has also determined as optimal number of groups the number of the species in the dataset in correspondence of the Variance Ratio Criterion.

The comparison between our proposed evolutionary clustering with standard clustering algorithms shows the validity of the approach in developing more efficient clustering algorithms which have strong convergence properties. These results encourage further research in improving the described algorithms and in testing and building new statistical criteria to be used. Moreover, analysis in a probabilistic framework could be further developed in testing different possible models coming from different distributions and in validating criteria, such as the BIC criterion, to automatically detect the number of groups, within an iterative approach.

## References


Bandyopadhyay S., Murthy C.A., Pal S.K., "Pattern classification using genetic algorithm: Determination of H", *Pattern Recognition Letters* 19,1171-1181, 1998.

Banfield J.D:, Raftery A.E., "Model-Based Gaussian and Non-Gaussian Clustering", *Biometrics* 49, 803-821, September 1993.

Baragona R., Calzini C., Battaglia F., "Genetic algorithms and clustering: an application to Fisher's iris data", *Advances in Classification and Data Analysis* , Springer, pp.65-68, 1999.

Bock H.H., Probabilistic models in cluster analysis , *Computational Statistics & Data Analysis* 23, pp.5-28, 1996.

Calinski T., Harabasz J., "A dendrite method for cluster analysis", *Communication in Statistics*, 3(1), pp.1-27, 1974.

Forgy E.W ., "Cluster Analysis of Multivariate Data: Efficiency versus Interpretability of classification", *Biometrics*, 21, 768-9, 1965.

Fraley C., Raftery A.E., "MCLUST:software for model-based cluster and discriminant analysis", *Journal of Classification*, 16, 297-306, 1999.

Friedman H.P. and Rubin J., "On some invariant criterion for grouping data", *Journal of the American Statistical Association* 63, 1159-1178, 1967.

Kim Y. ,Street W.N., and Menczer F. Feature selection in unsupervised learning via evolutionary search, in Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-00), pages 365-369, 2000.

Holland J.H., *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Harbor, 1975.

Liu G.L., *Introduction to combinatorial mathematics*, McGraw Hill, 1968.

Marriott F.H.C., "Optimization methods of cluster analysis", *Biometrics*, 69, 2, pp.417-422, 1982.

Raghavan V.V., Birchand K., " A clustering strategy based on a formalism of the reproductive process in a natural system", *in Proceedings of the Second International Conference on Information Storage and Retrieval*, 10-22, 1979.

Raymer M.L. et AAVV, "Dimensionality Reduction using Genetic Algorithms", *IEEE Transaction on Evolutionary Computation*.

Ricolfi L., HELGA *Nuovi principi di analisi dei gruppi*, FrancoAngeli s.r.l., Milano, Italy,1992.

Rudolph G., "Convergence analysis of canonical genetic algorithm", *IEEE Transactions on Neural Network*, 5(1):96-101,January 1994.

Srikanth R., George R., Warsi N., Prabhu D., Petry F.E., Buckles B.P., "A variable-length genetic algorithm for clustering and classification", *Pattern Recognition Letters* 16, 789-800, 16, 1995.

Tseng Y.L. e Yang S.B., " A genetic approach to the automatic clustering problem" , *Pattern Recognition*, Vol.34 (2), pp.415-424 (2001).