Introducing Interactive Evolutionary Computation in Data Clustering

Anna Russo[1], Onofrio Gigliotta[1](✉),
Francesco Palumbo[1], and Orazio Miglino[1,2]

[1] Natural and Artificial Cognition Laboratory, University of Naples Federico II,
Naples, Italy
onofrio.gigliotta@unina.it
[2] Institute of Cognitive Sciences and Technologies, CNR, Rome, Italy

Abstract. Data clustering consists in finding homogeneous groups in a dataset. The importance attributed to cluster analysis is related to its fundamental role in many knowledge fields. Often data clustering techniques are the *ghost host* of many innovative applications for a wide range of problems (i.e. biology, marketing, customers segmentation, *intelligent machines*, machine translation, etc.). Recently, there is an emerging interest in *Data Clustering* community to develop bio-inspired algorithms in order to find new methods for clustering. It is widely observed that bio-inspired algorithms and the Evolutionary Computation (EC) techniques reach solutions similar to others computational approaches but using a bigger computational power. This limitation represents a concrete obstacle to an extensive use of Evolutionary (or bio-inspired) approach to data clustering applications. In the present paper we propose to use Interactive Evolutionary Computation (IEC) techniques where a human being (the breeder) selects *Cluster configurations* (genotypes) on the basis of their graphical visualizations (phenotypes). We describe a first version of a software, called Revok, that implements the IEC basic principles applied to data clustering. In the *conclusion* section we outline the necessary steps to reach a mature IEC tool for data clustering.

Keywords: Interactive Evolutionary Computation · Data mining

1 Introduction

Data clustering consists in finding homogeneous groups in a dataset. A very general Data Clustering definition is the following: given a set of n points in a p dimensional space, cluster analysis aims at grouping data into k groups that are maximally homogeneous within each one and maximally heterogeneous between them [6]. The importance attributed to cluster analysis is related to its fundamental role in many knowledge fields. In fact, on natural side, human and animal learning/adapting processes produce clusters of objects and/or actions that we refer to as concepts, mental categories, perceptual patterns, behavioural strategies and so on. On the other hand, data clustering techniques are the *ghost host*

C. Pizzuti and G. Spezzano (Eds.): WIVACE 2014, CCIS 445, pp. 26–36, 2014.
DOI: 10.1007/978-3-319-12745-3_3

of many innovative applications for a wide range of problems (i.e. genetics, marketing, customer segmentation, *intelligent* machines, etc.). Because of their basic importance in different fields, methods for data clustering have been independently developed in very different scientific contexts where a variety of methods have been proposed [6]. Recently, there is an emerging interest in Data Clustering community to build-up algorithms using bio-inspired techniques (Neural Networks, Genetic Algorithm, Evolutionary Computation, Ant algorithms, Swarm algorithms, etc.) [5,17,20,21]. More specifically, there is a consolidated scientific literature concerning EC techniques applied to data clustering [1]. In general, bio-inspired techniques, as all clustering methods, measure and optimize the efficiency of a given algorithm on the basis of an evaluation function. In the case of EC techniques it corresponds to the fitness function. The novelty of EC techniques, respect to more traditional approaches, is represented by the possibility to have a parallel competition between many different clustering variants. The fittest variants are selected for reproduction. The selected individuals generate new variants ready for a new evaluation (or evolution) phase. The evaluation, selection and reproduction processes could be iterated until an efficient solution emerges. It is widely observed that EC techniques produce results similar to other approaches but they require bigger computational resources. Perhaps, all computational approaches (either bio-inspired and traditional) have reached ceiling performances that cannot be overcome anymore. If we adopt a natural cognition point of view this limitation could have an explanation that leads to possible new approaches in Data Clustering. In psychological terms multidimensional data clustering is a kind of categorization based on the analysis of explicit information and quantitative variables (dimensions) that describe a given phenomenon. On the contrary, the best clustering systems for multidimensional domains actually known, the human beings, do not work only on the basis of explicit information. They use also (or mainly) latent and implicit information captured by (neuro)cognitive mechanisms that work under of consciousness threshold. Psychological literature refer to these categorization mechanisms (i.e. sensory analysis, perception, mental reasoning, etc.) as Cognitive Unconscious [9]. According to these research findings all cognitive representations emerge from a dynamic interplay between unconsciousness and consciousness processes. Moreover, it is showed that Cognitive Unconscious mechanisms and mental schemas are very hard to report at conscious level. The interplay of conscious and unconscious of our neurocognitive structures is particular evident in the case of human experts of some specific domain. The human experts are usually trained for many years to recognize (categorize) natural phenomena on both explicit and latent information. For example, take into consideration the classification abilities of a doctor that make a diagnosis. In this condition a doctor reaches his decision (diagnosis) integrating explicit biological indexes captured by technological tools with implicit knowledge that is produced by his/her professional experiences and training. In other words, humans are extremely able to apply sophisticated clustering algorithms but they are not able to explicit them. This gap between explicit and implicit level of cognition could explain why the categorization performance of artificial systems are poorer compared to those performed by human beings. In this work, we propose to improve EC clustering techniques introducing some qualitative evaluation in selection phase. In concrete terms we propose to apply the paradigm of Interactive Evolutionary Computation (IEC) [14,18] to multidimensional data analysis domains. IEC is a well-know technique used to perform optimization through the selection process of a human evaluator. In the framework of data clustering, a genotype describes a clustering configuration (a data partition) and/or its parameters (number of clusters, clusters size, clusters centroids, etc.); the environment is a dataset extracted from a data universe for a given problem; the phenotypic traits are quantitative (i.e.: performances indexes) and/or qualitative (i.e.: pictorial data aggregations) representations of organisms outputs. Individuals are selected for reproduction taking in account phenotypic traits. Following the metaphors, we have now three possibilities:

1. selecting the organisms on the basis of quantitative traits
2. selecting the organisms on the basis of qualitative traits
3. selecting the organism combining qualitative and quantitative traits

To our knowledge, the first option is almost the only method used in Evolutionary Computation techniques applied to data clustering. In concrete terms, a given performance criteria or fitness function is used to automatically select the organisms and the EC techniques are used to minimize or maximize that fitness measure. The other two selection procedures are (to our knowledge) completely new in the field of data clustering while are currently used in other EC applications. These selection techniques require the intervention of a human operator that interacts with the Artificial Evolution process. In other words, a *Breeder* analyses the qualitative phenotypic traits and decides which *organism* is ready for the reproduction. This procedure is usually called Interactive Evolutionary Computation (for a review see [4]; for a recent application to robotics see [13]). To our knowledge there have been only two attempts to introduce IEC in clustering processes. In the first, IEC has been used in order to retrieve images from a database according to a set of features extracted by a set of images presented and evaluated by a human evaluator [10]. In the second paper, IEC has been used in exploratory cluster analysis by showing 2d visualization maps of a set of SOMs [19]. In both works the proposed solutions were suited to specific problems while in the present paper we introduce a general solution feasible for a wide range of clustering problems.

2 Clustering and GA

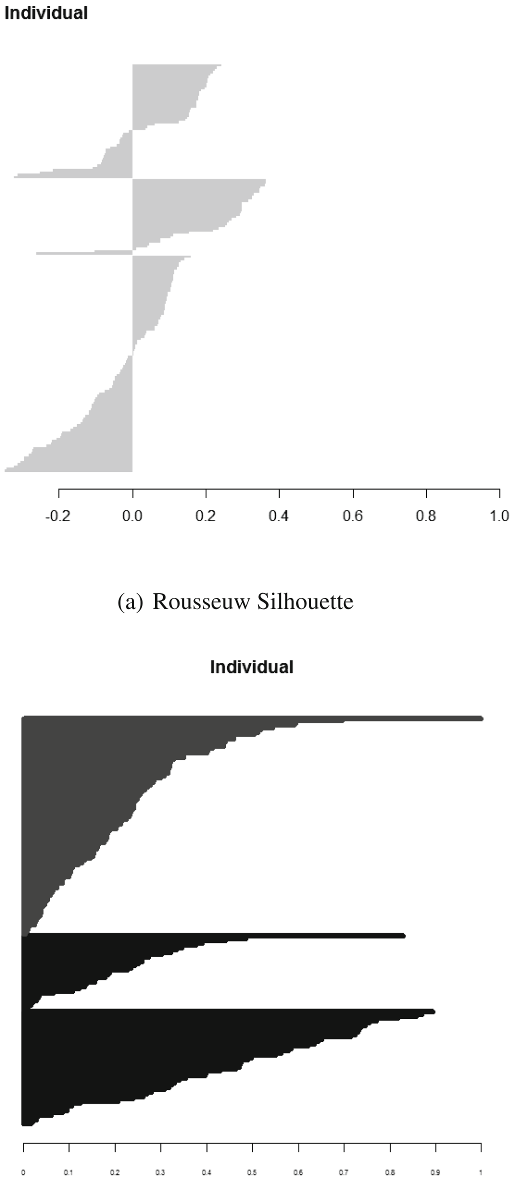
Clustering is an unsupervised classification technique whose objective is to partition a set of data points into groups or clusters. Resulting clusters should show homogeneity within groups and heterogeneity between groups so that the distance (or any similarity measure) between data points belonging to different clusters be greater than the distance between data points belonging to the same cluster [8]. The most known clustering algorithm is the K-means, which aim is to minimize the within-cluster sum of squares. Two problems affect the K-means algorithm: the first is related to the number of clusters, a choice that a user has to make a priori. The second problem refers to the sensibility to the initial conditions. Different initial conditions led to different partitions by stopping the algorithm in local optima. In order to overcome such kind of problems the use of Genetic Algorithms (GA) has been proposed for clustering data. In particular, GKA (Genetic K-means algorithm), FGKA (Fast GKA) and IGKA have proven to be able to find a global optimal partition given a prefixed number of clusters [17] while Bandyopadhyay and Maulik [2] have used GA to discover automatically also the number of clusters. Clustering, as stated above, is an unsupervised technique that partitions data in order to minimize variance within groups and maximizing variance between groups. This statistical principle is well grounded on the information contained in the data but completely ungrounded on the needs and on the perceptual ability of who have to make use of such partition. In order to solve this issue, in this paper, we introduce an interactive GA clustering. The use of evolutionary search along with the user guide, especially during an exploration phase, can lead to the discovery of interesting and meaningful solutions [14].

3 Interactive GA

In this paper we present an interactive evolutionary algorithm to clustering datasets according to users' needs and perceptual skills. In particular, it permits a user to explore the search space in order to find a suitable partition without specifying the number of clusters. During the initialization phase, a random population of n partitions is created, where n is determined by the users. Each partition is encoded in a genotype string and is displayed to the user through its phenotype: a graphical representation of the found clusters. For each iteration the user is asked to choose its favourite g partitions. Then the g chosen genotypes undergo to a reproduction and mutation process in order to recreate a new whole population. The evolutionary process ends when the user is satisfied by a specific partition.

3.1 Genotype

Each genotype encodes a list of c centroids of $d+1$ dimensions (genes), where d is the number of attributes of a given dataset while the further binary dimension is used to encode the centroid activation. In order to facilitate the clustering process, data points within the dataset are ranged between [0,1]. Only one genetic operator is applied to the genotype: mutation. Mutation is achieved by adding, with a probability p, a value extracted from a normal distribution to each gene. For each genotype then, a partition is created by assigning a data item to the closest centroid.



(a) Rousseuw Silhouette

(b) Probabilistic Silhouette

Fig. 1. Silhouette graphs of the same individual (three clusters). In (a) the Rousseuw method in (b) the probabilistic method.

3.2 Phenotype

The genotype-phenotype mapping represents a crucial factor because the phenotype has to describe a solution as clear as possible to the users, even those that are not familiar with sophisticated statistical knowledge, moreover a suitable graphical representation has to allow a user to exploit his perceptual skills in order to discover relations and clusters. For this reasons we decided to use the Silhouette method to graphically represents a partition expressed by a genotype. Originally described by Rousseeuw [16], the Silhouette, computed on distance measures, provides a compact information about how well data items fit into their assigned clusters. Silhouette comes in two formats: based on distance measures [16] and based on a posteriori probability [12]. In the first case, each data item is represented as an horizontal line whose maximum length is 1. A positive length indicates that the item has been correctly clustered while a negative length indicates that the data item is in the wrong cluster (see Fig. 1a). In the probability based Silhouette there are non negative values. This features produces more linear and simple graph (see Fig. 1b). Through a pilot test we observed that the probabilistic approach is more clear in the moment of choice. The principal reasons are:

- the plotted values are positive (a posteriori probability of membership to a specific cluster), in this way the cluster separation is more definite and the user choice is more rapid as we see in Fig. 1. The negative part of classic Silhouette graph is misleading to recognize the number of classes.
- the concept of the probability of membership to a cluster is more intuitive compared to the average of the distances of the classical approach, therefore it is more understandable even in a non-expert user.

4 Revok: Simple Application for IEC in Data Clustering

The IEC data clustering has been implemented in Revok, an application developed in R [11][1] (an interpreted language suitable for statistical purposes) for clustering data through the feedback of a human evaluator. Revok allow users to load their datasets and to set IEC parameters such as, for example, the number of displayable individuals/partitions and mutation rate.

Once started, Revok presents users a series of graphical representations of the partitions found according to the GA. The graphical representation we have chosen for its mathematical properties and its understandable form (humans are very good in perceptual categorization) is the Silhouette calculated according to the method described in [12] by using a posteriori probability computed according to the following equation [3]:

[1] The R environment can be downloaded from the internet along with its complete documentation for free. It is an integrated set of software assets for the manipulation and visualization of data. It allows object oriented programming which is particularly suitable for iterative algorithms.

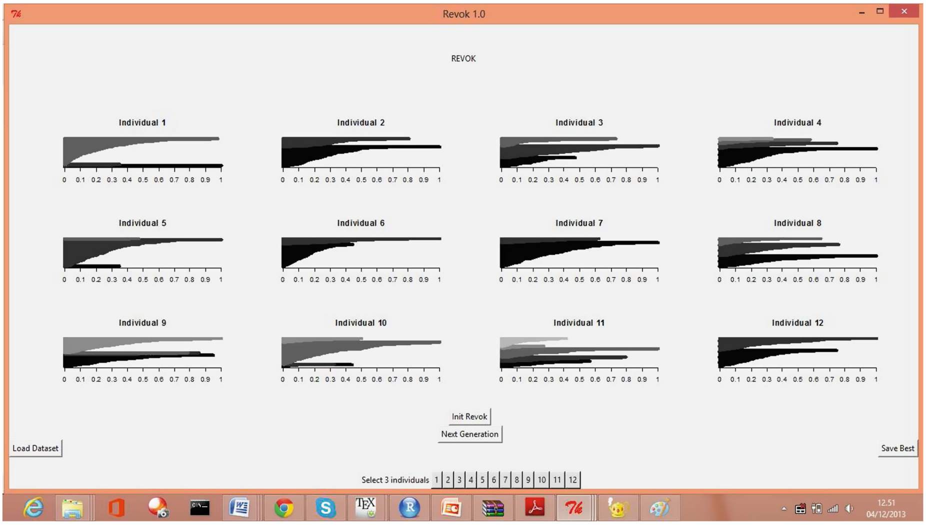


Fig. 2. Display window

$$P_k(x) = \frac{\prod_{j \neq k} d_j(x)}{\sum_{t=1}^{K} \prod_{j \neq t} d_j(x)} \tag{1}$$

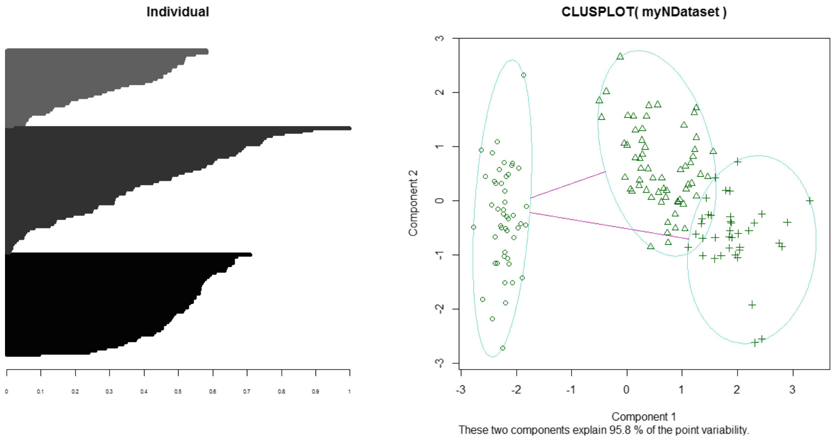
Where $d_j(x)$ is the distance of the data item x from the j_{th} centroid and K is the number of centroids. The computed Silhouette is then presented to the user in the display window (see Fig. 2).

In our simple evolutionary algorithm, as previously mentioned, each individual corresponds to a single partition which is the result of a different arrangement of centroids in the solutions space.

The individuals of the next generation are generated from those chosen by the user through a process of mutation. Mutation is applied by adding to each gene an independent random Gaussian number $N(0, \sigma)$, where σ is defined by the user. Revok has been designed to allow users, without any particular mathematical skill, to carry out their own clusterization relying only on their perceptual skills. For this reasons we tested Revok with undergraduate students in psychology, asking them to partition two UCI datasets, Iris and Haberman, following only the request to obtain graphs with a low number of clusters and long horizontal lines. Making use of the mentioned datasets allow us to compare revok users' partitions with those provided with the UCI datasets.

5 Results

Revok has been tested with two datasets: *Iris and Haberman* downloaded from the UCI machine learning repository (http://archive.ics.uci.edu/ml/). In particular we asked to 10 undergraduate students in psychology to partition the two



(a) Silhouette graph

(b) Cluaplot graph

Fig. 3. The Silhouette and Clusplot graph for the Iris dataset after five generations with Revok

datasets following only an aesthetic criterion. The following plotted results refer to the typical partitions achieved by those students.

The characteristics of Iris dataset are:

- Multivariate and real dataset
- Number of instance: 150
- Number of attribute: 4
- Number of clusters: 3

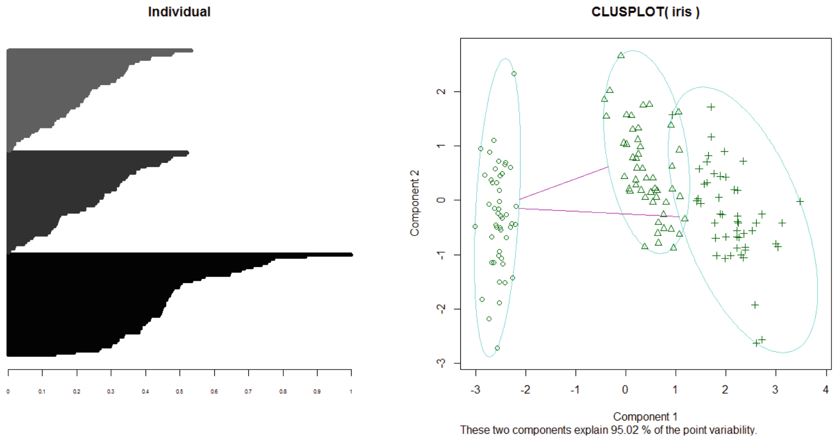
The Fig. 3 is a typical result after five generations with the following parameters:

- maximum number of clusters: 6
- mutation rate: 0.2
- σ: 0.1
- number of parents: 1

The figure reports the silhouette and the clusplot graph. The clusplot graph creates a bivariate visualization of the clustered data [15]. All observations are represented by points in the plot, using principal components (PCA). Clusters, then are represented by the ellipses. To represents the distances between the clusters, the clusplot method permits to draw segments of the lines between the cluster centers as reported in our graphs.

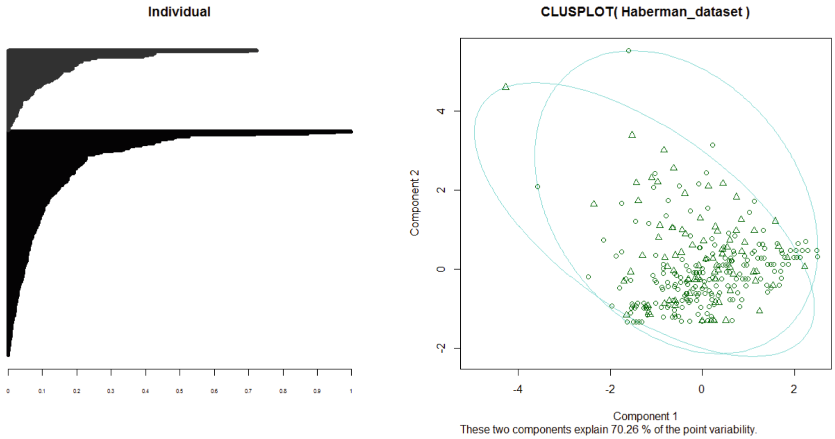
The characteristics of Haberman dataset are:

- Multivariate and integer dataset
- Number of instance: 306
- Number of attribute: 3
- Number of clusters: 2



(a) Silhouette graph (b) Clusplot graph

Fig. 4. The reference Silhouette and Clusplot graph for the Iris dataset



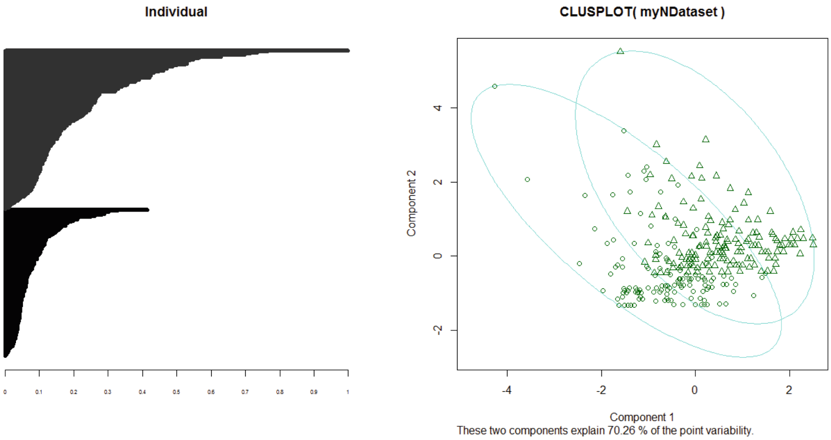
(a) Silhouette graph (b) Clusplot graph

Fig. 5. The reference Silhouette and Clusplot graph for the Haberman dataset

The Fig. 6 reports a typical result after five generations with the following parameters:

- maximum number of clusters: 6
- mutation rate: 0.2
- σ: 0.1
- number of parents: 1

In the Figs. 4 and 5 are shown the reference silhouette and clusplot graphs for the Iris and Haberman dataset.



(a) Silhouette graph

(b) Clusplot graph

Fig. 6. The Silhouette and Clusplot graph for the Haberman dataset after five generations with Revok

Results after few iterations, as can be seen from plotted graphs, approximate well the real partitions of the two datasets.

6 Conclusions

Revok is still a proof of concept and we need to test it with many more datasets, but clearly shows how IEC, along with compact visualization tools, can help users in clustering data by simply relying on their perceptual skills. Moreover, IEC can help users to ground the clustering process on their needs and knowledge (even if this knowledge is almost unconscious [9]). In fact, when living animals categorize incoming information, they categorize not only by virtue of the statistical property of the data flow but, above all, for their purposive behaviour [7]. The importance of IEC in data clustering resides in this human value. In order to improve Revok thus, we can act at different levels. A first level refers to distance between data points, usually we utilise euclidean distances but studies about the formation of concepts in humans, suggest that we may use different metrics: future version of Revok will be able to capture such a feature. A second level is about the visualization. At the moment each data point is shown as a line in the silhouette graph, but it can be useful to see data items in terms of physical features (i.e. color, shape etc.) in order to facilitate the recognition of similarity between elements.

References

1. Abul Hasan, M.J., Ramakrishnan, S.: A survey: hybrid evolutionary algorithms for cluster analysis. Artif. Intell. Rev. **36**(3), 179–204 (2011)
2. Bandyopadhyay, S., Maulik, U.: Genetic clustering for automatic evolution of clusters and application to image classification. Pattern Recogn. **35**(6), 1197–1208 (2002)
3. Ben-Israel, A., Iyigun, C.: Probabilistic d-clustering. J. Classif. **25**(1), 5–26 (2008)
4. Brintrup, A., Ramsden, J., Tiwari, A.: A review on design optimisation and exploration with interactive evolutionary computation. In: Tiwari, A., Roy, R., Knowles, J., Avineri, E., Dahal, K. (eds.) Applications of Soft Computing. AISC, vol. 36, pp. 111–120. Springer, Heidelberg (2006)
5. Du, K.L.: Clustering: a neural network approach. Neural Netw. **23**(1), 89–107 (2010)
6. Everitt, B., Landau, S., Leese, M.: Cluster Analysis. A Hodder Arnold Publication. Wiley, New York (2001)
7. Goodwin, C.J.: A History of Modern Psychology. Wiley, New York (2002)
8. Hruschka, E.R., Campello, R.J.G.B., Freitas, A.A., De Carvalho, A.C.P.L.F.: A survey of evolutionary algorithms for clustering. Trans. Syst. Man Cybern. Part C **39**(2), 133–155 (2009)
9. Kihlstrom, J.: The cognitive unconscious. Science **237**(4821), 1445–1452 (1987)
10. Lee, J.Y., Cho, S.B.: Sparse fitness evaluation for reducing user burden in interactive genetic algorithm. In: 1999 IEEE International Fuzzy Systems Conference Proceedings, 1999. FUZZ-IEEE '99, vol. 2, pp. 998–1003 (1999)
11. Lumley, T.: R Fundamentals and Programming Techniques. Chapman & Hall/CRC, Boca Raton (2006)
12. Menardi, G.: Density-based silhouette diagnostics for clustering methods. Stat. Comput. **21**(3), 295–308 (2011)
13. Miglino, O., Gigliotta, O., Ponticorvo, M., Stefano, N.: Breedbot: an evolutionary robotics application in digital content. Electron. Libr. **26**(3), 363–373 (2008)
14. Parmee, I., Bonham, C.: Cluster-oriented genetic algorithms to support interactive designer/evolutionary computing systems. In: Proceedings of the 1999 Congress on Evolutionary Computation, 1999. CEC 99, vol. 1, pp. 546–553 (1999)
15. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2008). http://www.R-project.org, ISBN 3-900051-07-0
16. Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)
17. Sheikh, R., Raghuwanshi, M.M., Jaiswal, A.: Genetic algorithm based clustering: a survey. In: First International Conference on Emerging Trends in Engineering and Technology, 2008. ICETET '08, pp. 314–319 (2008)
18. Takagi, H.: Interactive evolutionary computation: fusion of the capabilities of EC optimization and human evaluation. Proc. IEEE **89**(9), 1275–1296 (2001)
19. Teh, C.S., Chen, C.J.: Interactive evolutionary computation and density-based clustering for data analysis. In: International Conference on Intelligent and Advanced Systems, 2007. ICIAS 2007, pp. 104–108 (2007)
20. Xu, R., Wunsch II, D.: Survey of clustering algorithms. Trans. Neural Netw. **16**(3), 645–678 (2005)
21. Yang, Y., Kamel, M.S.: An aggregated clustering approach using multi-ant colonies algorithms. Pattern Recogn. **39**(7), 1278–1289 (2006)