What is a Cluster?
Perspectives from Game Theory

Marcello Pelillo
University of Venice, Italy
pelillo@dsi.unive.it

> "Since no paradigm ever solves all the problems it defines and since no two paradigms leave all the same problems unsolved, paradigm debates always involve the question: Which problems is it more significant to have solved?"
>
> Thomas S. Kuhn, *The Structure of Scientific Revolutions* (1962)

Extended Abstract

There is no shortage of clustering algorithms, and recently a new wave of excitement has spread across the machine learning community mainly because of the important development of spectral methods. At the same time, there is also growing interest around fundamental questions pertaining to the very nature of the clustering problem (see, e.g., [17, 1, 28]). Yet, despite the tremendous progress in the field, the clustering problem remains elusive and a satisfactory answer even to the most basic questions is still to come.

Upon scrutinizing the relevant literature on the subject, it becomes apparent that the vast majority of the existing approaches deal with a very specific version of the problem, which asks for *partitioning* the input data into coherent classes. In fact, almost invariably, the problem of clustering is *defined* as a partitioning problem, and even the classical distinction between hierarchical and partitional algorithms [15] seems to suggest the idea that partitioning data is, in essence, what clustering is all about (as hierarchies are but nested partitions). This is unfortunate, because it has drawn the community's attention away from different, and more general, variants of the problem and has led people to neglect underdeveloped foundational issues. As J. Hartigan clearly put it more than a decade ago: "We pay too much attention to the details of algorithms. [...] We must begin to subordinate engineering to philosophy." [10, p. 3].

The partitional paradigm (as I will call it, following Kuhn) is attractive as it leads to elegant mathematical and algorithmic treatments and allows us to employ powerful ideas from such sophisticated fields as linear algebra, graph theory, optimization, statistics, information theory, etc. However, there are several (far too often neglected) reasons for feeling uncomfortable with this oversimplified formulation. Probably the best-known limitation of the partitional approach is the typical (algorithmic) requirement that the number of clusters be known in advance, but there is more than that.

To begin, the very idea of a partition implies that *all* the input data will have to get assigned to some class. This subsumes the old philosophical view which gives categories an *a priori* ontological status, namely that they exist independent of human experience, a view which has now been discredited by cognitive scientists, linguists, philosophers, and machine learning researchers alike (see, e.g., [18, 7, 9]). Further, there are various applications for which it makes little sense to force all data items to belong to some group, a process which might result either in poorly-coherent clusters or in the creation of extra spurious classes. As an extreme example, consider the classical figure/ground separation problem in computer vision which asks for extracting a coherent region (the figure) from a noisy background [12, 23]. It is clear that, due to their intrinsic nature, partitional algorithms have no chance of satisfactorily solving this problem, being, as they are, explicitly designed to partition all the input data, and hence the unstructured clutter items too, into coherent groups. More recently, motivated by practical applications arising in document retrieval and bioinformatics, a conceptually identical problem has attracted some attention within the machine learning community and is generally known under the name of one-class clustering [8, 5].

The second intrinsic limitation of the partitional paradigm is even more severe as it imposes that each element cannot belong to more than one cluster. There are a variety of important applications, however, where this requirement is too restrictive. Examples abound and include, e.g., clustering micro-array gene expression data (wherein a gene often participate in more than one process), clustering documents into topic categories, perceptual grouping, and segmentation of images with transparent surfaces. In fact, the importance of dealing with overlapping clusters has been recognized long ago [16] and recently, in the machine learning community, there has been a renewed interest around this problem [3, 11]. Typically, this is solved by relaxing the constraints imposed by crisp partitions in such a way as to have "soft" boundaries between clusters.

Finally, I would like to mention another limitation of current state-of-the-art approaches to clustering which, admittedly, is not caused in any direct way by the partitioning assumption but, rather, by the intrinsic nature of the technical tools typically used to attack the problem. This is the symmetry assumption, namely the requirement that the similarities between the data being clustered be symmetric (and non-negative). Indeed, since Tversky's classical work [26], it is widely recognized by psychologists that similarity is an asymmetric relation. Further, there are many practical applications where asymmetric (or, more generally, non-metric) similarities do arise quite naturally. For example, such (dis)similarity measures are typically derived when images, shapes or sequences are aligned in a template matching process. In image and video processing, these measures are preferred in the presence of partially occluded objects [14]. Other examples include pairwise structural alignments of proteins that focus on local similarity [2], variants of the Hausdorff distance [6], normalized edit-distances, and probabilistic measures such as the Kullback-Leibler divergence. A common method to deal with asymmetric affinities is simply to symmetrize them, but in so doing we might lose important information that reside in the asymmetry. As argued in [14], the violation of metricity is often not an artifact of poor choice of features or algorithms, but it is inherent in the problem of robust matching when different parts of objects (shapes) are matched to different images. The same argument may hold for any type of local alignments. Corrections or simplifications of the original affinity matrix may therefore destroy essential information.

Although probabilistic model-based approaches do not suffer from several of the limitations mentioned above, here I will suggest an alternative strategy. Instead of insisting on the idea of determining a partition of the input data, and hence obtaining the clusters as a by-product of the partitioning process, in this presentation I propose to reverse the terms of the problem and attempt instead to derive a rigorous formulation of the very notion of a cluster. Clearly, the *conceptual* question "what is a cluster?" is as hopeless, in its full generality, as is its companion "what is an *optimal* clustering?" which has dominated the literature in the past few decades, both being two sides of the same coin. An attempt to answer the former question, however, besides shedding fresh light into the nature of the clustering problem, would allow us, as a consequence, to naturally overcome the major limitations of the partitional approach alluded to above, and to deal with more general problems where, e.g., clusters may overlap and clutter elements may get unassigned, thereby hopefully reducing the gap between theory and practice.

In our endeavor to provide an answer to the question raised above, we found that game theory offers a very elegant and general perspective that serves well our purposes. Hence, in the second, constructive part of the presentation I will describe a game-theoretic framework for clustering [21, 25, 22] which has found applications in fields as diverse as computer vision and bioinformatics. The starting point is the elementary observation that a "cluster" may be informally defined as a maximally coherent set of data items, i.e., as a subset of the input data C which satisfies both an *internal* criterion (all elements belonging to C should be highly similar to each other) and an *external* one (no larger cluster should contain C as a proper subset). We then formulate the clustering problem as a non-cooperative *clustering game*. Within this context, the notion of a cluster turns out to be equivalent to a classical equilibrium concept from (evolutionary) game theory, as the latter reflects both the internal and external cluster conditions mentioned above.

Evolutionary game theory originated in the early 1970's as an attempt to apply the principles and tools of game theory to biological contexts, with a view to model the evolution of animal, as opposed to human, behavior (see the classical work by J. Maynard Smith [19] who pioneered the field). It considers an idealized scenario whereby pairs of individuals are repeatedly drawn at random from a large, ideally infinite, population to play a symmetric two-player game. In contrast to conventional game theory, here players are not supposed to behave rationally or to have complete knowledge of the details of the game. They act instead according to an inherited behavioral pattern, or pure strategy, and it is supposed that some evolutionary selection process operates over time on the distribution of behaviors. Economists and social scientists have soon recognized the advantages of this new branch of game theory, as it allows one to elegantly get rid of the much-debated assumptions underlying the traditional approach, concerning the full rationality and complete knowledge of players. It also offered a *dynamical* perspective to game theory, an element which was totally missing in the classical theory, and provided new tools to deal with the equilibrium selection problem (namely, to explain reasons for players of a game to choose a certain equilibrium over another). Nowadays, evolutionary game theory is a well-established field on its own and has become one of the most active and rapidly growing areas in economics and social sciences. We refer the reader to [13, 27] for classical introductions to this rapidly expanding field. A central concept in evolutionary game theory is the notion of an evolutionary stable strategy (ESS), which is essentially a Nash equilibrium satisfying an additional stability property which guarantees that if an ESS is established in a population, and if a small proportion of the population adopts some mutant behavior, then the selection process will eventually drive them to extinction.

Now, to get back to our problem, the (pairwise) clustering problem can be formulated as the following (two-player) game. Assume a pre-existing set of objects O and a (possibly asymmetric and even negative) matrix of affinities A between the elements of O. Two players with complete knowledge of the setup play by simultaneously selecting an element of O. After both have shown their choice, each player receives a payoff, monetary or otherwise, proportional to the affinity that the chosen element has with respect to the element chosen by the opponent. Clearly, it is in each player's interest to pick an element that is strongly supported by the elements that the adversary is likely to choose. As an example, let us assume that our clustering problem is one of figure/ground discrimination, that is, the objects in O consist of a cohesive group with high mutual affinity (figure) and of non-structured noise (ground). Being non-structured, the noise gives equal average affinity to elements of the figures as to elements of the ground. Informally, assuming no prior knowledge of the inclination of the adversary, a player will be better-off selecting elements of the figure rather than of the ground.

Within this framework, clusters correspond to the ESS's of our non-cooperative game. The hypotheses that each object belongs to a cluster compete with one-another, each obtaining support from compatible edges and competitive pressure from the others. Competition will reduce the population of individuals that assume weakly supported hypotheses, while allowing populations assuming hypotheses with strong support to thrive. Eventually, all inconsistent hypotheses will be driven to extinction, while all the surviving ones will reach an equilibrium whereby they will all receive the same average support, hence exhibiting the internal coherency characterizing a cluster As for the extinct hypotheses, they will provably have a lower support, thereby hinting to external incoherency. The stable strategies can be found using *replicator dynamics*, a classic formalization of a natural selection process [27, 13].

In a nutshell, our game-theoretic perspective has the following attractive features:

1. it makes no assumption on the underlying (individual) data representation: like spectral (and, more generally, graph-based) clustering, it does not require that the elements to be clustered be represented as points in a vector space;
2. it makes no assumption on the structure of the affinity matrix, being it able to work with asymmetric and even negative similarity functions alike;
3. it does not require *a priori* knowledge on the number of clusters (since it extracts them sequentially);
4. it leaves clutter elements unassigned (useful, e.g., in figure/ground separation or one-class clustering problems)
5. it allows extracting overlapping clusters [24];
6. it generalizes naturally to hypergraph clustering problems, i.e., in the presence of high-order affinities [22], in which case the clustering game is played by more than two players.

The approach outlined above is but one example of using purely game-theoretic concepts to model *generic* machine learning problems (see [4] for another such example in a totally different context), and the potential of game theory to machine learning is yet to be fully explored. Other areas where game theory could potentially offer a fresh and powerful perspective include, e.g., semi-supervised learning, multi-similarity learning, multi-task learning, learning with incomplete information, learning with context-dependent similarities. The concomitant increasing interest around the algorithmic aspects of game theory [20] is certainly beneficial in this respect, as it will allow useful cross-fertilization of ideas.

Acknowledgments. I acknowledge financial support from the FET programme within EU FP7, under the SIMBAD project (contract 213250), and I am grateful to Mário Figueiredo and Tiberio Caetano for providing useful comments on a preliminary version of this paper.

References

[1] M. Ackerman and S. Ben-David. Measures of clustering quality: A working set of axioms for clustering. *NIPS 2008*.

[2] S. F. Altschul, W. Gish, W. Miller, E. W. Meyers, and D. J. Lipman. Basic local alignment search tool. *J. Molecular Biology*, 215:403–410, 1990.

[3] A. Banerjee, C. Krumpelman, S. Basu, R. J. Mooney, and J. Ghosh. Model-based overlapping clustering. *KDD 2005*.

[4] N. Cesa Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[5] K. Crammer, P. P. Talukdar, and F. Pereira. A rate-distortion one-class model and its applications to clustering. *ICML 2008*.

[6] M. P. Dubuisson and A. K. Jain. Modified Hausdorff distance for object matching. *ICPR 1994*.

[7] U. Eco. *Kant and the Platypus: Essays on Language and Cognition*. Harvest Books, 2000.

[8] G. Gupta and J. Ghosh. Robust one-class clustering using hybrid global and local search. *ICML 2005*.

[9] I. Guyon, U. von Luxburg, and R. C. Williamson. Clustering: Science or art? Opinion paper, 2009.

[10] J. Hartigan. Introduction. In P. Arabie, L. J. Hubert, and G. de Soete, editors, *Clustering and Classification*, River Edge, NJ, 1996. World Scientific.

[11] K. A. Heller and Z. Ghahramani. A nonparametric Bayesian approach to modeling overlapping clusters. *AISTATS 2007*.

[12] L. Herault and R. Horaud. Figure-ground discrimination: A combinatorial optimization approach. *IEEE Trans. Pattern Anal. Machine Intell.*, 15(9):899–914, 1993.

[13] J. Hofbauer and K. Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, Cambridge, UK, 1998.

[14] D. W. Jacobs, D. Weinshall, and Y. Gdalyahu. Classification with nonmetric distances: Image retrieval and class representation. *IEEE Trans. Pattern Anal. Machine Intell.*, 22:583–600, 2000.

[15] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, NJ, 1988.

[16] N. Jardine and R. Sibson. The construction of hierarchic and non-hierarchic classifications. *Computer Journal*, 11:177–184, 1968.

[17] J. M. Kleinberg. An impossibility theorem for clustering. *NIPS 2002*.

[18] G. Lakoff. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. The University of Chicago Press, 1987.

[19] J. Maynard Smith. *Evolution and the Theory of Games*. Cambridge University Press, Cambridge, UK, 1982.

[20] N. Nisan, T. Roughgarden, É. Tardos, and V. Vazirani, editors. *Algorithmic Game Theory*. Cambridge University Press, 2007.

[21] M. Pavan and M. Pelillo. Dominant sets and pairwise clustering. *IEEE Trans. Pattern Anal. Machine Intell.*, 29(1):167–172, 2007.

[22] S. Rota Bulò and M. Pelillo. A game-theoretic approach to hypergraph clustering. *NIPS 2009*.

[23] A. Shashua and S. Ullman. Structural saliency: The detection of globally salient features using a locally connected network. *ICCV 1988*.

[24] A. Torsello, S. Rota Bulò, and M. Pelillo. Beyond partitions: Allowing overlapping groups in pairwise clustering. *ICPR 2008*.

[25] A. Torsello, S. Rota Bulò, and M. Pelillo. Grouping with asymmetric affinities: A game-theoretic perspective. *CVPR 2006*.

[26] A. Tversky. Features of similarity. *Psychol. Rev.*, 84:327–352, 1977.

[27] J. W. Weibull. *Evolutionary Game Theory*. MIT Press, Cambridge, MA, 1995.

[28] R. B. Zadeh and S. Ben-David. A uniqueness theorem for clustering. *UAI 2009*.