Pattern Recognition 81 (2018) 95–111

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Clustering ensembles: A hedonic game theoretical approach

Nelson C. Sandes, André L.V. Coelho*

Graduate Program in Applied Informatics, Center of Technological Sciences, University of Fortaleza, Brazil

ARTICLE INFO

Article history:
Received 16 May 2017
Revised 9 March 2018
Accepted 20 March 2018
Available online 26 March 2018

MSC:
62H30
91A12
91A80

Keywords:
Data clustering
Clustering ensemble
Hedonic game
Nash stability
Evidence accumulation

ABSTRACT

Clustering ensembles (CE) comprise a class of pattern recognition methods that take a set of data clusterings (base partitions) as input and generate a consensus, better-quality partition as output. This work tackles the CE problem from a hedonic game theoretical perspective. In the modeled cooperative game, data points are viewed as players while clusters are regarded as coalitions. Interestingly, we show that by using an evidence-accumulation based similarity measure our novel Hedonic Game based Clustering Ensemble (HGCE) algorithm always converges to a Nash stable coalition structure, that is, to a clustering solution that cannot be unilaterally improved from the standpoint of each data point. A variant of the algorithm is also introduced, which is insensitive to the way the data points are ordered in the data set. In order to assess the potentials of HGCE and contrast its performance with that exhibited by a number of CE methods, experiments have been conducted on several artificial and real-world data sets, the majority of which related to bioinformatics. Overall, the empirical results and statistical tests relative to two well-known external validity measures ratify the usefulness and competitiveness of the proposed approach, also showing that HGCE is computationally efficient and resilient to random perturbations to the set of base partitions.

1. Introduction

Cluster analysis is a well-known unsupervised pattern recognition task that has been instrumental for leveraging a range of data analytical activities [1,2]. In a nutshell, the data clustering problem consists in finding a partition of a set of data points (a.k.a. patterns, objects) in such a way that similar points are grouped together into the same cluster, while those that are dissimilar to each other are assigned to different clusters [3,4].

Although of easy statement, the data clustering task is computationally hard to pursue, mainly due to its unsupervised, combinatorial nature [2,5,6] and to the subjective, contextual notions of what a good cluster (and clustering) should really be [3,7]. As a consequence, several algorithms have been proposed and new algorithms continue to emerge, each associated with different assumptions on the data and different similarity and validation criteria [1,8].

Recently, non-cooperative and evolutionary game theory concepts have been adopted in the development of new clustering algorithms [9,10]. For instance, Gupta and Ranganathan [11] modeled the clustering problem as a competitive game with pure strategies, aiming at the concurrent optimization of two complementary criteria, namely compaction and equipartitioning. By achieving the Nash equilibrium [9] of the game, the final data partition is obtained. Badami et al. [12] extended the previous work by also considering mixed strategies to be available to the players, which might lead to better equilibria and thus clustering solutions. On the other hand, Pelillo et al. [13] formulated the clustering problem in terms of a non-cooperative clustering game and showed that a natural interpretation of a cluster turns out to be equivalent to an evolutionary game-theoretic equilibrium concept. A similar approach was conducted by Rota Bulò and Pelillo [14] for hypergraph clustering, which refers to the process of extracting maximally coherent groups from a set of objects using high-order (rather than pairwise) similarities.

Cooperative game theory (CGT) [15] has also been investigated for the purpose of clustering. For example, Dhamal et al. [16] modeled the data clustering problem as a characteristic form game for which four well-known solution concepts (Nucleolus, Shapley value, Gately point and τ-value [9,10]) coincide. In the conceived algorithm, named as DRAC (after *Density-Restricted Agglomerative Clustering*), the Shapley value of the data points are directly related to their density. DRAC outperformed other conventional clustering algorithms on standard data sets. On the other hand, Garg et al. [17] mapped the data cluster formation to coalition formation in cooperative games, and then took advantage of the Shap-

* Corresponding author.
E-mail addresses: nelson.sandes@edu.unifor.br (N.C. Sandes), acoelho@unifor.br, coelho.alv@gmail.com (A.L.V. Coelho).

https://doi.org/10.1016/j.patcog.2018.03.017
0031-3203/

ley value of the patterns to identify clusters and their prototypes. Since the underlying game is convex, the resulting biobjective clustering algorithm (referred to as BiGC) is efficient and can yield high-quality partitions with respect to both average point-to-center distance and average intra-cluster point-to-point distance. The authors also proved that BiGC satisfies key clustering properties, such as order independence and scale invariance.

More recently, Feldman et al. [18] considered the class of hedonic games for dealing with clustering problems, even though their work did not focus specifically on data clustering. In a nutshell, hedonic games model scenarios where players have explicit preferences over coalitions. Each player only cares about which other players compose its coalition, and does not care about how the other coalitions are formed [19]. Ideally, the outcome of a hedonic game is a coalition structure (that is, a partition of the set of players into coalitions) that conforms to all players' preferences. In particular, Feldman et al. [18] investigated two different settings of hedonic clustering games, namely fixed clustering, which subdivides into *K*-medians and *K*-centers, and correlation clustering. The authors provided a thorough theoretical analysis of these games, characterizing Nash equilibria, and proving upper and lower bounds on the price of anarchy and price of stability. Albeit very encompassing, no experiments on real (data) clustering problems were conducted in [18] to empirically validate the potentials of the hedonic clustering game abstraction.

In this paper, we also resort to the solid theoretical framework made available by CGT to deal with the data clustering problem. However, our focus is more specifically on clustering ensembles (CE), that is, on combinations (aggregations) of input clusterings (known as base partitions) aiming at producing consensual, better-quality partitions [20–24].

Despite the fact that conventional, single clustering algorithms have been successfully applied in a range of scenarios [1,3], the choice of the algorithm best suited to a given data set is still a non-trivial task to pursue. One chief reason for this is that different algorithms abide by different optimization and similarity criteria, and are sensitive to the way their control parameters are effectively set up. For example, the familiar *K*-means algorithm [4] is guided by the minimization of the distance between data points and the centroids of their respective clusters, and its good performance is very contingent upon the appropriate choice of the number of clusters *K* and similarity measure. In such circumstances, the aggregation of several partitions of the same data, possibly generated by different algorithms or the same algorithm with different parameterizations, comes to be a much useful strategy [22,24].

Although the technical literature has demonstrated, both theoretically and empirically, the suitability of the CE approach, there is still room for improvement in this class of algorithms, both in terms of effectiveness and efficiency. In order to help filling this gap, this work tackles the CE problem from a hedonic game theoretical perspective [15,18]. Arguably, this is the first research initiative effectively investigating CGT concepts in the CE context.

In our modeled cooperative game, data points are viewed as players whereas clusters are regarded as coalitions. Being a hedonic setting [18], the utility of each player is determined by the identity of the other members composing its cluster. In our model, each coalition is comprised of players that are similar to each other based on the frequency that they have been grouped together in the base partitions.

Interestingly, we show in the sequel that by using an evidence-accumulation based similarity measure between data points [21,25], our novel *Hedonic Game based Clustering Ensemble* (HGCE) algorithm always converges to a Nash stable coalition structure (data partition), where no player can improve its utility by unilaterally changing its own coalition (cluster) [15,26]. Moreover, we also provide a variant of HGCE that is insensitive to the way the data points are actually ordered in the data set.

In order to validate the potentials of HGCE and contrast its performance with that exhibited by relevant CE methods, experiments have been conducted on several artificial and real-world data sets, the majority of which related to bioinformatics [27,28]. Overall, the empirical results and statistical tests relative to two well-known external validity measures ratify the usefulness and competitiveness of the proposed approach. In this context, the experiments reveal that HGCE is computationally efficient and robust to random perturbations to the set of base partitions. Moreover, the satisfactory performance of HGCE has not varied significantly when we changed the way the base partitions are generated and when we modified the initial clustering (coalition structure) conditions, since the different Nash equilibria that may be achieved at convergence yield final clustering solutions of similar good quality.

The rest of the paper is structured as follows. In Section 2, we formally characterize the CE problem and briefly overview prominent CE algorithms that have been used for performance comparison in our experiments. In Section 3, we focus on the topic of coalition formation and hedonic games, providing the theoretical background for our approach. Especially, we draw attention to some relevant results that apply to the subclass of hedonic games that is associated with our novel algorithm (namely, additively separable hedonic games – ASHG [19]). Section 4 is devoted to the detailed description of HGCE and its main conceptual ingredients. We also discuss some relevant issues related to its best response dynamics and present an order-insensitive variant. Experimental results are discussed in detail in Section 5, where a thorough statistical analysis is conducted concerning the clustering performance of our algorithm vis-à-vis the alternative CE methods. Section 6 summarizes our main contributions and concludes the paper.

2. Clustering ensembles

Formally, the CE problem can be stated as follows[1] [21,28]. Let $X = \{x_1, x_2, \ldots, x_n\}$ be a data set of patterns and be$CS = \{C_1, C_2, \ldots, C_{|CS|}\}$ be a partition (clustering) of this data set, where C_k is a cluster of CS, $1 \leq k \leq |CS|$. From the definition of a partition, we have $C_{k_1} \cap C_{k_2} = \varnothing$, for any $1 \leq k_1, k_2 \leq |CS|$, $k_1 \neq k_2$, and $\bigcup_{k=1}^{|CS|} C_k = X$. Also, let $\Pi = \{CS^1, CS^2, \ldots, CS^M\}$ be a set of base partitions of X and let $\mathbb{P}$ be the set of all possible partitions in X, such that $\Pi \subset \mathbb{P}$. The CE problem aims at finding a new and better (according to some validity criterion) partition $CS^+ \in \mathbb{P}$ by using the information available in Π. Alternatively, it may also target a reliable partition, especially when the statistical distribution of the data is unknown and the most appropriate individual clustering algorithm cannot be determined.

The related literature has discussed several attractive properties displayed by the class of CE algorithms, such as consistence, improved quality solution, stability, knowledge reuse, data independence, and privacy protection [21,22,24,29,30]. In general, CE techniques comprise two steps (refer to Fig. 1) [22,24]: the generation step, when the base partitions are induced; and the consensus function step, when a new partition is obtained via some consensus function operating on the base partitions, extracting useful information (evidence) from them. The main differences between CE algorithms usually lie in the approach to generate the input clusterings and the function to implement the consensus step.

[1] Notice that in the context of HGCE we render as equivalent the concepts of "cluster" and "coalition" as well as the concepts of "clustering" and "coalition structure". So, these corresponding concepts share the same mathematical notation along the paper (refer to Section 3).