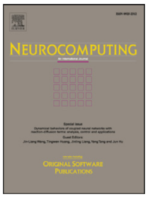
Neurocomputing 000 (2017) 1–13

Contents lists available at ScienceDirect

Neurocomputing

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



A multi-act sequential game-based multi-objective clustering approach for categorical data

Imen Heloulou[a,*], Mohammed Said Radjef[a], Mohand Tahar Kechadi[b]

[a] LaMOS Research Unit, Department of Operational Research, University of Abderrahmane Mira, Road Targua Ouzemour, Bejaia 06000, Algeria
[b] School of Computer Science and Informatics, University of College Dublin (UCD), Belfield, Dublin 4, Ireland

ARTICLE INFO

Article history:
Received 1 April 2016
Revised 2 January 2017
Accepted 5 June 2017
Available online xxx

Communicated by Deng Cai

Keywords:
Multi-objective clustering
Categorical data
Multi-act sequential game
Nash equilibrium

ABSTRACT

Clustering categorical data, where no natural ordering can be found among the attributes values, has started drawing interest recently. Few clustering methods have been proposed to satisfy the categorical data requirements. Most of these methods have focused on optimizing a single measure, however, several applications in different areas need to consider multiple incommensurable criteria, often conflicting, during clustering. Motivated by this, we developed a multi-objective clustering approach for categorical data based on sequential games. It can automatically generate the correct number of clusters. The approach consists of three main phases. The first phase identifies initial clusters according to an initialization mechanism which has an important effect in the final clustering result. The second phase uses multi-act multi-objective sequential two-player games in order to determine the appropriate number of clusters. A methodology based on backward induction is used to calculate a pure Nash equilibrium for each game. Finally, the third phase constructs homogenous clusters by optimizing intra-cluster inertia. The performance of this algorithm has been studied on both simulated and real-world datasets. Comparisons with other clustering algorithms illustrate the effectiveness of the proposed approach.

1. Introduction

Clustering can be defined as the process of partitioning a given dataset into K distinct groups based on some similarity/dissimilarity measures such that there are high intra-cluster similarity and low inter-cluster similarity. Cluster analysis has been widely used in numerous applications, including pattern recognition, image processing, biology, text query interface [1], etc. Several customized clustering algorithms have been developed to satisfy these problems requirements. Each algorithm tries to optimize an objective function based on a particular criteria. Traditionally, the important clustering objectives have been defined on the basis of compaction, connectedness and spatial separation metrics. First, clustering with the compaction objective attempts to identify clusters with minimum intra-cluster variation. K-means [2] is a widely used approach in this category. PAM or K-medoids [3] is a variation of the K-means, which instead of the cluster center, the cluster medoid is determined. These kind of methods, mainly K-means, tend to be very effective for spherical and/or well-separated clusters; but they may fail for more complicated cluster structures [4]. The connectedness metric is based on the idea that neighboring objects should share the same cluster. Density-based methods [5] or single linkage agglomerative [4] implement this principle to provide clusters with arbitrary shapes and a natural protection against outliers. However, they can lack robustness when there is little spatial separation between clusters. Finally, algorithms based on spatial separation serve to maximize the inter-cluster separation. However, they give little guidance during the clustering process and identify outliers as individual clusters, while merging the bulk of data items into one cluster [4].

Most of the proposed clustering methods handle datasets with numeric attributes, where the dissimilarity between any two points of the dataset can be computed using standard distance measures such as Euclidean distance. However, data mining applications frequently involve data of different types; interval-scaled, ordinal, categorical, and binary variables, on which some well known distance measures can not be applied. Recently, much effort has been put on the development of new techniques for clustering categorical data, due to increasing number of applications in market basket study [6], text mining [7], etc. One of the most common ways to tackle categorical data clustering is to extend existing algorithms with an appropriate distance measure for categorical data. For example, K-modes [8,9] extends K-means by replacing the Euclidean distance with the simple matching dissimilarity measure

[*] Corresponding author.
E-mail addresses: heloulou.imen@gmail.com (I. Heloulou), radjefms@yahoo.fr (M.S. Radjef), tahar.kechadi@ucd.ie (M.T. Kechadi).

http://dx.doi.org/10.1016/j.neucom.2017.06.011
0925-2312/© 2017 Elsevier B.V. All rights reserved.

and the cluster center by a virtual object, called mode, with the most frequent category encountered in the cluster. A fuzzy version of the K-modes algorithm is proposed in [10].

ROCK [11] is a bottom-up algorithm which heuristically optimizes a criterion function defined in terms of the number of “links” between tuples given by the Jaccard coefficient [12]. Two data objects are more similar if they have more common neighbors. ROCK clusters a randomly sampled dataset and then partitions the entire dataset based on these clusters. CACTUS [13] is an agglomerative algorithm that conducts clustering from attribute relationship. The central idea is that summary information constructed from the dataset is sufficient for discovering well-defined clusters. Hierarchical clustering algorithms such as Single, Complete or Average linkage have also been used to cluster categorical data [4]. A Hamming distance vector-based categorical data clustering algorithm (CCDV) has been also developed in [14]. CCDV sequentially extracts the clusters from a given dataset based on the Hamming distance vectors, with automatic evolution of the number of the clusters. Few more techniques were also reported in [15–18]. However, all these algorithms rely on optimizing a single measure of the clustering goodness. Several applications require the simultaneous examination of multiple objectives, which are often antagonistic. Therefore, clustering should be considered as a multi-objective rather than single-objective optimization problem. The multi-objective clustering methods attempt to identify clusters in such a manner that several objectives are optimized during the process [19].

Multi-objective Pareto-optimization-based algorithms for clustering categorical data gained prominence recently. These methods perform simultaneous optimization of complementary objectives. Handl and Knowles [20] have proposed an evolutionary multi-objective clustering technique, where the partitioning criteria are chosen as the overall compactness and connectivity. The algorithm is capable of handling categorical and continuous datasets with automatic K-determination. Mukhopadhyay et al. [21] have proposed a fuzzy multi-objective algorithm for clustering categorical data by optimizing the fuzzy compactness and the fuzzy separation of the clusters. The power of genetic algorithm is exploited to search suitable clusters and their modes, in such a way that intra-cluster and inter-cluster distances are simultaneously optimized [22]. An Incremental Learning based on multi-objective fuzzy clustering for categorical data is also developed in [23] and it optimizes two conflicting objectives, namely J_m and XB indices. However, Pareto optimal solutions are not socially equitable and intend to identify solutions targeting the overall system optimization, rather than the optimization of individual objectives. Game theory approaches ensure that each metric in the problem is optimized with respect to the others [24]. When we have conflicting objectives, the clustering problem can be modeled as a game consisting of players with conflicting objectives competing to optimize their payoffs. Each player decision is based upon the decisions of all other players involved in the game.

There are many approaches that use game theory for pattern clustering. A multi-objective clustering based-noncooperative game is developed in [24] to optimize compaction and equipartitioning metrics. However, pure Nash equilibrium does not always exist which limits the performance of this algorithm. Based on this method, Badami et al. [25] have proposed a novel formulation of the payoffs function which models both compaction and equipartitioning objectives in equal priority. Their approach can be applied to mixed strategies as well as to pure ones. Therefore, it could identify better clusters due to the existence of Nash equilibrium in the space of mixed strategies. However, both algorithms have very high complexity in terms of both computational time and memory requirements, with respect to the number of players, strategies and consequently payoff matrices. Cooperative game theory with characteristic function and transferable utility, is also used to resolve the initialization clustering problem [26]. Two criteria are optimized; the average intra-cluster point-to-point distance using Shapley value, and the average distance between each point and its center using K-means. However, this approach is restricted to K-means, which is not always desirable especially when the clusters have unequal variances or when they have non-convex shapes. Bulò and Pelillo [27] have used the concept of evolutionary game theory for hypergraph clustering. They have proved that the notion of a cluster turns out to be equivalent to a classical equilibrium concept from (evolutionary) game theory, which incorporates the two basic properties of a cluster; internal and external coherency.

Recently, a multi-objective clustering based-sequential games was proposed in [28]. This approach tends to optimize R-squared, connectivity and intra-cluster inertia objectives and could be applied only for continuous data. Motivated by this, in this paper we propose a novel algorithm for multi-objective clustering of categorical data which is the first work addressing the issue of multi-objective clustering based-game theory to deal with categorical attributes of data. The basic idea behind our approach is to establish that the categorical data clustering problem can be posed as one of multi-objective sequential game. More precisely, in this paper, we provide a detailed description of the game theoretic model for multi-objective clustering of categorical data. We also present a novel initialization phase and a novel formulation of the payoffs function. In this way, not only the number of clusters is determined dynamically, but also the distribution of objects to clusters will be also done by negotiation. Finally, we study the performance of the proposed algorithm and compare it with widely used algorithms for categorical data.

The remainder of this paper is organized as follows: In the next section, we introduce the basic notations of clustering process and game theory. Then, Section 3 describes the problem of multi-objective clustering for categorical data. Section 4 discusses the basic concepts of the proposed multi-objective clustering approach for categorical data. The experimental results are provided in Section 5. Finally, Section 6 presents some concluding remarks and suggestions for future work.

2. Notations

2.1. Multi-objective clustering of categorical data

Let $X = \{x_1, \ldots, x_i, \ldots, x_n\}$ be a set of n objects, where each object x_i is described by a set of δ attributes $A_1, \ldots, A_\delta$. Each attribute A_j has a set of admissible categorical values; we denote it by $DOM(A_j) = \{a_j^1, \ldots, a_j^{q_j}\}$. It consists of different q_j categories. Hence, the ith object x_i is described as: $x_i = (x_{i1}, \ldots, x_{i\delta})$, where $x_{ij} \in DOM(A_j)$, $1 \leq j \leq \delta$.

Instead of using the mean as the center of the cluster, which does not exist for categorical variables, it is replaced by the mode [8,9]. Assume that $C_i = \{x_1^i, \ldots, x_l^i\}$ has a set of l objects. The mode of cluster C_i is a vector $m_i = (m_{i1}, \ldots, m_{i\delta})$, such that $m_{ij} \in DOM(A_j)$ is the most frequently encountered value in $\{x_{1j}^i, \ldots, x_{lj}^i\}$.

For categorical variables, the similarity measure between two objects x_i, $x_j \in X$, is usually defined as the number of matches among the categorical attributes values. This measure is also a kind of generalized Hamming distance [29].

$$d(x_1, x_2) = \sum_{j=1}^{\delta} \varphi(x_{1j}, x_{2j}), \tag{1}$$

where:

$$\varphi(x_{1j}, x_{2j}) = \begin{cases} 0, & \text{if } x_{1j} = x_{2j} \\ 1, & \text{if } x_{1j} \neq x_{2j}. \end{cases}$$