Journal of Computational Science 4 (2013) 219–231

Contents lists available at SciVerse ScienceDirect

Journal of Computational Science

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

Competitive clustering algorithms based on ultrametric properties

S. Fouchal[a,*,1], M. Ahat[b,1], S. Ben Amor[c,2], I. Lavallée[a,1], M. Bui[a,1]

[a] University of Paris 8, France
[b] Ecole Pratique des Hautes Etudes, France
[c] University of Versailles Saint-Quentin-en-Yvelines, France

ARTICLE INFO

Article history:
Received 30 July 2011
Received in revised form 16 October 2011
Accepted 24 November 2011
Available online 12 January 2012

Keywords:
Clustering
Ultrametric
Complexity
Amortized analysis
Average analysis
Ordered space

ABSTRACT

We propose in this paper two new competitive unsupervised clustering algorithms: the first algorithm deals with ultrametric data, it has a computational cost of $O(n)$. The second algorithm has two strong features: it is fast and flexible on the processed data type as well as in terms of precision. The second algorithm has a computational cost, in the worst case, of $O(n^2)$, and in the average case, of $O(n)$. These complexities are due to exploitation of ultrametric distance properties. In the first method, we use the order induced by an ultrametric in a given space to demonstrate how we can explore quickly data proximity. In the second method, we create an ultrametric space from a sample data, chosen uniformly at random, in order to obtain a global view of proximities in the data set according to the similarity criterion. Then, we use this proximity profile to cluster the global set. We present an example of our algorithms and compare their results with those of a classic clustering method.

1. Introduction

The clustering is useful process which aims to divide a set of elements into a set of finite number of groups. These groups are organized such as the similarity between elements in a same group is maximal, while similarity between elements from different groups is minimal [15,17].

There are several approaches of clustering, hierarchical, partitioning, density-based, which are used in a large variety of fields, such as astronomy, physics, medicine, biology, archaeology, geology, geography, psychology, and marketing [24].

The clustering aims to group objects of a data set into a set of meaningful subclasses, so it can be used as a stand-alone tool to get insight into the distribution of data [1,24].

The clustering of high-dimensional data is an open problem encountered by clustering algorithms in different areas [31]. Since the computational cost increases with the size of data set, the feasibility can not be fully guaranteed.

We suggest in this paper two novel unsupervised clustering algorithms: The first is devoted to the ultrametric data. It aims to show rapidly the inner closeness in the data set by providing a general view of proximities between elements. It has a computational cost of $O(n)$. Thus, it guarantees the clustering of high-dimensional data in ultrametric spaces. It can, also, be used as a preprocessing algorithm to get a rapid idea on behavior of data with the similarity measure used.

The second method is general, it is applicable for all kinds of data, it uses a metric measure of proximity. This algorithm provides rapidly the proximity view between elements in a data set with the desired accuracy. It is based on a sampling approach (see details in [1,15]) and ultrametric spaces (see details in [20,23,25]).

The computational complexity of the second method is in the worst case, which is rare, of $O(n^2)+O(m^2)$, where n is the size of data and m the size of the sample. The cost in the average case, which is frequent, is equal to $O(n)+O(m^2)$. In both cases m is insignificant, we give proofs of these complexities in Proposition 9. Therefore, we use $O(n^2)+\varepsilon$ and $O(n)+\varepsilon$ to represent respectively the two complexities.

This algorithm guarantees the clustering of high-dimensional data set with the desired precision by giving more flexibility to the user.

Our approaches are based in particular on properties of ultrametric spaces. The ultrametric spaces are ordered spaces such that data from a same cluster are "*equidistant*" to those of another one (e.g. in genealogy: two species belonging to the same family, "brothers", are at the same distance from species from another family, "cousins") [8,9].

We utilize ultrametric properties in the first algorithm to cluster data without calculating all mutual similarities. The structure

* Corresponding author.
E-mail addresses: said.fouchal@laisc.net (S. Fouchal), murat.ahat@laisc.net (M. Ahat), soufian.ben-amor@uvsq.fr (S. Ben Amor), ivan.lavallee@laisc.net (I. Lavallée), marc.bui@laisc.net (M. Bui).
[1] Laboratoire d'Informatique et des Systèmes Complexes, 41 rue Gay Lussac 75005 Paris, France, http://www.laisc.net.
[2] Laboratoire PRiSM, 45 avenue des Etats-Unis F-78 035 Versailles, France, http://www.prism.uvsq.fr/.

1877-7503/$ – see front matter.
doi:10.1016/j.jocs.2011.11.004

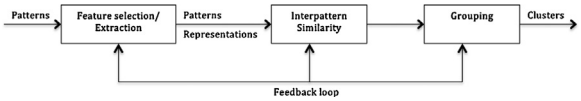

Fig. 1. Steps of clustering process.

induced by ultrametric distance allows us to get a general information about proximity between all elements from just one data, consequently we reduce the computational cost. We can find ultrametric spaces in many kinds of data sets such as: phylogeny where the distance of evolution is constant [16], genealogy, library, information and social sciences . . . to name a few.

In the second algorithm, we use ultrametric to acquire a first insight of the proximity between elements in just a sample data w.r.t. the similarity used. Once this view obtained, we expand it to cluster the whole data set.

The rest of the text is organized as follows. In Section 2, we present a brief overview of the clustering strategies. In Section 3, we introduce the notions of metric and ultra-metric spaces, distance and balls. Our first algorithm is presented in Section 4. We present an example of this first algorithm in Section 5. The second algorithm is introduced in Section 6. In Section 7, we present an example of the second algorithm and we compare our results with those of a classic clustering algorithm. Finally, in Section 8 we give our conclusion and future work.

2. Related work

The term "clustering" is used in several research communities to describe methods for grouping of unlabeled data. The typical pattern of this process can be summarized by the three steps of Fig. 1 [21].

Feature selection and extraction are preprocessing techniques which can be used, either or both, to obtain an appropriate set of features to use in clustering. Pattern proximity is calculated by similarity measure defined on data set between a pairs of objects. This proximity measure is fundamental to the clustering, the calculations of mutual measures between element are essential to most clustering procedures. The grouping step can be carried in different way, the most known strategies are defined bellow [21].

Hierarchical clustering is either agglomerative ("bottom-up") or divisive ("top-down"). The agglomerative approach starts with each element as a cluster and merges them successively until forming a unique cluster (i.e. the whole set) (e.g. *WPGMA* [9,10], *UPGMA* [14]). The divisive begins with the whole set and divides it iteratively until it reaches the elementary data. The outcome of hierarchical clustering is generally a dendrogram which is difficult to interpret when the data set size exceeds a few hundred of elements. The complexity of these clustering algorithms is at least $O(n^2)$ [28].

Partitional clustering creates clusters by dividing the whole set into k subsets. It can also be used as divisive algorithm in hierarchical clustering. Among the typical partitional algorithms we can name K-means [5,6,17] and its variants K-medoids, PAM, CLARA and CLARANS. The results depend on the k selected data in this kind of algorithms. Since the number of clusters is defined upstream of the clustering, the clusters can be empty.

Density-based clustering is a process where the clusters are regarded as a dense regions leading to the elimination of the noise. DBSCAN, OPTICS and DENCLUE are typical algorithms based on this strategy [1,4,7,18,24].

Since, the major clustering algorithms calculate similarities between all data prior to the grouping phase (for all types of similarity measure used), the computational complexity is increased to $O(n^2)$ before the execution of the clustering algorithm.

Our first approach deals with *ultrametric spaces*, we propose the first – as our best knowledge – unsupervised clustering algorithm on this kind of data without calculating similarities between all pairs of data. So, we reduce the computational cost of the process from $O(n^2)$ to $O(n)$. We give proofs that; since the data processed are described with an ultrametric distance we do not need to calculate all mutual distances to obtain information about proximity in the data set (*cf*. Section 4).

Our second approach is a new flexible and fast unsupervised clustering algorithm which costs mostly $O(n)+\varepsilon$ and seldom $O(n^2)+\varepsilon$ (in rare worst case), where n is the size of data set and ϵ is equal to $O(m^2)$ where m is the size of an insignificant part (sample) of the global set.

Even if the size of data increases, the complexity of the second proposed algorithm, the amortized complexity [30,32,34], remains of $O(n)+\varepsilon$ in the average case, and of $O(n^2)+\varepsilon$ in the worst case, thus it can process dynamic data such as those of Web and social network with the same features.

The two approaches can provide overlapped clusters, where one element can belong to more than one or more than two clusters (more general than weak-hierarchy), see [2,4,7] for detailed definitions about overlapping clustering.

3. Definitions

Definition 1. A metric space is a set endowed with distance between its elements. It is a particular case of a *topological space*.

Definition 2. We call a distance on a given set E, an application d: $E \times E \to \mathbb{R}^+$ which has the following properties for all $x, y, z \in E$:

1 (Symmetry) $d(x, y) = d(y, x)$,
2 (Positive definiteness) $d(x, y) \geqslant 0$, and $d(x, y) = 0$ if and only if $x = y$,
3 (Triangle inequality) $d(x, z) \leqslant d(x, y) + d(y, z)$.

Example 1. The most familiar metric space is the *Euclidean* space of dimension n, which we will denote by $\mathbb{R}^n$, with the standard formula for the distance: $d((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = ((x_1 - y_1)^2 + \cdots + (x_n - y_n)^2)^{1/2}$ (1).

Definition 3. Let (E, d) be a metric space. If the metric d satisfies the strong triangle inequality:

$$\forall x, y, z \in E, d(x, y) \leqslant max\{d(x, z), d(z, y)\}$$

then it is called ultrametric on E [19]. The couple (E, d) is an ultrametric space [11,12,29].

Definition 4. We name open ball centered on $a \in E$ and has a radius $r \in \mathbb{R}^+$ the set $\{x \in E : d(x, a) < r\} \subset E$, it is called $B_r(a)$ or $B(a, r)$.

Definition 5. We name closed ball centered on $a \in E$ and has a radius $r \in \mathbb{R}^+$ the set $\{x \in E : d(x, a) \leqslant r\} \subset E$, it is called $B_f(a, r)$.

Remark 1. Let d be an ultrametric on E. The closed ball on $a \in E$ with a radius $r > 0$ is the set: $B(a, r) = \{x \in E : d(x, a) \leqslant r\}$

Proposition 1. *Let d be an ultrametric on E, the following properties are true* [11]*:*

1 *If* $a, b \in E$, $r > 0$, *and* $b \in B(a, r)$, *then* $B(a, r) = B(b, r)$,
2 *If* $a, b \in E$, $0 < i \leqslant r$, *then either* $B(a, r) \cap B(b, i) = \emptyset$ *or* $B(b, i) \subseteq B(a, r)$. *This is not true for every metric space*,
3 *Every ball is clopen (closed and open) in the topology defined by d (i.e. every ultrametrizable topology is zero-dimensional). Thus, the parts are disconnected in this topology.*

Hence, the space defined by d is homeomorphic to a subspace of countable product of discrete spaces (c .f Remark 1) (see the proof in [11]*).*