Information Fusion 39 (2018) 81–95

Contents lists available at ScienceDirect

Information Fusion

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

Collaborative clustering: Why, when, what and how

Antoine Cornuéjols[a], Cédric Wemmert[b,*], Pierre Gançarski[b], Younès Bennani[c]

[a] UMR MIA-Paris, AgroParisTech, INRA - Université Paris-Saclay, Paris, France
[b] UMR ICube - Unistra, CNRS - Université de Strasbourg, Illkirch, France
[c] LIPN - UMR, CNRS - Université Paris 13, Sorbonne Paris Cité, Villetaneuse, France

ARTICLE INFO

Article history:
Received 2 January 2017
Revised 31 March 2017
Accepted 22 April 2017
Available online 24 April 2017

Keywords:
Collaborative clustering
Clustering combining
Cooperative clustering

ABSTRACT

Clustering is one type of unsupervised learning where the goal is to partition the set of objects into groups called clusters. Faced to the difficulty to design a general purpose clustering algorithm and to choose a good, let alone perfect, set of criteria for clustering a data set, one solution is to resort to a variety of clustering procedures based on different techniques, parameters and/or initializations, in order to construct one (or several) final clustering(s). The hope is that by combining several clustering solutions, each one with its own bias and imperfections, one will get a better overall solution.

In the cooperative clustering model, as Ensemble Clustering, a set of clustering algorithms are used in parallel on a given data set: the local results are combined to get a hopefully better overall clustering. In the collaborative framework, the goal is that each local computation, quite possibly applied to distinct data sets, benefit from the work done by the other collaborators.

This paper is dedicated to collaborative clustering. In particular, after a brief overview of clustering and the major issues linked to, it presents main challenges related to organize and control the collaborative process.

1. Introduction

Unsupervised learning is often defined in contrast with supervised learning. In *supervised learning*, the goal is to make predictions about output value y given an input object or instance $\mathbf{x}$. This is done through a decision procedure $h : \mathcal{X} \longrightarrow \mathcal{Y}$ that is learned from a training set $\mathcal{S} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)\}$ and some prior knowledge, where each example of $\mathcal{S}$ is composed of an object $\mathbf{x}_i \in \mathcal{X}$ and a corresponding output value $y_j \in \mathcal{Y}$.

By contrast, the objective of *unsupervised learning* is not to make predictions from as yet unknown input values to output values, but to reveal possible hidden structures in the available data set, $\mathcal{S} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. In a way, this can be compared to signal analysis by which one seeks a decomposition of the signal into underlying basis functions. If these putative structures or regularities may sometimes be extrapolated to make predictions about future events, this is not the primary goal of unsupervised learning. Another crucial distinction with supervised learning is that *there is no absolute way to measure the relevance of the uncovered regularities*, whatever their form [1]. In supervised learning, one can use validation sets or cross-validation to estimate the predictive value of the learned decision function. If the predictive performance is low, then either the data or the learning algorithm is wanting. Unfortunately, there is no equivalent to the predictive performance in unsupervised learning. The algorithms can only find the kind of underlying structures that the user has predefined either implicitly or explicitly in their code. In the best of worlds, the methods also provides some level of significance of the discovered structure. But there is no objective way of measuring the value of the findings, that is whether they correspond to some "true" underlying structure of the data set or if they are just figments of the imagination of the user and the algorithm chosen. Indeed, the significance tests that are often used as referees are themselves, by necessity, biased towards some types of regularities. This is this property that makes unsupervised learning so challenging, both to find a solid theoretical theory about what is a good or best technique, and to apply it with some level of confidence to data in need of interpretability.

Clustering is one type of unsupervised learning where the goal is to partition the set of objects into groups called clusters. These groups can be mutually exclusive or they may overlap, depending on the approach used. Clusters are defined by the fact that the objects within are more similar to each other than to objects from

* Corresponding author.
E-mail addresses: antoine.cornuejols@agroparistech.fr (A. Cornuéjols), cedric.wemmert@unistra.fr, wemmert@unistra.fr (C. Wemmert), pierre.gancarski@unistra.fr (P. Gançarski), younes.bennani@lipn.univ-paris13.fr (Y. Bennani).

http://dx.doi.org/10.1016/j.inffus.2017.04.008
1566-2535/

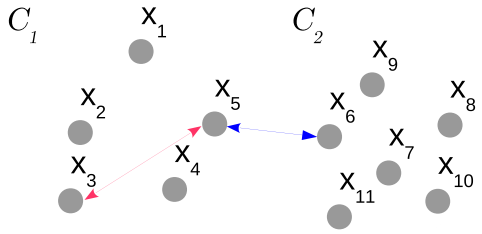

Fig. 1. Clustering is an ill-defined problem.

other groups. The similarity measure is of course of paramount importance to define the kind of structures or clusters that can be uncovered in the data, and hundreds of distances have been proposed in the literature according to the problem and context at hand.

Two major approaches of clustering exist: *generative* and *discriminative*, both relying more or less directly on a chosen distance. The former supposes that a generative model has been defined, often in the form of a statistical model, and the goal is to find the model parameters maximizing the probability that the data was generated by the model. The latter relies on similarity measures and on optimization criteria to find groups in the data. In either case, before an algorithm can be properly defined, numerous questions have to find an answer.

1.1. The questions raised in clustering

The exploration of ill-known data sets and the uncovering of hidden regularities are marred with an array of questions and potential pitfalls.

The question arising immediately is: what is clustering? Is there a clear definition and hence, hopefully, some measurable criterion that ought to be optimized?

Intuitively, clustering is the grouping of objects such that similar objects end up in the same group and dissimilar objects are assigned to different groups. Formally, clustering a data set $\mathcal{S}$ of N objects means finding a partition $\{\mathcal{C}_1, \ldots, \mathcal{C}_K\}$ of $\mathcal{S}$ such that:

$$\bigcup_{k=1}^{K} \mathcal{C}_k = \mathcal{S},$$

where the groups $\mathcal{C}_k$ are:

1. As ***homogeneous*** as possible (small intra-group variability)
2. As ***distinct*** as possible (large inter-group variability)

Most clustering techniques output *partitions* (disjoint clusters):

$$\mathcal{C}_k \cap \mathcal{C}_{k'} = \varnothing \quad \text{if} \quad k \neq k'$$

which is not always desirable.

For all its seemingly clear definition, clustering is *an ill-defined problem*. One fundamental issue is that clustering is based on the idea that similar objects should be clustered together while dissimilar objects should be separated in different groups. But, mathematically, similarity is not a transitive relation, while belonging to the same cluster is.

Thus, on Fig. 1, which seems a reasonable clustering of the given data points, $\mathbf{x}_1$ appears to be close to $\mathbf{x}_2$, and $\mathbf{x}_2$ to $\mathbf{x}_3$ and so on until $\mathbf{x}_{11}$, and, as a consequence, they should all be put in the same cluster. But, if the shown clustering is correct, it violates the first requirement (all similar elements should end up in the same cluster): $\mathbf{x}_5$ and $\mathbf{x}_6$ should belong to the same cluster; as well the second one (dissimilar elements should be put in distinct clusters): $\mathbf{x}_5$ and $\mathbf{x}_3$ should not belong to the same cluster.

There is therefore an ambiguity in the definition of clustering that can only be removed through some additional bias. For instance, the distance used for measuring the inter-group dissimilarity (e.g. single linkage, average linkage, complete linkage, and so on) will favor one type of structure over others. However, this bias impacts the clustering process and not the optimization criterion which remains therefore intrinsically ambiguous.

Another major source of problems is that an ideal clustering would entail the exploration of an impossibly large space of possible answers. Thus, the number of partitions of N objects in K groups is:

$$S_{N,K} = \frac{1}{K!} \sum_{k=0}^{K} (-1)^k \, (K-k)^N \binom{K}{k} \simeq \frac{K^N}{K!} \text{ as } N \to \infty \tag{1}$$

If the number of partitions K is not known beforehand, then the number of all partitions to be examined is given by the Bell number:

$$B_N = \sum_{k=1}^{N} S_{N,k} \tag{2}$$

As an illustration, a computer handling one million partitions per second would take more than 147,000 years to study all partitions of a set of only 25 elements: there are indeed 4,638,590,332,229,999,353 possible partitions of such a set!

One therefore has two perspectives: either finding an optimization criterion such that the optimization problem becomes convex in the search space, or designing a heuristic search algorithm that can search the space of solutions efficiently and, to some extent, escape local minima. No convex optimization criterion is known, and thus one must solve the second alternative. As it happens, most of the resulting optimization problems are NP-hard.

To sum up, clustering, is not only an *ill-defined* problem [2,3], it is also an *ill-posed problem* that requires some prior bias in order to be practically solved. Different algorithms may yield dramatically different outputs for the same input sets. Additionally, the entailed computational costs are huge if no proper heuristics is employed.

Consequently, several concrete questions must be answered before a clustering method can be defined and applied.

1.1.1. Formally defining the types of clusters looked for

In clustering, we wish to organize the data in some meaningful way, but "meaningful" depends on the context and on our focus of interest. The same given set of objects can be clustered in various different meaningful ways. For instance, we could be interested in categorizing speakers by the language they speak, or by the topic of discussion, or by gender. Accordingly, one would concentrate on different descriptors in the spoken signal, and use different distances in order to group the speakers.

The *distance* is a critical part in the definition of what types of clusters will be looked for. Actually, in many algorithms, several distances must be decided upon: a distance between instances in the input space, but also a distance between an instance and a cluster, and a distance between clusters. As is well-known by practitioners, any single difference in these choices points may alter considerably the result of a clustering.

Another problem is the choice of the relevant *number of clusters* when the number of "true" underlying categories is not known beforehand. This is related to the model selection problem [4]. Often, what is looked for are clusters that are compact (within inertia) and well separated (extra-inertia). It happens that these two requirements tend to be inversely correlated, when one improves, the other deteriorates. The relative weights put on these aspects control therefore the result in the same manner that the choice of the hypothesis space or of a regularization term controls the result in supervised learning. However, unlike for supervised learning, there is no ground truth in the data that can help choosing the