Pattern Recognition 71 (2017) 26–35

Contents lists available at ScienceDirect

Pattern Recognition

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

Consistency of mean partitions in consensus clustering

Brijnesh J. Jain

TU Berlin, Ernst-Reuter-Platz 7, Berlin, 10587, Germany

ARTICLE INFO

Article history:
Received 25 November 2016
Revised 11 April 2017
Accepted 17 April 2017
Available online 11 May 2017

Keywords:
Consensus clustering
Mean partition
Consistency
Central limit theorem

ABSTRACT

This article studies the asymptotic behavior of mean partitions in consensus clustering. We show that the mean partition approach is consistent and asymptotic normal under mild assumptions. To derive both results, we represent partitions as points of some geometric space, called orbit space. Then we draw on results from the theory of Fréchet means and stochastic programming. The asymptotic properties hold for continuous extensions of standard cluster criteria (indices). The results justify consensus clustering using finite but sufficiently large sample sizes. Furthermore, the orbit space framework provides a mathematical foundation for studying further statistical, geometrical, and analytical properties of sets of partitions.

1. Introduction

Clustering is a standard technique for exploratory data analysis that finds applications across different disciplines such as computer science, biology, marketing, and social science. The goal of clustering is to group a set of unlabeled data points into several clusters based on some notion of dissimilarity. Inspired by the success of classifier ensembles, consensus clustering has emerged as a research topic [9,22]. Consensus clustering first generates several partitions of the same dataset. Then it combines the sample partitions to a single consensus partition. The assumption is that a consensus partition better fits to the hidden structure in the data than individual partitions.

One standard approach of consensus clustering combines the sample partitions to a mean partition [4–7,10,14,18,20,21]. A mean partition best summarizes the sample partitions with respect to some (dis)similarity function. A natural question is the choice of sample size. If the sequence of mean partitions fails to converge stochastically for growing sample size n, then picking a reasonable value for n becomes an additional parameter selection problem. Otherwise, if the mean partitions converge stochastically to an expected partition, the problem of selecting a sample size n simplifies to the problem of selecting a sufficiently large n, because we have high confidence that nothing unexpected will happen when sampling further partitions. In other words, stochastic convergence justifies the common practice to draw finite but sufficiently large sample sizes.

Though there is an extensive literature on consensus clustering [22], little is known about the asymptotic behavior of the mean partition approach. Topchy et al. [19] studied the asymptotic behavior of the mean partition approach under the following simplifying assumptions:

(A1) The underlying distance is a semi-metric.[1]
(A2) Partitions are hard (crisp) partitions.
(A3) The expected partition is unique.
(A4) Partitions strongly concentrate on the expected partition.

In this contribution, we study the asymptotic behavior of the mean partition approach without drawing on assumptions (A2)–(A4). We show (i) consistency of the mean partitions, (ii) strong consistency of the variations, and (iii) a modified version of the Central Limit Theorem for mean partitions. We present two variants of results (i) and (ii). The first variant assumes that partitions form a compact metric space. The second variant requires the Euclidean space as ambient space and assumes that partitions are compared by a continuous cluster criterium. We also draw on continuity of the cluster criterium for showing result (iii). since standard criteria for comparing partitions are defined on the discrete space of hard partitions, we present examples of their continuous extensions. We can apply the generalized standard criteria to soft partitions and we can analyze the asymptotic behavior of the mean partition approach for hard and soft partitions in a unified manner.

The basic idea to derive the results is to represent partitions as points of a geometric space, called orbit space. Orbit spaces are well explored, possess a rich mathematical structure and have a natural connection to Euclidean spaces [3,11,16]. For the first variant of results (i) and (ii), we link the consensus function of the

[1] A semi-metric satisfies all axioms of a metric, but not necessarily the triangle inequality.

E-mail addresses: brijnesh.jain@gmail.com, jain@dai-labor.de

http://dx.doi.org/10.1016/j.patcog.2017.04.021
0031-3203/

mean partition approach to Fréchet functions [8], which are well explored in mathematical statistics [1,2]. The second variant of results (i) and (ii) as well as result (iii) apply results from stochastic programming [17].

The rest of this paper is structured as follows: Section 2 constructs the orbit space of partitions and introduces metric structures. In Section 3, we introduce Fréchet consensus functions and study their asymptotic behavior. Section 4 presents examples of continuous extensions of standard cluster criteria. Finally, Section 5 concludes with a summary of the main results and with an outlook to further research. We present proofs in the appendix.

2. Geometry of partition spaces

tictice? In this section, we show that a partition can be represented as a point in some geometric space, called orbit space. Then we endow orbit spaces $\mathcal{P}$ with metrics δ derived from the Euclidean space and study their properties.

2.1. Partitions

Let $\mathcal{Z} = \{z_1, \dots, z_m\}$ be a set of m data points. A partition X of $\mathcal{Z}$ with ℓ clusters $\mathcal{C}_1, \dots, \mathcal{C}_\ell$ is specified by a matrix $\boldsymbol{X} \in [0, 1]^{\ell \times m}$ such that $\boldsymbol{X}^T \mathbf{1}_\ell = \mathbf{1}_m$, where $\mathbf{1}_\ell \in \mathbb{R}^\ell$ and $\mathbf{1}_m \in \mathbb{R}^m$ are vectors of all ones.

The rows $\boldsymbol{x}_{k:}$ of matrix $\boldsymbol{X}$ refer to the clusters $\mathcal{C}_k$ of partition X. The columns $\boldsymbol{x}_{:j}$ of $\boldsymbol{X}$ refer to the data points $z_j \in \mathcal{Z}$. The elements x_{kj} of matrix $\boldsymbol{X} = (x_{kj})$ represent the degree of membership of data point z_j to cluster $\mathcal{C}_k$. The constraint $\boldsymbol{X}^T \mathbf{1}_\ell = \mathbf{1}_m$ demands that the membership values $\boldsymbol{x}_{:j}$ of data point z_j across all clusters must sum to one.

By $\mathcal{P}_{\ell,m}$ we denote the set of all partitions with ℓ clusters over m data points. Since some clusters may be empty, the set $\mathcal{P}_{\ell,m}$ also contains partitions with less than ℓ clusters. Thus, we consider $\ell \leq m$ as the maximum number of clusters we encounter. If the exact numbers ℓ and m do not matter or are clear from the context, we also write $\mathcal{P}$ for $\mathcal{P}_{\ell,m}$. A hard partition X is a partition with matrix representation $\boldsymbol{X} \in \{0, 1\}^{\ell \times m}$. The set $\mathcal{P}^+ \subset \mathcal{P}$ denotes the subset of all hard partitions.

2.2. The orbit space of partitions

The representation space $\mathcal{X}$ of the set $\mathcal{P} = \mathcal{P}_{\ell,m}$ of partitions is a set of the form

$$\mathcal{X} = \{\boldsymbol{X} \in [0, 1]^{\ell \times m} \,:\, \boldsymbol{X}^T \mathbf{1}_\ell = \mathbf{1}_m\}.$$

Then we have a natural projection

$$\pi : \mathcal{X} \to \mathcal{P}, \quad \boldsymbol{X} \mapsto X = \pi(\boldsymbol{X})$$

that sends matrices $\boldsymbol{X}$ to partitions X they represent. The map π conveys two properties: (1) π is surjective: each partition can be represented by at least one matrix, and (2) π is not injective: a partition may have several matrix representations.

Suppose that matrix $\boldsymbol{X} \in \mathcal{X}$ represents a partition $X \in \mathcal{P}$. The subset of all matrices representing X forms an equivalence class $[\boldsymbol{X}]$ that can be obtained by permuting the rows of matrix $\boldsymbol{X}$ in all possible ways. The equivalence class of $\boldsymbol{X}$, called orbit henceforth, is of the form

$$[\boldsymbol{X}] = \{\boldsymbol{PX} \,:\, \boldsymbol{P} \in \Pi\},$$

where Π is the group of all $(\ell \times \ell)$-permutation matrices. The orbit space of partitions is the set

$$\mathcal{X}/\Pi = \{[\boldsymbol{X}] \,:\, \boldsymbol{X} \in \mathcal{X}\}.$$

The orbit space consists of all orbits $[\boldsymbol{X}]$, we can construct as described above. Mathematically, the orbit space $\mathcal{X}/\Pi$ is the quotient space obtained by the action of the permutation group Π on the set $\mathcal{X}$. The orbits $[\boldsymbol{X}]$ are in 1-1-correspondence with the partitions $X = \pi(\boldsymbol{X})$. Therefore, we can identify partitions with orbits and occasionally write $\boldsymbol{X} \in X$ if $X = \pi(\boldsymbol{X})$.

2.3. Metric structures

This section endows the partition space $\mathcal{P}$ with metrics that are derived by a generic construction principle. As examples, we consider metrics δ_p derived from l_p-metrics of Euclidean spaces. We show that $(\mathcal{P}, \delta_p)$ is a compact metric space for $p \geq 1$, and $(\mathcal{P}, \delta_2)$ is a geodesic space.

Every metric d on the representation space $\mathcal{X} \subset \mathbb{R}^{\ell \times m}$ induces a distance function

$$\delta : \mathcal{P} \times \mathcal{P} \to \mathbb{R}, \quad (X, Y) \mapsto \min\{d(\boldsymbol{X}, \boldsymbol{Y}) \,:\, \boldsymbol{X} \in X, \boldsymbol{Y} \in Y\}. \tag{1}$$

Note that the minimum in (1) exists, because the orbits $[\boldsymbol{X}]$ and $[\boldsymbol{Y}]$ are finite. As an example, we consider distance functions induced by the l_p-norm. The l_p-norm for matrices $\boldsymbol{X} \in \mathcal{X}$ is defined by

$$||\boldsymbol{X}||_p = \left(\sum_{k=1}^{\ell} \sum_{j=1}^{m} \left|x_{kj}\right|^p\right)^{1/p}$$

for every $p \geq 1$. The l_p-norm induces the distance function

$$\delta_p : \mathcal{P} \times \mathcal{P} \to \mathbb{R}, \quad (X, Y) \mapsto \min\{||\boldsymbol{X} - \boldsymbol{Y}||_p \,:\, \boldsymbol{X} \in X, \boldsymbol{Y} \in Y\},$$

called l_p-distance on $\mathcal{P}$, henceforth.

To show that distances δ on $\mathcal{P}$ induced by metrics d on $\mathcal{X}$ are also metrics, we demand that metric d is permutation invariant. We say, a metric d on $\mathcal{X}$ is permutation invariant, if

$$d(\boldsymbol{PX}, \boldsymbol{PY}) = d(\boldsymbol{X}, \boldsymbol{Y})$$

for all permutations $\boldsymbol{P} \in \Pi$. Permutation invariance means that the metric d is invariant under simultaneously relabeling the clusters of $\boldsymbol{X}$ and $\boldsymbol{Y}$. An example of permutation invariant metrics are the l_p-metrics. The next result shows that permutation invariant metrics on $\mathcal{X}$ induce distances on $\mathcal{P}$ that are again metrics.

Theorem 2.1. *Let $(\mathcal{X}, d)$ be a metric space and let $(\mathcal{P}, \delta)$ be the partition space endowed with distance function δ induced by metric d.* Suppose that d is permutation invariant. Then we have*:*

1. *The distance δ is a metric.*
2. *$(\mathcal{P}, \delta)$ is a compact space.*
3. *$(\mathcal{P}, \delta_2)$ is a geodesic space.*

Theorem 2.1 *presents a generic way to construct metrics on $\mathcal{P}$. Being a compact metric space is a strong property for consistency statements. Being a geodesic space means that any pair of partitions X and Y have a midpoint partition M such that*

$$\delta_2(X, M) = \delta_2(Y, M) = \frac{1}{2}\delta_2(X, Y).$$

Note that being a geodesic space is a necessary and sufficient condition for guaranteeing the midpoint property for all pairs of partitions.

3. Fréchet consensus clustering

This section first formalizes the problem of consensus clustering using the mean partition approach and then studies its asymptotic behavior. For this, we link the consensus function of the mean partition approach to the Fréchet function [8] from mathematical statistics. Then we show that under normal conditions the mean partition approach is consistent and asymptotically normal.