Neurocomputing 72 (2009) 1379–1389

Contents lists available at ScienceDirect

Neurocomputing

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

Nearly homogeneous multi-partitioning with a deterministic generator

Michaël Aupetit

CEA, DAM, DIF, F-91297 Arpajon, France

ARTICLE INFO

Available online 13 January 2009

Keywords:
Homogeneous partition
Distribution
Seriation
Divergence
Deterministic sampling
Multi-partition
Random sampling
Nearest neighbor
Homogeneity measure
Reproducibility

ABSTRACT

The need for homogeneous partitions, where all parts have the same distribution, is ubiquitous in machine learning and in other fields of scientific studies. Especially when only few partitions can be generated. In that case, validation sets need to be distributed the same way as training sets to get good estimates of models' complexities. And when standard data analysis tools cannot deal with too large data sets, the analysis could be performed onto a smaller subset, as far as its homogeneity to the larger one is good enough to get relevant results. However, pseudo-random generators may generate partitions whose parts have very different distributions because the geometry of the data is ignored. In this work, we propose an algorithm which deterministically generates partitions whose parts have empirically greater homogeneity on average than parts arising from pseudo-random partitions. The data to partition are seriated based on a nearest neighbor rule, and assigned to a part of the partition according to their rank in this seriation. We demonstrate the efficiency of this algorithm on toys and real data sets. Since this algorithm is deterministic, it also provides a way to make reproducible machine learning experiments usually based on pseudo-random partitions.

1. Introduction

1.1. The need for homogeneous partitions

In machine learning and in many domains of scientific studies, data sets need to be partitioned to get reliable estimates of some statistics, either through training and validation sets or through patient and control groups [12]; to save computation time through the study of a sufficiently small subset of the entire set [7]; or to perform opinion poles. In all these cases, it happens eventually that averaging the estimates obtained over many different partitions of the same data set is not possible, either because it would exceed some time or money budget (getting good estimates from huge data sets with some machine learning model or acquiring the labels of all the data from human experts, performing a census, etc.) or it is not physically feasible (giving drug and placebo to the same patient because of different partitions). Then if only one partition can be used, it should be homogeneous, i.e. all its parts should have the same distribution, in order to reduce bias of the estimated quantities. Usually, practitioners either rely on simple random draws with equal probability for each datum to generate such a partition or they attempt to control the odds to get a partition with higher homogeneity.

1.2. State of the art

Stratification [4] is a common way to bias randomness in favor of homogeneous partitions. A partition is generated by hand according to some variable, or automatically using vector quantization (e.g. K-means), then data are selected at random inside each stratum to get a part of the partition. However, there is no general rule to obtain a good stratification.

Another way which has been recommended in a machine learning context [6, p. 135] is similar to the accept–reject sampling method: draw several random partitions, and keep the one which minimizes the Kullback–Leibler (KL) divergence between the density functions estimated on each part.[1] Measuring homogeneity can also be carried out using two-sample tests [8]. However, running such tests or building such density estimates is time consuming (up to $O(N^3)$). So the accept–reject approach may need many runs, and so, much time, with no insurance of getting partitions with high homogeneity.

Another way to get homogeneous partitions is to pose the partitioning problem as an optimization one and try to solve it with some optimization method. For example in a simple local search setting, a first partition is generated at random, and then the elements of different parts are permuted so as to maximize the homogeneity of the partition. So for a partition of N data in

[1] Computing the KL-divergence requires the true probability density functions (*pdf*), so the authors proposed to approximate the density of each part with normal densities, parameters being obtained analytically from the data set.

E-mail address: michael.aupetit@cea.fr

0925-2312/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2008.12.024

two parts with equal size, there are $N/2$ possible permutations from the initial state (a binary vector with N bits assigning each datum to one of the two parts), and the diameter of the search space is $N/2$ (no more than $N/2$ permutations are needed to pass from one state to any other). So reaching a local optimum of the search space takes N^2 times the complexity of the homogeneity testing function. However, testing the homogeneity is usually high (up to N^3 if graph-based two-sample tests are used [8]), leading to a $O(N^5)$ suboptimal algorithm. Moreover, the size of the space to explore is huge: $C_N^{N/2} = N!/(N/2)!(N/2)!$ ways to take $N/2$ elements out of a set of N, so there is a high probability to get stuck into a local optimum, without knowing how far it is to the global one.

Another example is the matched random sampling approach devised in [3,12]. The graph non-bipartite matching (NBM) algorithm of Greevy et al. [12] is designed for optimal pairing of treatment and control patients in a medical context [17]. It is based on pairing all the data so as to minimize the sum of the distances within the pairs [10]. This problem is encoded as a weighted graph matching problem, where pairs of vertices (the data) have to minimize the total weight of the edges connecting them. It may provide partitions in two parts with equal size, by assigning randomly each datum of a pair to a different part. The NBM solves a specific instance of the *general maximum weight graph* [15] problem, and algorithms designed to solve it could be used to find partitions with more than two parts. The complexity of optimal algorithms scales in $O(|E|\,|V|^{1/2})$ [15] where $|E|$ and $|V| = N$ are, respectively, the number of edges and vertices in the graph. However, if the complete graph of the data is used, then $|E| = N(N+1)/2$ and so complexity scales in $O(N^{5/2})$. While if a more reasonable proximity graph is built such as the K-nearest neighbor graph (KNNG), or the Gabriel graph (GG) [9], number of edges scales in $O(N)$ but the graph building process itself scales in $O(N^2)$ (KNNG) or $O(N^3)$ (GG). At last, the objective function optimized in these approaches has not been clearly related to an overall homogeneity measure of the partition.

Our work follows yet another way, which is an attempt to build incrementally an homogeneous partition. We propose a deterministic algorithm designed to generate nearly homogeneous partitions in two or more parts, with possibly unequal size of their parts. This algorithm is a heuristic which experimentally increases the homogeneity between the parts of the partition it generates. It is based on seriating the data and subsequently assigning them to each part of the partition according to their rank in the seriation.

In the next section, we define a new divergence measure between continuous *pdf*. Then we show that minimizing this divergence and minimizing a Hausdorf-like distance between finite samples drawn from these *pdf* are equivalent in the limit of an infinite sample size at least for univariate densities and the Euclidean metric. Then in Section 3, we propose a heuristic to minimize this distance measure between finite samples, and study its properties. Finally in Section 4 we demonstrate its efficiency in getting homogeneous partitions on multivariate artificial and real data sets.

2. Measuring homogeneity

2.1. A new divergence measure

Here we propose a new symmetric divergence measure between two *pdf*. It exists several divergence measures [11] intending to measure how far two *pdf* are from each other. Given *pdf* p and q with the same support X, the KL divergence D_{KL} is given by

$$D_{KL}(p,q) = \int_X p(x)\log\left(\frac{p(x)}{q(x)}\right)\mathrm{d}x,$$

the Jensen–Shannon divergence is given by

$$D_{JS} = \frac{1}{2}\left(D_{KL}\left(p,\frac{p+q}{2}\right) + D_{KL}\left(q,\frac{p+q}{2}\right)\right)$$

and the Renyi's divergence is defined as

$$D_R(p,q) = \frac{1}{\alpha-1}\ln\left(\int_X [p(x)]^{\alpha}[q(x)]^{1-\alpha}\,\mathrm{d}x\right), \quad \alpha\neq 1,\ \alpha>0$$

which corresponds to Bhattacharya's divergence for $\alpha = \frac{1}{2}$. The main problem with these divergence measures is the need for a functional form of the *pdf*. Here we propose the following divergence measure:

Definition.

$$D_H(p,q) = \int_X \log\left(\frac{1}{2}\left(\frac{p(x)}{q(x)} + \frac{q(x)}{p(x)}\right)\right)\mathrm{d}x$$

We prove the following properties:

Property 1. $\forall p,q>0, D_H(p,q) = D_H(q,p)$.

Property 2. $\forall p>0, D_H(p,p) = 0$.

Property 3. $\forall p,q>0, D_H(p^*,q^*) = \min_{\forall p,q>0}(D_H(p,q)) \Leftrightarrow p^* = q^*$.

Property 4. $\forall p,q>0, D_H(p,q) \geqslant 0$.

Proofs are reported in the appendix. The triangular inequality is not true in general, so $D_H(p,q)$ is not a true distance function between *pdf* p and q, but it is a symmetric divergence measure between these *pdf*.

2.2. A homogeneity criterion

Now we define a homogeneity criterion between finite data sets and show that minimizing this criterion for points on the real line is equivalent to minimizing the divergence measure D_H defined above.

It is expected that two parts of a data set have the same distribution if each datum in one part is close enough to a datum in the other one. The closer the points the closer the distribution densities. At best, if both parts are exactly superimposed (case of a data set where data are replicated twice, each one in a different part), then distances between data of one part and their nearest neighbor (their clone) in the other part are zeros. Obviously, both parts being identical have the same distribution.

So, given a set of data $S = (x_1,\ldots,x_N) \in (\mathbb{R}^D)^N$, and two subsets $P_1 \subset S$ and $P_2 \subset S$, we compute the homogeneity measure between P_1 and P_2 as the sum of the distances between each datum of one set and its nearest datum in the other set:

$$H_d(P_1,P_2) = \sum_{u\in P_1}\min_{v\in P_2}(d(u,v)) + \sum_{v\in P_2}\min_{u\in P_1}(d(v,u)),$$

where d is the distance measure. One can see that if we replace all the sums by the max operator, $H_d(P_1,P_2)$ corresponds to the Hausdorff distance *wrt* d, between the finite sets P_1 and P_2.

Now we show that H_d is relevant to measure the homogeneity between two sets on the real line *wrt* the Euclidean metric. Indeed, we show that finding sets P_1 and P_2 which minimize $H_d(P_1,P_2)$ is equivalent to finding sets P_1 and P_2 whose respective densities p_1 and p_2 minimize the divergence $D_H(p_1,p_2)$.

Let $S \in (\mathbb{R})^N$ be a set of N points on the real line, with even N. Let P_1 and P_2 be two subsets partitioning S such that $|P_1| = |P_2| = N/2$, $P_1 \cup P_2 = S$ and $P_1 \cap P_2 = \emptyset$. Let p_1 and p_2 be absolutely continuous probability measures with common