Information Sciences 181 (2011) 2316–2339

Contents lists available at ScienceDirect

Information Sciences

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

Ectropy of diversity measures for populations in Euclidean space

Bakir Lacevic *, Edoardo Amaldi

Politecnico di Milano, Dipartimento di Elettronica e Informazione, Piazza Leonardo da Vinci 32, 20133 Milan, Italy

ARTICLE INFO

Article history:
Received 10 May 2010
Received in revised form 10 October 2010
Accepted 11 December 2010
Available online 31 December 2010

Keywords:
Population diversity measure
Ectropy
Klee's measure
Minimum spanning tree
Singular values
Discrepancy

ABSTRACT

Measures to evaluate the diversity of a set of points (population) in Euclidean space play an important role in a variety of areas of science and engineering. Well-known measures are often used without a clear insight into their quality and many of them do not appropriately penalize populations with a few distant groups of collocated or closely located points. To the best of our knowledge, there is a lack of rigorous criteria to compare diversity measures and help select an appropriate one. In this work we define a mathematical notion of ectropy for classifying diversity measures in terms of the extent to which they tend to penalize point collocation, we investigate the advantages and disadvantages of several known measures and we propose some novel ones that exhibit a good ectropic behavior. In particular, we introduce a quasi-entropy measure based on a geometric covering problem, three measures based on discrepancy from uniform distribution and one based on Euclidean minimum spanning trees. All considered measures are tested and compared on a large set of random and structured populations. Special attention is also devoted to the complexity of computing the measures. Most of the novel measures compare favorably with the classical ones in terms of ectropy. The measure based on Euclidean minimum spanning trees turns out to be the most promising one in terms of the tradeoff between the ectropic behavior and the computational complexity.

1. Introduction

Given a population, i.e., a set of points (vectors, individuals) in a metric space, its diversity is undoubtedly one of the most important indicators of its state. Measures to evaluate the diversity of a population play an important role and are studied (or just used) in a variety of areas such as evolutionary algorithms [65,62,1,21,51,6,4,3,77,78,39,22,32,29,70,68,12,15,33,11,44,50,72,53,71,48,58]; swarm intelligence [59,30,61,9]; numerical integration, quasi Monte Carlo methods and motion planning [28,38,66,18,47,19,42,43,13,24]; classifier systems [79,23,55,34,64,17], biology, environmental sciences and econometrics [69,73,67,52,14,45,63]. Surprisingly, not much effort has been undertaken so far for a systematic analysis and comparison of existing population diversity measures. Moreover, there is a lack of rigorous criteria to assess the quality of population diversity measures and compare them.

In the seminal paper [69], Weitzman discusses the existence of a diversity function that captures the "value of diversity" of a collection of points and proposes a function based on the dissimilarities between points. Assuming ultrametricity of the dissimilarity measure, the diversity function can theoretically be generated recursively using dynamic programming. In [50] Morrison and Jong give a brief history of population diversity measures and introduce a measure that extends the concept of moment of inertia to arbitrarily high dimensional spaces. Wineberg and Oppacher informally define a diversity measure as the answer to the question "how different is everybody from everybody else?" [72]. In [72,53,71] they show that commonly

* Corresponding author.
E-mail addresses: lacevic@elet.polimi.it (B. Lacevic), amaldi@elet.polimi.it (E. Amaldi).

0020-0255/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2010.12.004

used measures (mainly from evolutionary computation) reduce to cumulated difference between all possible pairs of points in the population (this also holds for the measures described in [50]). In [37] Lacevic et al. point out that the measures based on the sum (average) of pairwise distances between points in $\mathbb{R}^m$ have substantial shortcomings. For instance, these measures may reach their maximum value when the population consists of very few (sometimes only two) mutually distant clusters of collocated or closely located points. However, the approach presented lacks generality w.r.t. the space dimension. Mattiussi et al. [48] analyze different diversity measures for populations of strings of possibly variable length, paying special attention to the computational complexity issue. In [34,64], the authors provide an extensive analysis of various diversity measures for constructing classifier ensembles and a deeper understanding of the principle of diversity, but the specific nature of the field substantially mitigates the generality of the concepts. Similarly, the notion of diversity is ubiquitous in ecology and related disciplines, and various (prevalently entropy-based) indices are used to estimate biodiversity [73,67,52,14,45,63]. The very interesting problem of uniform design is considerably studied in the literature (see e.g. [28,38,66,18,47,19,42,43]) with a focus on uniform sampling (used for numerical integration, quasi Monte Carlo method, computer graphics, etc.). Various discrepancy functions are designed to measure the deviation of a given set of points from a uniform distribution and many generators of "small-discrepancy" sequences are available, e.g. [38,66]. However, we know of no attempt to explicitly bring this theory together with the related concept of diversity.

In many applications, a high-quality diversity measure is needed to evaluate the state of a population. Diversity measures are typically used for pure a posteriori analysis or estimates are used to predict some suitable parameter setups that implicitly affect diversity [6,4,3,77]. An accurate diversity measure is also particularly important when some decisions are influenced by the diversity via feedback, for instance, when parameters or the structure of an algorithm change (in terms of measured diversity) during the execution [65,62,77,59,30,79]. Furthermore, a proper insight into the nature of a diversity measure could help improve algorithms that aim at diversity enhancement [57,76,46].

Several approaches are widely used for measuring population diversity. In a first approach, the diversity is evaluated in genotypic space (when the population consists of binary strings or strings of "genes" from any alphabet) using either deterministic formulas (e.g., Hamming distance or another metric) [62,1,21,72,53,71,48] or entropy calculations [51,15,49,72,53,71,48,73,67,52,14,45,63]. In a second approach, the diversity is computed in parameter (phenotypic) space, usually in $\mathbb{R}^m$, where $m \in \mathbb{N}$. The two most frequently used measures are based on summing (averaging) the distances of all the points to the centroid point (see [65,6,72,53,71,59,30,61,46]) or the distances between all pairs of points in the population (see [51,4,3,72,53,71,48,61,46]). Similar techniques include the column-based variance and the moment of inertia diversity measure [50]. In spite of the different expressions, all of these measures rely on the distances between all possible pairs of points [72]. Apparently, this is also true for the Shannon entropy-based diversity measure [72].

The main focus of this paper is on diversity measures in $\mathbb{R}^m$, although some attention is also devoted to measures for binary coded populations. The main contributions are the following[1]:

- We introduce a mathematical notion of ectropy that allows classification of different diversity measures. The level of ectropy indicates to what extent the diversity measure tends to penalize the collocation/proximity of points. The definition of ectropy is then used to point out some drawbacks of the classical measures.
- We propose several novel measures that overcome the drawbacks of the classical ones. They are designed to capture how uniformly the population is distributed in the domain space. In particular, we introduce a quasi-entropy measure based on a geometric covering problem, three measures based on discrepancy from uniform distribution and one based on Euclidean minimum spanning trees. Furthermore, special attention is devoted to the complexity of computing the diversity measures.
- An extensive computational study is conducted in order to establish the mutual dependencies among different measures. This study supports the theoretical analysis performed for a variety of measures. On the other hand, it provides valuable insight into the behavior of those measures for which the rigorous analysis could not be carried out.

The remainder of the paper is organized as follows. In Section 2, the ectropic property of the diversity measure is defined, while in Section 3, some drawbacks of the classical diversity measures are pointed out. In Section 4 we analyze several alternative diversity measures and prove that they overcome some of the shortcomings of classical measures. Our setup for the extensive computational study is presented in Section 5, along with the numerical results and comments. Finally, Section 6 contains some concluding remarks and future work directions.

2. Ectropic property of diversity measures

2.1. Notation

Let $X \subseteq \mathbb{R}^m$, with $m \in \mathbb{N}$, and let $\mathbf{P}(X,n) = (\mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_n)$, with $n \cdots\ n \in \mathbb{N}$, be a population of n points, given in a matrix form. The column $\mathbf{x}_k$ represents the k-th point $\mathbf{x}_k = (x_{1k}\ x_{2k} \cdots x_{mk})^T \in X$, with $k = 1,2,\ldots,n$. Whenever convenient, a population $\mathbf{P}$ is represented in the form of the set $\mathbf{P} = \{\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n\}$. Then the cardinality $|\mathbf{P}(X,n)|$ of the population $\mathbf{P}(X,n)$ is at most n and

[1] Selected preliminary results of this paper can be found in [36].