Population Diversity Index: A New Measure for Population Diversity

S.K. Smit
Vrije Universiteit Amsterdam
Dept. of Computer Science
Amsterdam, The Netherlands
sksmit@cs.vu.nl

Z. Szlávik
Vrije Universiteit Amsterdam
Dept. of Computer Science
Amsterdam, The Netherlands
z.szlavik@vu.nl

A.E. Eiben
Vrije Universiteit Amsterdam
Dept. of Computer Science
Amsterdam, The Netherlands
gusz@cs.vu.nl

ABSTRACT

A number of diversity measures used in evolutionary computing suffer from 'mis-measuring' the diversity of populations. In this paper, we identify and demonstrate this problem using a carefully engineered test suite of six differently arranged populations. We also propose a new measure, called Population Diversity Index (PDI), that solves the problem. We show that sorting the test configurations by their PDI value we obtain a correct ranking (i.e., a natural one, conformant with the human-perceived order). PDI also allows for a comparison between populations of different sizes and genome-dimensions, and its relation to the uniform distribution makes the calculated diversity values easy to interpret.

Categories and Subject Descriptors

I.2.8 [**Artificial Intelligence**]: Problem Solving, Control Methods, and Search

General Terms

Theory

Keywords

population, diversity, measurements

1. INTRODUCTION

In this paper we identify, demonstrate and solve a problem with the current measures of population diversity in evolutionary computing. Existing diversity measures can sometimes 'mis-measure' the diversity of populations. To be specific, comparing the diversity of two given populations can lead to a ranking different from the natural one – the one a human experimenter would make by looking the population plots. Obviously, this could be seen as the fault of the human eye, and one could welcome a mathematically well founded comparison as a fix. However, we argue that this is not the right interpretation. To substantiate our argument, we hand-craft a number of test configurations (differently arranged populations in a 2D space) in such a way that the spread of the population in each configuration is clearly different.

On a conceptual level, we introduce the notion of *conformity* with respect to population diversity measures. We call a diversity measure *conformant* if it ranks different populations according to the human eye. Furthermore, we demonstrate that the five most commonly used diversity measures in the genotype space, all have conformity problems. Finally, we introduce a new diversity measure called Population Diversity Index (PDI), intended to fix this. The experimental comparison shows that it is indeed highly correlated with the perceived diversity of a population.

GECCO'11, July 12–16, 2011, Dublin, Ireland.
ACM 978-1-4503-0690-4/11/07.

2. CURRENT DIVERSITY MEASURES

The diversity of a population can be measured either in the genotypic space (measuring the variance in genomes), or in the phenotypic space (measuring the variance in fitness). Since an infinite number of genomes can map to a single fitness, measures based on the genotypic space provide more information about the spread of the population than phenotypic measures. Therefore, we focus on genotypic diversity measures. The class of genotypic diversity measures can be divided into two sub-classes, namely *Genome-based measures* and *Gene-based measures*. Genome-based measures treat the genome as a whole, while Gene-based measure assess one gene at a time and calculate the average of these values.

Most often, genome-based measures are based on a distance measure to indicate the difference between two genomes. For example, the commonly used pairwise Hamming distance can be calculated by summing the Hamming distances of all possible pairs in the population. It is mainly used when the genome is a binary vector, but we can use the Manhattan distance (PMD) or Euclidean distance (PED) instead to deal with real-values representations. Another approach is to measure Distance to Population Centroid (PDC)[4]. The most commonly used gene-based measures are: Population Inertia (PI) [2] and Discretized Gene-based Entropy (DGE) [1].

Although each measure is valid way of calculating diversity, their levels of conformity differ, and their values, in some cases, are not intuitive. To illustrate conformity of these measures, we use six populations of 100 individuals in the $[0, 1]^2$ genotypic space (Figure 1). The figures 1.1 to 1.6 are ordered by decreasing *perceived* diversity of their corresponding populations. Arrangement 1 shows a uniformly distributed population over the search space (highest diversity, as one would intuitively determine), arrangement 6 depicts a fully converged population (with, naturally, lowest diversity), while arrangements 2 to 5 show populations with decreasing perceived diversity. Corresponding diversity values, by the measures we discussed so far in this paper, are presented in Table 1. None of the measures order the example populations of Figure 1 in a descending order of perceived diversity, i.e. they fail to capture perceived diversity in at least one sense. In general, the results show that Gene-based measures can not deal with coordinate aligned pattern (such as in the test cases), and centroid-based approaches can not deal with multiple clusters. The most conform is Discretized Gene-based Entropy, since it captures most of the perceived diversity changes. However, its gene-based approach causes problems in case 5. To overcome this limitation, we introduce a new diver-

Table 1: Diversity Values of the Arrangements

	Arrangement					
Measure	**1**	**2**	**3**	**4**	**5**	**6**
PMD	0.330	0.328	0.418	0.423	0.423	0.00
PED	0.328	0.386	0.434	0.443	0.591	0.00
PC	0.269	0.265	0.392	0.406	0.406	0.00
PI	0.165	0.169	0.310	0.330	0.330	0.00
DGE ($b = 10$)	0.721	0.616	0.434	0.428	0.428	0.00
PDI ($\varepsilon = 0.001$)	0.972	0.856	0.679	0.590	0.394	0.00

sity measure called Population Diversity Index (PDI) that treats the genome as a whole, in contrast to one gene at a time.

3. POPULATION DIVERSITY INDEX

To make this new measure conform to the perceived diversity, and easy to interpret, we want to define it in such a way that it is equal to one on a uniformly distributed population, and equal to zero if the data is concentrated in one point. Since the Shannon entropy has exactly these properties, this new measure of diversity is derived using that as a basis. Furthermore, we extend it to cope with real-valued genes and to conform to perceived diversity, we use the perceived similarity between pairs of individuals. Let us consider a set of m discrete samples P, and the set of unique values of P denoted by U. The Shannon entropy of P is calculated as follows.

$$H(P) = -\sum_{\bar{u} \in U} (occ(\bar{u})/m) \cdot log\,(occ(\bar{u})/m)$$

where $occ(\bar{u})$ is the number of times the vector $\bar{u}$ ($\in [0,1]^n$) occurs in P. Since, by definition, each vector $\bar{u}$ from U is present in P $occ(\bar{u})$ times, rather than enumerating over the unique vectors $\bar{u}$ in U, $H_\nu(P)$ can be rewritten to enumerate over all vectors $\bar{x}$ in P. Furthermore, as the maximum entropy of a sample of size m is equal to $log(m)$, we can define the normalized entropy (H_ν), independent of the size of P, as:

$$H_\nu(P) = -\,(\sum_{i=1}^{m} log\,(occ(\bar{x}_i)/m))/(m \cdot log(m)) \qquad (1)$$

In order to handle continuous valued samples, i.e. to produce entropy values other than $log(m)$, we would like to replace the $occ(\bar{x}_i)$ function in Equation 1 with another function we denote by $p(\bar{x}_i)$. $p(\bar{x}_i)$ is to be based on how close a sample $\bar{x}_i$ is to members of P (including itself). Population Diversity Index (PDI) is defined by inserting $p(\bar{x}_i)$ into Equation 1:

$$PDI = -\,(\sum_{i=1}^{m} log(p(\bar{x}_i)))/(m \cdot log(m)) \qquad (2)$$

To calculate how close two individuals $(\bar{x}_1, \bar{x}_2)$ are, we consider the Euclidean distance ($d(\bar{x}_1, \bar{x}_2)$) between them (other distance measures are also possible). Since perceived diversity is closely related to the perceived similarity, we base our measure of diversity on Shepard's Universal Law of Generalization [3]. Accordingly, the perceived similarity between two objects has an exponential relation to their corresponding distance. Thus, we define the similarity between two individuals in the population as:

$$s(\bar{x}_1, \bar{x}_2) = e^{-\omega \cdot d(\bar{x}_1, \bar{x}_2)}, \text{ with } \omega = -log(\varepsilon)/\alpha_n \qquad (3)$$

where $\varepsilon \in (0,1)$ is a scaling parameter, and α_n the expected distance between any two individuals if they were uniformly distributed over $[0,1]^n$. This definition of similarity ensures that if the distance between two samples is lower than the expected distance

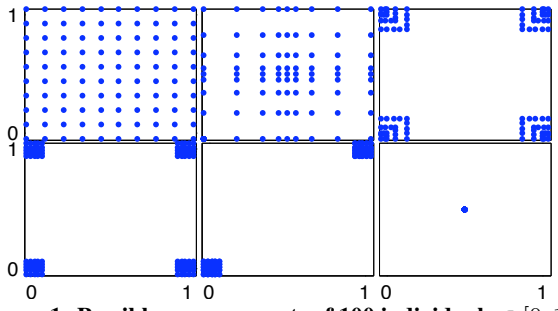

Figure 1: Possible arrangements of 100 individuals $\in [0,1]^2$

α_n, then the corresponding similarity will be higher than ε. Similarly, if the distance is higher than α_n, similarity will be lower than ε. Also, if two samples are identical, their similarity will be equal to one. For each individual $\bar{x}_i$ in the population, we calculate $\hat{p}_i$, the average similarity of $\bar{x}_i$ to members of the population:

$$\hat{p}(\bar{x}_i) = \sum_{j=1}^{m} s(\bar{x}_i, \bar{x}_j)/m \qquad (4)$$

Furthermore, we denote the expected similarity when the distribution of the population is uniform by $\beta_{\varepsilon,n}$. Since we would like the value for $p(\bar{x}_i)$ in Equation 2 to be equal to $\frac{1}{m}$, if and only if the distribution of the population is uniform, $\hat{p}(\bar{x}_i)$ needs to be transformed in such a way that $\beta_{\varepsilon,n}$ is transformed into $\frac{1}{m}$. A function that accomplishes this is: $f(\hat{p}(\bar{x}_i)) = \hat{p}(\bar{x}_i)^\varsigma$, where $\varsigma$is equal to:

$$\varsigma = -\,log(m)/log(\beta_{\varepsilon,n}) \qquad (5)$$

Therefore PDI becomes:[1]

$$PDI = -\frac{\sum_{i=1}^{m} log(p(\bar{x}_i))}{m \cdot log(m)} = -\frac{\sum_{i=1}^{m} log(\hat{p}(\bar{x}_i)^\varsigma)}{m \cdot log(m)} \qquad (6)$$

PDI is not restricted to boolean or real-valued genotypes, but can be extended to incorporate any genome for which an appropriate distance measure can be defined (e.g., a population of trees). Generalizability and its high level of conformity are the main advantages of using the Population Diversity Index for measuring and assessing diversity.

4. REFERENCES

[1] N. Mori, H. Kita, and Y. Nishikawa. Adaptation to a changing environment by means of the thermodynamical genetic algorithm. In H.-M. Voigt et al., editors, *PPSN IV*, *Lecture Notes in Computer Science*, pages 513–522. Springer, 1996.

[2] R. W. Morrison and K. A. D. Jong. Measurement of population diversity. In *Selected Papers from the 5th European Conference on Artificial Evolution*, pages 31–41, London, UK, 2002. Springer.

[3] R. N. Shepard. Toward a universal law of generalization for psychological science. *Science*, 237(4820):1317–1323, 1987.

[4] R. Ursem. Diversity-guided evolutionary algorithms. In J. Guervós et al., editors, *PPSN VII*, *Lecture Notes in Computer Science*, pages 462–471. Springer, 2002.

[1]Since neither α nor β have closed forms, they need to be approximated. An approximation for various dimensions can be found at: http://www.cs.vu.nl/~sksmit